Big-Data-Technologien – Wissen für Entscheider

Leitfaden

BITKOM

Impressum

Herausgeber:	BITKOM Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. Albrechtstraße 10 A | 10117 Berlin-Mitte Tel.: 030.27576-0 | Fax: 030.27576-400 bitkom@bitkom.org | www.bitkom.org
Ansprechpartner:	Dr. Mathias Weber Tel.: 030.27576-121 m.weber@bitkom.org
Verantwortliches Gremium:	BITKOM-Arbeitskreis Big Data
Projektleitung:	Guido Falkenberg, Senior Vice President Product Marketing, Software AG Dr. Holger Kisker, Vice President & Research Director, Forrester Germany GmbH Jürgen Urbanski, Managing Director, TechAlpha
Copyright:	BITKOM 2014
Grafik/Layout:	Design Bureau kokliko/ Astrid Scheibe (BITKOM)
Titelbild:	© fotolia.com.com – agsandrew

Big-Data-Technologien – Wissen für Entscheider

Leitfaden

BITKOM

Inhaltsverzeichnis

Verzeichnis der Abbildungen

Verzeichnis der Tabellen

Autoren des Leitfadens

- Jörg Bartel, IBM Deutschland GmbH
- Arnd Böken, Graf von Westphalen Rechtsanwälte Partnerschaft
- Florian Buschbacher, PricewaterhouseCoopers AG Wirtschaftsprüfungsgesellschaft
- Guido Falkenberg, Software AG
- Johannes Feulner, fun communications GmbH
- Dr. Georg Fuchs, Fraunhofer IAIS Institut für Intelligente Analyse- und Informationssysteme
- Nadine Gödecke, Fraunhofer MOEZ Zentrum für Mittel- und Osteuropa
- Dr. Holmer Hemsen, Technische Universität Berlin
- Stefan Henß, Technische Universität Darmstadt
- Ralph Kemperdick, Microsoft Deutschland GmbH
- Dr. Holger Kisker, Forrester Germany GmbH
- Dr. Sebastian Klenk, EXASOL AG
- Hardy Klömpges, Atos IT Solutions and Services GmbH
- Holm Landrock, Experton Group AG
- Dr. Mario Lenz, Empolis Information Management GmbH
- Stefan Lipp, Talend Germany GmbH
- Dr. Dirk Mahnkopf, SAS Institute GmbH
- Prof. Dr. Volker Markl, Technische Universität Berlin
- Axel Mester, IBM Deutschland GmbH
- Dr. Gerhard Paaß, Fraunhofer IAIS Institut für Intelligente Analyse- und Informationssysteme
- Dr. Andreas Ribbrock, Teradata GmbH
- Oliver Roser, SEMANTIS GmbH
- Dr. Stefan Rüping, Fraunhofer IAIS Institut für Intelligente Analyse- und Informationssysteme
- Dr. Fritz Schinkel, Fujitsu Technology Solutions GmbH
- Karl-Heinz Sylla, Fraunhofer IAIS Institut für Intelligente Analyse- und Informationssysteme
- Georg Urban, Microsoft Deutschland GmbH
- Jürgen Urbanski, TechAlpha
- Prof. Dr. Holger K. von Jouanne-Diedrich, Hochschule Aschaffenburg
- Dr. Angi Voß, Fraunhofer IAIS Institut für Intelligente Analyse- und Informationssysteme
- Dr. Mathias Weber, BITKOM e.V.
- Hans Wieser, Microsoft Deutschland GmbH

An der Entwicklung des Leitfadens haben weiterhin mitgewirkt:

- Markus Brandes, Atos Information Technology GmbH
- Dr. Mark von Kopp, SAP AG
- Wulf Maier, Hewlett-Packard GmbH

Liste der Abkürzungen

ACL	Access Control List
ACORD	Association for Cooperative Operations Research and Development
ANSI	American National Standards Institute
API	Application Programming Interface
BI	Business Intelligence
BPM	Business Process Management
CAP	Consistency – Availability – Partition Tolerance
CEP	Complex Event Processing
CMS	Content-Management-System
CRAN	Comprehensive R Archive Network
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRM	Customer Relationship Management
CTO	Chief Technology Officer
DASD	Direct Access Storage Devices
DBMS	Database Management System
DOLAP	Desktop OLAP
DSCP	Distributed Stream Computing Platform
DWH	Data Warehouse
EDA	Explorative visuellen Datenanalyse
EDW	Enterprise Data Warehouse
EIT	Europäisches Institut für Innovation und Technologie
ELT	Extract – Load – Transform
EPL	Event Processing Language
ERP	Enterprise Resource Planning
ESB	Enterprise Service Bus
ETL	Extract – Transform – Load
FISMA	Federal Information Security Management Act
FTP	File Transfer Protocol
GB	Gigabyte
GIS	Geoinformationssystem
GLBA	Gramm-Leach-Bliley Act
GNU	GNU's Not Unix
GPFS	General Parallel File System
GPFS	General Parallel File System
GPS	Global Positioning System
GUI	Graphical User Interface
HANA	High Performance Analytic Appliance
HCM	Human Capital Management
HDSF	Hadoop Distributed File System
HFAT	Hochfrequentes algorithmisches Trading
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level 7
HOLAP	Hybrid OLAP
HTTP	Hypertext Transfer Protocol
IE	Informationsextraktion
IEC	International Electrotechnical Commission
IMDG	In-Memory Data Grid
IO	Input-Output
IoT	Internet of Things
ISO	International Organization for Standardization
JDBC	Java Database Connectivity
JSON	JavaScript Object Notation
LDAP	Lightweight Directory Access Protocol
M2M	Machine-to-Machine
MB	Megabyte
MDM	Master Data Management
MDX	Multidimensional Expressions
MOLAP	Multidimensionales OLAP
MPP	Massively Parallel Processing
MTBF	Mean Time Between Failures
NAS	Network-Attached Storage
NFS	Network File System (auch Network File Service)
NTSB	National Transportation Safety Board
OCR	Optical Character Recognition
ODBC	Open Database Connectivity
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
OTC	Over-the-Counter

PACT	Parallelization Contracts
PAM	Pluggable Authentication Module
PCI	Peripheral Component Interconnect
PCI DSS	Payment Card Industry Data Security Standards (
PDD	Persönliche digitale Datenwirtschaft
POS	Part-of-Speech
RAM	Random Access Memory
RDBMS	Relational Database Management System
REST	Representational State Transfer
ROLAP	Relational OLAP
SaaS	Software as a Service
SATA	Serial Advanced Technology Attachment
SCM	Supply Chain Management
SELinux	Security-Enhanced Linux
SMP	Symmetrisches Multiprozessorsystem
SOX	Sarbanes-Oxley
SQL	Structured Query Language
SSD	Solid State Disk
SSL/TLS	Secure Sockets Layer/Transport Layer Security
SWIFT	Society for Worldwide Interbank Financial Telecommunication
TCO	Total Cost of Ownership
UDF	User Defined Function
UIMA	Unstructured Information Management Architecture
VA	Visual Analytics
VPN	Virtual Private Network
XML	Extensible Markup Language
YARN	Yet Another Resource Negotiator

Geleitwort

Prof. Dieter Kempf – BITKOM Präsident,
Vorsitzender des Vorstands Datev eG

In der modernen Wirtschaft werden Daten immer wichtiger. Verbraucher äußern sich in Online-Foren oder sozialen Netzwerken über Produkte und Services, die Verbreitung mobiler Endgeräte steigt rasant und ihr Einsatz wird immer vielfältiger. Medizinische Kleinstgeräte überwachen Vitalfunktionen von Patienten und melden verdächtige Veränderungen. Sensoren an Produktionsmaschinen, Turbinen, Fahrzeugen oder wissenschaftlichen Geräten erfassen den Zustand hunderter Parameter.

Die intelligente Auswertung der Daten kann Organisationen wichtige Informationen liefern. Unternehmen gewinnen zum Beispiel neue Erkenntnisse über Marktentwicklungen und Kundenbedürfnisse. Es ist offensichtlich, dass solche Unternehmen einen Wettbewerbsvorteil erlangen, die aus der Vielzahl der Daten geschäftsrelevante Informationen filtern können. Das ist das Feld von Big Data Analytics.

Der vorliegende Leitfaden des BITKOM-Arbeitskreises Big Data soll Entscheidern die Orientierung in dieser komplexen Materie erleichtern. An der Erstellung der Publikation waren mehr als 30 Autoren beteiligt: IT-Spezialisten, Markt- und Technologie-Analysten, Juristen und Wirtschaftsprüfer, Wissenschaftler sowie Experten aus Organisationen der Aus- und Weiterbildung. Besonders wichtig war uns die Mitwirkung der Rechtsexperten, denn Big-Data-Analysen, die personenbezogene Daten einbeziehen, müssen schon in der Konzeptphase auf ihre Gesetzeskonformität geprüft werden.

Big Data wird in Unternehmen nur selten völlig neu aufgesetzt. In der Regel werden die bestehenden IT-Lösungen genutzt und erweitert. Der Leitfaden zeigt zum Beispiel, wie Unternehmen vorhandene Business-Intelligence-Anwendungen durch neue Ansätze anreichern können. Unternehmen stehen beim Einsatz von Big Data häufig vor einer Hürde: Spezialisten für Data Science sind Mangelware und müssen oft erst ausgebildet werden. Inzwischen gibt es erste Hochschulprogramme und Weiterbildungskurse. Dazu gibt der Leitfaden ebenfalls einen Überblick.

Big-Data-Technologien können nicht isoliert betrachtet werden. Big Data, Cloud Computing, Mobile Computing und Industrie 4.0 befruchten sich gegenseitig und können die Wettbewerbsfähigkeit der deutschen Unternehmen nachhaltig stärken. Ihr Einsatz kann auch einen Beitrag zur Ertüchtigung unserer Infrastrukturen liefern: Intelligente Netze für Energie, Verkehr, Gesundheit oder Verwaltung sind ohne Big Data kaum noch denkbar. Daher begrüßen wir, dass die Politik mit einem Technologieprogramm bereits erste Schritte zur Beschleunigung des Big-Data-Einsatzes in verschiedenen Sektoren unternommen hat.

Prof. Dieter Kempf
BITKOM Präsident

1 Management Summary

Einordnung

Dieser Leitfaden ist Bestandteil einer Serie von BITKOM-Publikationen über Big Data. Der erste Leitfaden mit dem Titel »Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte« erschien 2012. Der zweite Leitfaden über das »Management von Big-Data-Projekten« folgte zum 1. Big Data Summit im Juni 2013. Der vorliegende dritte Leitfaden richtet sich in erster Linie an Entscheidungsträger aus der Wirtschaft, gibt einen Überblick über die Big-Data-Technologien und soll so Technologieentscheidungen erleichtern. Aus Gründen des Umfangs spart der Leitfaden 3 den Aspekt aus, welche neuen Datenprodukte und –dienstleistungen rund um Big Data entstehen: Den Startschuss für den vierten Leitfaden hat der BITKOM im Januar 2014 gegeben. Die für den Sommer 2014 geplante Publikation soll an Beispielen aus der Wirtschaft zeigen, welche neuen Geschäftsmodelle sich bei den Big-Data-Nutzern herausbilden.

Begriffsbestimmung

Im Leitfaden 1 wurde Big Data als Einsatz großer Datenmengen aus vielfältigen Quellen mit einer hohen Verarbeitungsgeschwindigkeit zur Erzeugung wirtschaftlichen Nutzens bezeichnet. Big Data hat vier wesentliche Facetten:

- **Datenmenge (Volume):** Immer mehr Organisationen und Unternehmen verfügen über gigantische Datenberge, die von einigen Terabytes bis hin zu Größenordnungen von Petabytes führen.
- **Datenvielfalt (Variety):** Unternehmen haben sich mit einer zunehmenden Vielfalt von Datenquellen und Datenformaten auseinanderzusetzen. Aus immer mehr Quellen liegen Daten unterschiedlicher Art vor, die sich grob in unstrukturierte, semistrukturierte und strukturierte Daten gruppieren lassen. Gelegentlich wird auch von polystrukturierten Daten gesprochen. Die unternehmensinternen Daten werden zunehmend durch externe Daten ergänzt, beispielsweise aus sozialen Netzwerken.
- **Geschwindigkeit (Velocity):** Riesige Datenmengen müssen immer schneller ausgewertet werden, nicht selten in Echtzeit. Die Verarbeitungsgeschwindigkeit hat mit dem Datenwachstum Schritt zu halten. Damit sind folgende Herausforderungen verbunden: Analysen großer Datenmengen mit Antworten im Sekundenbereich, Datenverarbeitung in Echtzeit, Datengenerierung und Übertragung in hoher Geschwindigkeit.
- **Analytics:** Analytics umfasst die Methoden zur möglichst automatisierten Erkennung und Nutzung von Mustern, Zusammenhängen und Bedeutungen. Zum Einsatz kommen u.a. statistische Verfahren, Vorhersagemodelle, Optimierungsalgorithmen, Data Mining, Text- und Bildanalytik. Bisherige Datenanalyse-Verfahren werden dadurch erheblich erweitert.

Vielfalt der Technologien

Big Data basiert nicht auf einer singulären Technologie, sondern ist vielmehr das Resultat des Zusammenwirkens einer ganzen Reihe von Innovationen in verschiedenen Gebieten. Insgesamt erlauben diese Fortschritte, aus immer mehr Daten einen immer höheren betriebswirtschaftlichen Nutzen zu ziehen. Je nach Anwendungsszenario können hierbei verschiedene Technologiekonzepte zum Einsatz kommen. (Kapitel 3)

Klassische Technologien – Big-Data-Anforderungen nicht gewachsen

Der vom Wettbewerb ausgehende Druck auf Unternehmen, schnell rapide zunehmende Datenmengen zu verarbeiten, beschleunigt sich weiter. Dafür reichen klassische Technologien wie traditionelle Datenbanken, Data Warehouse oder Reporting nicht mehr aus. Heute gilt es, sehr viel mehr Informationen über den Markt und die Kunden zu sammeln und auszuwerten, um weiterhin einen Wettbewerbsvorteil zu erarbeiten. (Kapitel 2)

Big Data – Hebel für neue Geschäftsmodelle

Für die Unternehmen geht es bei Big Data nicht nur um die Verbesserung bestehender Produkte oder Prozesse – viele sehen die Umsatz-Relevanz von Big Data für neue Geschäftsfelder. Analyseergebnisse werden zu neuen Produkten führen, die wiederum neue Vertriebs- und Geschäftsmodelle mit sich bringen werden. (Kapitel 2)

Daten – vierter Produktionsfaktor

Daten werden für viele Branchen auch außerhalb der Informationswirtschaft zu einem Differenzierungsmerkmal und Asset werden. Für Unternehmen wird es in Kürze normal sein, Big-Data-Analysen zur Unterstützung ihrer Entscheidungsprozesse einzusetzen. Auch in der öffentlichen Verwaltung wird Big Data eine zunehmende Rolle spielen. Klare gesetzliche Regelungen können die Ausbreitung von Big Data in Deutschland positiv beeinflussen und sich auf die Wettbewerbsfähigkeit der deutschen Industrie auswirken. (Kapitel 2)

Vier Technologiesegmente – grobe Orientierung in Technologielandschaft

In Abhängigkeit von den konkreten Anforderungen aus dem Anwendungsszenario können verschiedene Architekturen oder auch Kombinationen von Architekturen die zielführende Lösung bilden.

Für eine erste Orientierung lassen sich vier Technologiesegmente unterscheiden. Für Anwendungen mit nicht zu hohen Anforderungen an Zeit und Datenvielfalt eignen sich standardisierte Analytics Appliances. Lösungen mit In-Memory-Technologie kommen zum Einsatz, wenn die Datenauswertung etwa um den Faktor 1000 beschleunigt werden muss. Liegt eine große Vielfalt an Datenformaten vor, empfehlen sich Lösungen auf der Basis von Hadoop. Diese Open-Source-Technologie kann sehr große Mengen unterschiedlich strukturierter Daten speichern und verarbeiten; sie skaliert nahezu unbegrenzt. Streaming und Complex Event Processing bieten sich an, wenn Daten praktisch im Augenblick ihres Anfallens zu erfassen und auszuwerten sind. Diese grobe vorgenommene Segmentierung der Technologielandschaft zeigt, warum unterschiedliche Technologien zum Einsatz kommen. (Kapitel 3)

Taxonomie der Big-Data-Technologien mit sechs Schichten

Eine Taxonomie der Big-Data-Technologien – quasi ein Technologie-Baukasten – bildet den wichtigsten Bezugspunkt für diesen Leitfaden. Die Taxonomie umfasst wesentliche Technologien, die für eine Big-Data-Komplettlösung benötigt werden. Sie ordnet die Technologien in sechs Schichten an.

Die vier Schichten (1) Daten-Haltung, (2) Daten-Zugriff, (3) Analytische Verarbeitung und (4) Visualisierung markieren den direkten Weg von den Rohdaten hin zu geschäftsrelevanten Erkenntnissen. Dieser Weg wird flankiert von (5) Daten-Integration und (6) Daten-Governance sowie Daten-Sicherheit. Diese flankierenden Schichten garantieren, dass sich der Weg von den Rohdaten zur Erkenntnis in existierende Standards großer Unternehmen einbettet und sich ein zielgerichtetes Zusammenwirken von Big Data mit existierenden Technologien, Prozessen und Compliance-Vorgaben einstellt. (Kapitel 3)

Als konstruktiv nutzbare Vorlage für Konzeption und Entwurf einer Big-Data-Anwendung eignet sich die Lambda-Architektur. Die in dieser Architektur vorgesehene Modularisierung spiegelt typische Anforderungen an Big-Data-Anwendungen wider und systematisiert sie. (Kapitel 3)

Hadoop – neue Big-Data-Denkweise

Herkömmliche Lösungen sind angesichts der mit Big Data assoziierten Herausforderungen (»3 V«) sowohl aus technischer als auch aus betriebswirtschaftlicher Sicht eng limitiert. Hadoop bietet eine Antwort auf diese Herausforderungen und hat sich daher als Kern einer modernen Datenarchitektur und Ergänzung herkömmlicher Lösungen fest etabliert.

Von einem preiswerten Daten-Parkplatz hat sich Hadoop so weiter entwickelt, dass es Unternehmensentscheidungen in nahezu Echtzeit durch verschiedenste Analyseverfahren unterstützt. Diese gehen über die Möglichkeiten traditioneller Business Intelligence Tools weit hinaus. Hadoop ist ein Open-Source-Framework zur parallelen Datenverarbeitung auf sehr hoch skalierenden Server-Clustern. Dieses Top-Level-Projekt der Apache Software Foundation umfasst mit seinen zahlreichen Unterprojekten mehrere Schichten in der Taxonomie (Daten-Haltung, -Zugriff, -Integration, -Sicherheit und Betrieb).

Hadoop hat sich für viele Problemstellungen als sehr tragfähig erwiesen und bringt wie ein Motor Big Data voran. Ein breites und innovatives Ökosystem aus Open Source und kommerziellen Produkten liefert in schnellem Tempo Verbesserungen am Hadoop-Framework; so kann sich Hadoop zunehmend zu einer unternehmensweiten, gemeinsamen Daten-Plattform entwickelt – einem Shared Service. (Abschnitt 4.1)

Investitionen in In-Memory- sowie NoSQL-Datenbanken prüfen

Frühere Investitionen der Unternehmen in relationale Datenbanken bilden eine gute Grundlage für Big-Data-Projekte, aber zur Erhaltung der Wettbewerbsfähigkeit reichen sie nicht aus, wenn die vielen innovativen Möglichkeiten in Betracht gezogen werden, um Daten in Geschäftswert und Gewinn umzuwandeln.

Unternehmen sollten Investitionen in In-Memory-Datenbanken (zur Beschleunigung des Zugriffs auf Daten) sowie in NoSQL (für mehr Flexibilität in den Datenstrukturen und bei der Verarbeitung) prüfen. Beide Technologien finden mittlerweile weiten Einsatz und ergänzen die relationalen Datenbanken. (Abschnitt 4.1)

Daten-Zugriff

Der Daten-Zugriff beinhaltet verschiedene Technologien, die es sehr unterschiedlichen analytischen Anwendungen ermöglichen, auf den Pool von großen Datenmengen zuzugreifen.

Der traditionelle Ansatz für Data Warehouse und Big Data analysiert ruhende Daten. Die Überwachung und Steuerung dynamischer Prozesse bedarf eines anderen Ansatzes. Hierbei werden zeitlich geordnete Ereignisse aus heterogenen Quellen überwacht, verdichtet, gefiltert und korreliert. Das ist das Feld von Streaming und Complex Event Processing.

Search- & Discovery-Technologien ermöglichen das Suchen und Entdecken von Informationen in meist unstrukturierten Daten analog zur Suchmaschine Google. (Abschnitt 4.2)

Analytische Verarbeitung

Die Analytische Verarbeitung bildet den eigentlichen Kern von Big-Data-Anwendungen. Die Analytische Verarbeitung umfasst ein ganzes Bündel von Technologien zur Verarbeitung der verschiedenen Datentypen sowie wichtige Themen wie Predictive Analytics, Data Mining und Maschinelles Lernen. (Abschnitt 4.3)

Fortgeschrittene Visualisierung

Fortgeschrittene Visualisierungen stellen ein mächtiges und hochgradig flexibles Werkzeug im Analyseprozess dar, das die algorithmischen Verfahren der Datenanalyse im Sinne von »Unsicherheit minimieren durch visuellen Check« entscheidend ergänzt. (Abschnitt 4.4)

Daten-Integration, Daten-Governance und Sicherheit

Die Big-Data-Denkweise impliziert einen neuen Umgang mit Daten und eine Neudefinition der Daten-Integration. Es findet ein Wandel vom »Extract-Transform-Load« zum »Extract-Load-Transform« statt. (Abschnitt 4.5)

Bei den Themen Daten-Governance und Sicherheit gibt es beim Übergang von BI zu Big Data zahlreiche neue Aspekte. Gerade in diesem Bereich dürfen keine Versäumnisse zugelassen werden. (Abschnitt 4.6).

Big-Data-Einsatzszenarien und -Lösungsarchitekturen

In konkreten Einsatzszenarien bestimmen in der Regel der Datentyp sowie die Anforderungen an die Verarbeitung die Auswahl der Bausteine in der Zielarchitektur. Daher orientieren die Zielarchitekturen an den Datentypen der verschiedenen Einsatz-Szenarien: Clickstream-Daten, Social-Media-Stimmungsdaten, Server-Logdaten, Sensordaten, Maschine-zu-Maschine-Kommunikation, Standortdaten und Freitext-Daten.

In der Unternehmenspraxis führt der Weg Big Data nicht selten über den Ausbau von Aktivitäten im Business Intelligence und Hybrid-Architekturen. (Kapitel 5)

Synergien zwischen Big Data, Cloud Computing, In-Memory Computing, Open Source

Als Basis für die Darstellung von Big-Data-Technologien in diesem Leitfaden leistet die entwickelte Taxonomie gute Dienste. Es gibt jedoch wichtige Entwicklungen, die eine ganze Reihe von Komponenten aus dem Baukasten betreffen – dazu gehören Cloud Computing, In-Memory Computing und Open Source.

Die Cloud bietet eine Vielzahl von Werkzeugen, um die Nutzung von Big Data zu vereinfachen, zu beschleunigen sowie die Kosten und Risiken zu verringern.

In-Memory Computing umfasst eine Anzahl von Technologien, die sich in unterschiedlichen Big-Data-Komponenten wiederfinden. Durch die Verlagerung der Datenverarbeitung von der Festplatte in den Rechner-Hauptspeicher (In-Memory) können Big-Data-Visualisierungen, -Analysen oder-Transaktionen massiv beschleunigt werden. Somit kann der geschäftliche Mehrwert schneller erbracht werden.

Unternehmen sollten sich außerdem gut überlegen, wo und wie sie Open-Source-Technologie in ihrer Big-Data-Strategie einsetzen wollen; ignorieren sollten sie Open Source nicht.
Big Data befindet sich in stürmischer Entwicklung. Es ist für Unternehmen empfehlenswert, sich über die Fortschritte in der Forschung auf dem Laufenden zu halten. (Kapitel 6)

Management der Big-Data-Risiken

Die Risiken, die Big-Data-Projekte mit sich bringen, sind nicht zu vernachlässigen. Mitunter handelt es sich um neue Risiken. Deshalb ist es wichtig ist, sich die Risiken und Gefahren bewusst zu machen. (Kapitel 7)

Rechtliche Anforderungen umsetzen

Eine besondere Herausforderung in Big-Data-Projekten stellen personenbezogene Daten dar. Im zweiten Big-Data-Leitfaden des BITKOM wurden dazu Möglichkeiten aufgezeigt.[1]

Technologieexperten, Rechts- und Organisationswissenschaftler stellen gemeinsam Überlegungen an, wie Big-Data-Analysen rechtskonform durchgeführt werden können (Kapitel 8).

Es gibt bereits Ansätze, mit denen Garantien über den Datenschutz in die Datenanalyse integriert werden können.

[1] Vgl. »Management von Big-Data-Projekten«, Leitfaden des BITKOM, Juni 2013

Eine wichtige Frage im Zusammenhang mit der Verwertung persönlicher digitaler Daten ist noch Gegenstand der Forschung: Wie kann ein Modell zur Verwertung persönlicher digitaler Daten aussehen, das Dateninhaber, Datenverwerter sowie Dienstleister der Datensicherung, Datenaufbereitung sowie des Datenschutzes zusammenführt? Als eine mögliche Antwort auf die Herausforderungen im Umgang mit persönlichen digitalen Daten wird die Etablierung einer Deutschen Daten-Treuhand vorgestellt. Diskutiert werden auch Konzepte, durch Rollenverteilung den Personenbezug von Daten zu vermeiden. Von großem Interesse für Big-Data-Projekte sind auch Erfahrungen bei der Implementierung von Open-Data-Ansätzen.

Herausforderungen im Betrieb von Big-Data-Lösungen

Im Betrieb geht es darum, Big-Data-Lösungen effizient und zuverlässig zu installieren, verwalten, erweitern und verändern. Es gilt, das Zusammenspiel der verschiedenen Technologien über alle Ebenen einer Lösung hinweg – von der physischen Infrastruktur, über die Daten-Haltung und -bereitstellung, die analytische Verarbeitung, die Visualisierung und die Daten-Integration bis hin zur Governance und Daten-Sicherheit – zu beherrschen. (Kapitel 9)

Neue Qualifikationsprofile zügig herausbilden

Bei der Adaption von Big-Data-Technologien und deren betriebswirtschaftlichem Einsatz wird die Verfügbarkeit von ausgebildeten Kräften mit Data-Science-Kenntnissen eine entscheidende Rolle spielen.

Wissen aus Analytik, IT und dem jeweiligen Fachbereich ist gefragt. Bislang gibt es nur wenige Fachkräfte, die diese Kompetenzen kombinieren. Solche Data Scientists werden von Unternehmen dringend gesucht. Sie vereinen die Rollen als Impuls- und Ideengeber, Lösungsarchitekt, Umsetzer, Analyst, Kommunikator und Überzeuger. Es existieren bereits einige Schulungskonzepte für Data Scientists. (Kapitel 10)

Ergänzende Informationen

Ein Glossar und ein Sachwortregister helfen bei der Orientierung in diesem Leitfaden. Weitere Hilfestellung können Interessierte u.a. bei den Unternehmen und Organisationen erhalten, deren Experten diesen Leitfaden erarbeitet haben.

2 Einleitung

Big Data basiert nicht auf einer singulären Technologie, sondern ist vielmehr das Resultat des Zusammenwirkens einer ganzen Reihe von Innovationen in verschiedenen Gebieten. Insgesamt erlauben diese Fortschritte, aus immer mehr Daten einen immer höheren betriebswirtschaftlichen Nutzen zu ziehen. Je nach Anwendungsszenario können hierbei verschiedene Technologiekonzepte zum Einsatz kommen.

Der vom Wettbewerb ausgehende Druck auf Unternehmen, schnell rapide zunehmende Datenmengen zu verarbeiten, beschleunigt sich immer mehr. Dafür reichen klassische Technologien wie traditionelle Datenbanken, Data Warehouse oder Reporting nicht mehr aus. Heute gilt es, sehr viel mehr Informationen über den Markt und die Kunden zu sammeln und auszuwerten, um weiterhin einen Wettbewerbsvorteil zu erarbeiten.

Für die Unternehmen geht es bei Big Data nicht nur um die Verbesserung bestehender Produkte oder Prozesse – viele sehen die Umsatz-Relevanz von Big Data für neue Geschäftsfelder. Analyseergebnisse werden zu neuen Produkten führen, die wiederum neue Vertriebs- und Geschäftsmodelle mit sich bringen werden.

Bei der Adaption von Big-Data-Technologien und deren betriebswirtschaftlichem Einsatz wird die Verfügbarkeit von ausgebildeten Kräften mit Data-Science-Kenntnissen eine entscheidende Rolle spielen.

Daten werden für viele Branchen auch außerhalb der Informationswirtschaft zu einem Differenzierungsmerkmal und Asset werden. Für Unternehmen wird es in Kürze normal sein, Big-Data-Analysen zur Unterstützung ihrer Entscheidungsprozesse einzusetzen.

Unternehmen nutzen im Schnitt nur etwa 12 % ihrer Daten für betriebswirtschaftlich relevante Analysen.[2] Diese Analysen wiederum beschränken sich meist auf die Auswertung historischer Daten transaktionaler Systeme[3].

Big Data bietet Unternehmen die Möglichkeit, mehr Daten zu nutzen und auch mehr Informationen aus diesen Daten für Entscheidungsprozesse zu gewinnen. Ermöglicht und vorangetrieben wird dieser Fortschritt durch eine Reihe innovativer Technologien sowie die Kombination existierender Technologien mit neuen Marktanforderungen und Trends auf der Anwenderseite.

2.1 Trends bei den Anbietern

Hadoop und In-Memory-Computing

Eine der größten Herausforderungen im Rahmen von Big Data ist die Speicherung und Verarbeitung riesiger Datenmengen zu überschaubaren Kosten. Die wichtigste Innovation in diesem Umfeld ist sicherlich Hadoop – eine neue skalierbare Technologie, mit der sich die Kosten für die Speicherung und Verarbeitung von Daten um etwa 95% im Vergleich zu traditionellen Speicher- und Data-Warehousing-Lösungen verringern lassen. Konkret ist dies

[2] Forrester Research BI/Big Data Survey Q3, 2012
[3] z.B. Finanzbuchhaltung, Auftragsbearbeitung, Beschaffung

die Fähigkeit des Hadoop Distributed File Systems (HDSF), Daten verteilt zu speichern, sowie von MapReduce, Daten parallel zu verarbeiten – alles Open-Source-Software, die auf allgemein verfügbarer Hardware läuft.[4]

Parallel dazu hat Moore's Law[5] zu drastischen Preis-Performance-Verbesserungen in der Geschwindigkeit der traditionellen Datenspeicherung und -verarbeitung geführt. In diesem Zusammenhang stellen In-Memory-Lösungen, bei denen die Daten-Haltung und verarbeitung komplett im Arbeitsspeicher stattfindet, eine inzwischen erschwingliche Alternative für Szenarien mit hohen Performanz-Ansprüchen dar.[6]

In-Memory-Technologien und Hadoop werden verstärkt ausgebaut und genutzt werden – beide Innovationen sind auf dem Weg von Nischentechnologien zum Mainstream.

Mobile Endgeräte und Internet of Things

Ein weiterer wichtiger Trend für die verbreitete Nutzung von Big-Data-Szenarien sind die fortschreitenden technischen Möglichkeiten mobiler Endgeräte und deren Verknüpfung. Das Internet der Dinge kommt mit großen Schritten auf uns zu. Die Verbreitung von mobilen Endgeräten und deren Nutzung wird weiter zunehmen. Auf der einen Seite führt dies zu einer Explosion zusätzlicher Datenquellen, die z. B. ortsspezifische Daten sammeln und weitergeben[7], zum anderen aber auch Daten für den mobilen Einsatz verarbeiten und zur Verfügung stellen – mit immer größerer Effizienz.[8] Viele Hersteller bieten heute Datenanalysen auf mobilen Endgeräten an, wesentlich weniger Hersteller reizen derzeit die Möglichkeiten ortsspezifischer Analysen[9] aus.

Doch nicht nur mobile Endgeräte tragen zur wachsenden Datenflut und vernetzung bei. Immer mehr Produkte und Produktionsanlagen erzeugen, versenden und analysieren Daten von der vernetzten Pumpe bis hin zu intelligenten Windkraftanlagen. Big Data verändert den industriellen Sektor und treibt die Industrialisierung 4.0 voran.[10]

Text Analyse und Prädiktive Analytik

Andere, neue Datenquellen wie z. B. Soziale Netzwerke sind für eine ganze Reihe von Big-Data-Szenarien von großer Wichtigkeit. Moderne Textanalyse erlaubt es, aus den Datenströmen sozialer Netzwerke relevante Informationen wie z. B. Meinungsbilder zu Produkten herauszufiltern und zu verarbeiten. Andere Analysetechniken für die Planung und Vorhersage von Ereignissen wie z. B. Predictive Analytics haben sich so weiter entwickelt, dass sie auch von Anwendern ohne tiefe mathematisch-statistische Kenntnisse eingesetzt werden können.

Vielfalt von Technologiekonzepten

Die Liste wichtiger technischer Innovationen und Verbesserungen, die die Verbreitung von Big-Data-Lösungen vorantreiben, ließe sich weiter fortsetzen. Big Data basiert nicht auf »der einen« Technologie, sondern ist vielmehr das Resultat des Zusammenwirkens einer ganzen Reihe von Innovationen in verschiedenen Gebieten. Insgesamt erlauben diese Fortschritte, aus immer mehr Daten einen immer höheren betriebswirtschaftlichen Nutzen zu ziehen. Je nach Anwendungsszenario können hierbei verschiedene Technologiekonzepte zum Einsatz kommen.

---

4 Vgl. hierzu den Unterabschnitt 4.1.1

5 Dieses empirische Gesetz besagt, dass sich die Performanz unserer IT-Systeme alle zwei Jahre verdoppelt – und das bei gleichbleibendem Preis.

6 Vgl. hierzu den Abschnitt 6.2

7 inklusive RFID und anderer Sensoren

8 Apple's M7 Chip illustriert, wie das Smart Phone die Basis für die persönliche Daten-Explosion wird.

9 GIS – Geo Information Systems

10 Weitere Informationen zum Thema Industrie 4.0 finden sich auf der Seite der BITKOM Arbeitsgruppe Industrie 4.0: http://www.bitkom.org/de/themen/74733.aspx

2.2 Trends bei den Anwendern

Obwohl Big Data erst durch eine Reihe verschiedener Technologien ermöglicht wird, steht natürlich der betriebswirtschaftliche Nutzen im Vordergrund. Der vom Wettbewerb ausgehende Druck auf Unternehmen, schnell rapide zunehmende Datenmengen zu verarbeiten, beschleunigt sich immer mehr. Dafür reichen eben klassische Technologien wie traditionelle Datenbanken, Data Warehouse oder Reporting nicht mehr aus. Heute gilt es, sehr viel mehr Informationen über den Markt und den Kunden zu sammeln und auszuwerten, um weiterhin einen Wettbewerbsvorteil zu erarbeiten. Kunden wollen nicht mehr als anonyme Profile, sondern als Persönlichkeiten mit individuellen Interessen und Bedürfnissen behandelt werden. Dafür müssen Unternehmen jedoch sehr viel mehr Informationen über ihre Kunden verarbeiten als bisher. Und nicht nur über Kunden stehen heute sehr viel mehr Daten zur Verfügung. Produkte, Anlagen oder Prozesse erzeugen immer mehr Daten, die für eine Optimierung genutzt werden können.

So spielt etwa in Banken die statistische Analyse von Kreditausfallrisiken eine wichtige Rolle bei der Bewertung von Krediten und dient dabei zunehmend zur Ableitung differenzierter Preismodelle. Im Versicherungswesen werden Schadenshäufigkeiten und Schadenssummen durch statistische Verteilungen nachgebildet und zur Grundlage von Tarifierungsmodellen gemacht. In der Industrie sind statistische Fragestellungen sehr häufig in der Qualitätssicherung zu finden. Regressionsmodelle helfen beispielsweise, Ursachen für Probleme einzugrenzen und wichtige Einflussquellen zu identifizieren. Simulationsverfahren für Warteschlangen-Probleme und Verfahren zur optimalen zeitlichen Planung von Ressourcen kommen im Projektmanagement zum Einsatz.

Big Data stimuliert neue Geschäftsmodelle

Es geht jedoch nicht nur um die Verbesserung bestehender Produkte oder Prozesse. Viele Firmen sehen die Umsatz-Relevanz von Big Data für neue Geschäftsfelder. Analyseergebnisse werden zu neuen Produkten[11] führen, die wiederum neue Vertriebs- und Geschäftsmodelle mit sich bringen werden. Das neue Marktsegment in der Informationswirtschaft wird ein Betätigungsfeld für neue Unternehmen, die Daten handeln oder anreichern. Es werden neue Geschäftsanwendungen und prozesse implementiert, die zu deutlichen Geschwindigkeits- und damit Wettbewerbsvorteilen führen.

Das Angebot an individualisierten und Echtzeit-Produkten wird zunehmen – in den verschiedenen Industrien mit unterschiedlicher Geschwindigkeit.

Wer die neuen Big-Data-Technologien bestmöglich nutzen will, sollte sich auf mathematisch-statistisches Know-how zur korrekten Datenmodellierung stützen können; dieses Wissensgebiet wird als Data Science (vgl. Kapitel 10) bezeichnet.

Wie bereits erwähnt, bemühen sich Hersteller, die Anforderungen in diesem Bereich durch vorpaketierte Lösungen zu reduzieren, können die Data Scientists jedoch nicht gänzlich ersetzen. Daher wird die Verfügbarkeit von ausgebildeten Kräften mit Data-Science-Kenntnissen im Markt eine entscheidende Rolle bei der Adaption von Big-Data-Technologien und deren betriebswirtschaftlichem Einsatz spielen.

[11] Dieser Aspekt wird im folgenden BITKOM-Leitfaden vertieft, der Ende 2014 erscheinen wird.

2.3 Schlussfolgerungen für die deutsche Wirtschaft und die öffentliche Verwaltung

Zweifellos gehört Big Data zu den wichtigsten Wachstumstreibern – sowohl für die IT-Industrie als auch im Anwendungsbereich vieler Industrien. So bearbeiten alle führenden Anbieter von Unternehmenssoftware dieses Marktsegment mit hoher Priorität[12]. Durch die Anwendung von Big-Data-Technologien können in der Wirtschaft viele neue Einsatzgebiete für IT erschlossen und neue Märkte geschaffen werden[13].

Daten werden auch für viele Branchen[14] außerhalb der Informationswirtschaft zu einem Differenzierungsmerkmal und Asset werden. Für Unternehmen ist es heute selbstverständlich, die Möglichkeiten von Internet und Smartphone zur Unterstützung von Geschäftsprozessen zu nutzen. Vollkommen analog wird es in Kürze normal sein, Big-Data-Analysen zur Unterstützung von Entscheidungsprozessen einzusetzen. So wird der Einsatz von Big Data in vielen Industrien zu einem entscheidenden Erfolgsfaktor werden, und Unternehmen ohne entsprechende Kunden-, Produkt-, oder Prozessinformationen drohen Nachteile in Wettbewerbsfähigkeit.

Big Data in der öffentlichen Verwaltung

Ganz analog zu Unternehmen, die mit Hilfe von Big Data ihre Produkte und Dienstleistungen für Kunden verbessern können, gilt das Gleiche für die öffentliche Verwaltung und deren Dienstleistungen für Bürger. Bessere Informationen (und Vorhersagen) über Bürger und deren Verhalten können Kommunen helfen,

- den Verkehrsfluß zu verbessern (z.B. durch Optimierung von Fahrplänen oder Ampelsystemen),
- die öffentliche Sicherheit zu verbessern (z.B. durch optimierte Einsatzpläne für Polizeikräfte) oder
- Verwaltungprozesse zu beschleunigen (z.B. Unterstützung von Genehmigungsverfahren durch automatische Erkennung von Betrugsmustern).

Wie in der Industrie wird auch in der öffentlichen Verwaltung Big Data eine zunehmend wichtige Rolle spielen.

Big Data und die Politik

Die Akzeptanz und Nutzung von Big Data steht und fällt mit den gesetzlichen Rahmenbedingungen, die diese Nutzung regeln. Aufgeschreckt durch immer neue Medienberichte über Datenskandale schwanken Konsumenten und Bürger zwischen der Angst vor Missbrauch ihrer persönlichen Daten und den Annehmlichkeiten individuell zugeschnittener Angebote und Dienstleistungen durch Unternehmen und Verwaltungen. Hier muss die Politik die entsprechenden gesetzlichen Regelungen vorgeben, die klarstellen, wer welche Daten wann und zu welchem Zweck verwenden kann und in welchen Fällen der Kunde bzw. Bürger über bestimmte Verwendungszwecke informiert bzw. sein Einverständnis eingeholt werden muss. Diese Regelungen dienen nicht nur dem Schutz der Privatsphäre der Kunden bzw. Bürger, sie geben auch der Industrie die Investitionssicherheit, Big-Data-Technologien zu implementieren und neue, innovative Geschäftsmodelle voran zu treiben. Fehlende klare gesetzliche Regelungen können die Ausbreitung von Big Data in Deutschland stark hemmen und die Wettbewerbsfähigkeit der deutschen Industrie negativ beeinflussen. Hier ist die Politik dringend gefragt, ihren Teil zum Erfolg von Big Data in Deutschland beizutragen.[15]

---

[12] Pressemitteilungen von EMC, IBM, Microsoft, Oracle, SAP, Software AG, Teradata, und anderen.

[13] Punktuell treten allerdings auch Kannibalisierungs-Effekte auf.

[14] In der Automobilwirtschaft ist das bereits klar erkennbar.

[15] Siehe auch den Blog von Forrester Research ‚Big Data And The German Dilemma' (http://blogs.forrester.com/holger_kisker/13-02-18-big_data_and_the_german_dilemma)

3 Technologieansätze im Big-Data-Umfeld

Im Kapitel 3 wird gezeigt, dass je nach Anwendungsszenario verschiedene Architekturen oder auch Kombinationen von Architekturen die zielführende Lösung bilden können – hier sind Hadoop und In-Memory nur zwei Komponenten. Zunächst wird eine grobe Segmentierung der Technologielandschaft vorgenommen, um zu zeigen, warum unterschiedliche Technologien zum Einsatz kommen. Anschließend wird eine generelle Taxonomie der Big-Data-Technologien eingeführt. Dieser Technologie-Baukasten bildet den wichtigsten Bezugspunkt für den gesamten Leitfaden.
Als konstruktiv nutzbare Vorlage für Konzeption und Entwurf einer Big-Data-Anwendung eignet sich die Lambda-Architektur. Die in dieser Architektur vorgesehene Modularisierung spiegelt typische Anforderungen an Big-Data-Anwendungen wider und systematisiert sie.

3.1 Big-Data-Technologien – vereinfachte Segmentierung

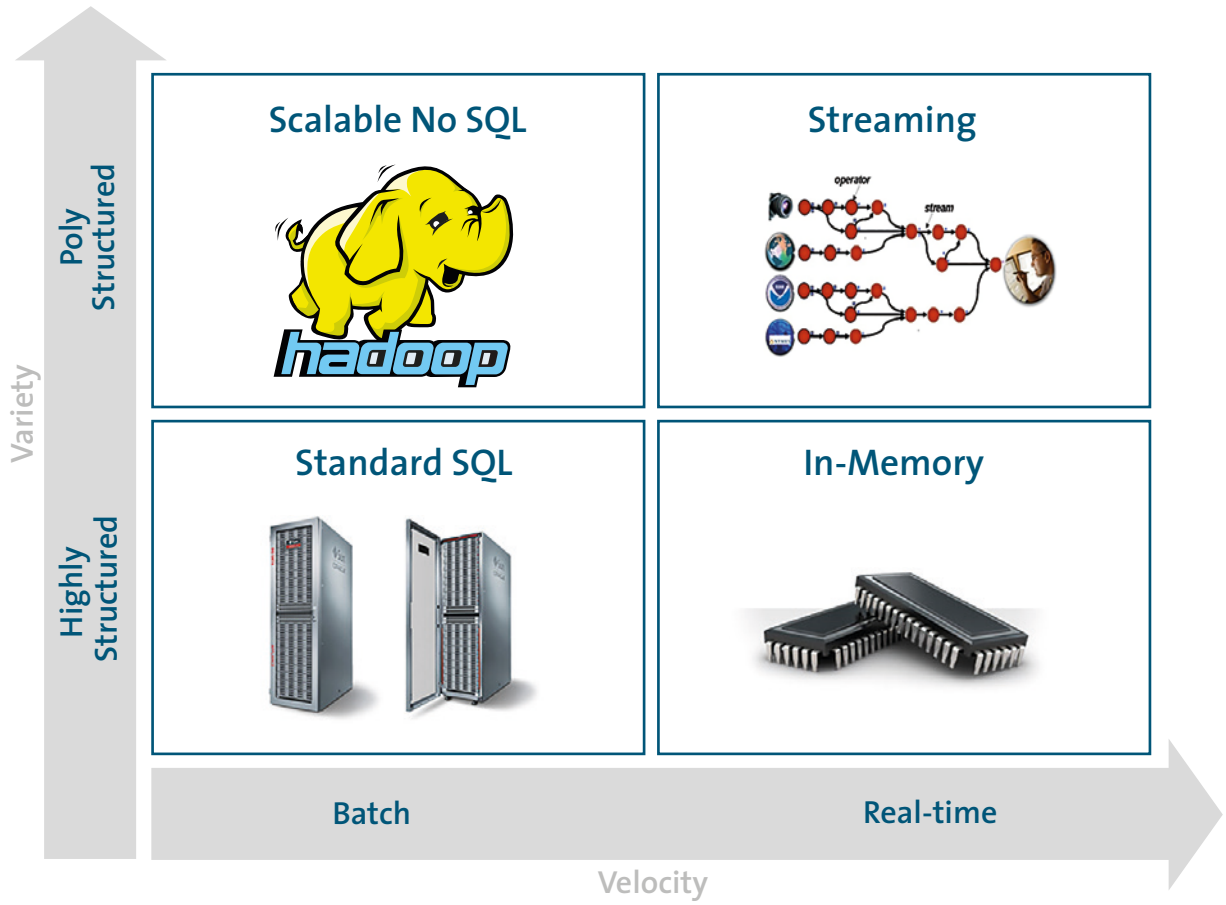


Abbildung 1: Big-Data-Anforderungen und Technologieansätze[16]

[16] Quelle: Forrester Webinar (Sept 19, 2013): Big Data: Gold Rush Or Illusion?

Entsprechend der BITKOM-Definition von Big Data[17] können Technologien zur Lösung verschiedener Big-Data-Anwendungsszenarien mit unterschiedlichen Herausforderungen genutzt werden. Die Herausforderungen können hierbei sowohl

- im Bereich großer Datenmengen (Volume),
- einer großen Datenvielfalt (Variety) oder
- einer hohen Geschwindigkeit der Datengenerierung oder -verarbeitung (Velocity)

liegen.

Auf den vierten Aspekt der BITKOM-Definition – die Daten-Analyse, die einen Mehrwert (Value) aus den Daten ziehen soll – wird im Detail im Abschnitt 4.3 eingegangen. An dieser Stelle werden zunächst die Technologien für die Daten-Haltung und den –Zugriff betrachtet, auf die dann Analyse-Tools aufsetzen können.
Je nach Anwendungsszenario kann eine Dimension – oder auch eine Kombination dieser Dimensionen – dazu führen, dass traditionelle Verfahren für Datenmanagement und –analyse nicht mehr greifen und Big-Data-Technologien eingesetzt werden müssen.
Geht man davon aus, dass große Datenmengen (Volume) zumeist eine der Herausforderungen von Big-Data-Anwendungsszenarien sind, so kann man die Wahl der passenden Technologie auf die Dimensionen

- Datenvielfalt (Variety) und
- Geschwindigkeit (Velocity)

reduzieren.

Die Abbildung 1 zeigt ein vereinfachtes Modell zur Bestimmung der passenden Datenmanagement-Technologie in Abhängigkeit von den Anforderungen in den beiden Dimensionen Variety und Velocity. Die Abbildung verdeutlicht, dass sich grundsätzlich nicht »die eine« Big-Data-Technologie empfiehlt. Vielmehr kann eine ganze Reihe relevanter Technologien (vgl. Tabelle 1) jeweils einzeln oder auch in Kombination zum Einsatz kommen.

Kerntechnologie	Erläuterung
Standard SQL bietet oft kostengünstige Lösungen.	Falls traditionelle Data-Warehouse- und Datenanalyse-Techniken nicht ausreichen, die Anforderungen bezüglich Variety und Velocity jedoch nicht sehr hoch sind[18] und eine Beschleunigung der Datenauswertung mit einem Faktor 10-100+ ausreicht, dann bieten standardisierte Analytics Appliances[19] verschiedener Markthersteller eine gute Lösung.
In-Memory bietet Innovation durch Top-Geschwindigkeit (vgl. Abschnitt 6.2).	Falls eine Beschleunigung der Datenauswertung mit Faktor 100 nicht ausreicht und ein Faktor von 1000 oder weit mehr benötigt wird, dann bieten Lösungen mit In-Memory-Technologie den richtigen Ansatz. So können Datenanalysen, die ansonsten viele Stunden dauern, innerhalb von Sekunden ablaufen und z. B. wöchentliche Planungsprozesse je nach Bedarf zu Ad-hoc-Prozessen zu jedem Zeitpunkt werden.
Hadoop ist das Arbeitspferd für alle Daten (vgl. Unterabschnitt 4.1.1).	Falls die größte Herausforderung eines Anwendungsszenarios in der Vielfalt der Datenformate liegt, dann wird zur Lösung eine Technologie benötigt, die alle Formate gleichermaßen verarbeitet und beliebig skaliert. Hadoop ist eine Open-Source-Technologie zur Speicherung und Verarbeitung sehr großer Datenmengen in allenDatenformaten. Basierend auf Hadoop sind vielfältige Lösungen[20] verfügbar, und eine ganze Reihe von IT-Dienstleistern bietet Unternehmen Unterstützung beim Einsatz von Hadoop an.
Streaming bietet Echtzeit-Lösungen (vgl. Unterabschnitt 4.2.2).	Falls Daten in dem Moment erfasst und ausgewertet werden sollen, in dem sie anfallen, kommen Complex-Event-Processing-Lösungen (CEP) zum Einsatz.

Tabelle 1: Bestimmung eines problemadäquaten Big-Data-Technologieansatzes

[17] Vgl.: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Leitfaden des BITKOM, Berlin 2012, S. 2
[18] weil man z. B. nur mit strukturierten Daten arbeitet
[19] Kombination aus Datenbank-Hardware und Datenanalyse-Software
[20] Für einige Anwendungsszenarien ist Hadoop allerdings beim Datenzugriff und auslesen nicht schnell genug.

Das Modell in Abbildung 1 stellt eine bewusste Vereinfachung der Realität dar. Die Anwendungsbereiche der vier verschiedenen Technologien sind keinesfalls scharf voneinander getrennt; oftmals bildet eine Kombination von Technologien die passende Lösung. Außerdem findet Hadoop zunehmend in allen ver Quadranten Anwendung. So bringt zum Beispiel die Erweiterung von Hadoop durch Storm und Spark im Laufe des Jahres 2014 auch Streaming- und In-Memory-Fähigkeiten in die Hadoop-Plattform.

3.2 Taxonomie der Big-Data-Technologien

Der Zweck jeder Big-Data-Lösung ist es, Daten in entscheidungsrelevante Informationen umzuwandeln. Die Vielfalt an Datentypen und Big-Data-Einsatz-Szenarien erfordert auch vielfältige Werkzeuge auf jeder Schicht einer Technologie-Landschaft. Der Abschnitt 3.2 setzt den Startpunkt für die Ausdifferenzierung der im Abschnitt 3.1 vorgestellten Kerntechnologien.

- Die in Abbildung 2 dargestellten Technologien zur Daten-Haltung werden im Abschnitt 4.1 ausführlich beschrieben; sie unterscheiden sich zum Beispiel nach dem Datenformat, dem Zweck der Daten-Haltung, der Performance sowie den Anforderungen an die Skalierbarkeit.

- Diese Vielfalt setzt sich auch bei den Technologien zum Daten-Zugriff fort. Sie sind Gegenstand des Abschnitts 4.2 und ermöglichen zum Beispiel sowohl die Stapelverarbeitung als auch Echtzeit-Verfahren sowie ein iteratives Entdecken der Daten (Unterabschnitt 4.2.3).

- Auch für die analytische Verarbeitung stellt der Leitfaden die relevanten Werkzeuge vor, welche sich zum großen Teil direkt am Einsatzszenario sowie am Datentyp orientieren (vgl. Abschnitt 4.3).

- Ferner müssen die Ergebnisse zielgruppengerecht präsentiert werden. Das leisten die Visualisierungstechnologien, die im Abschnitt 4.4 erläutert werden.

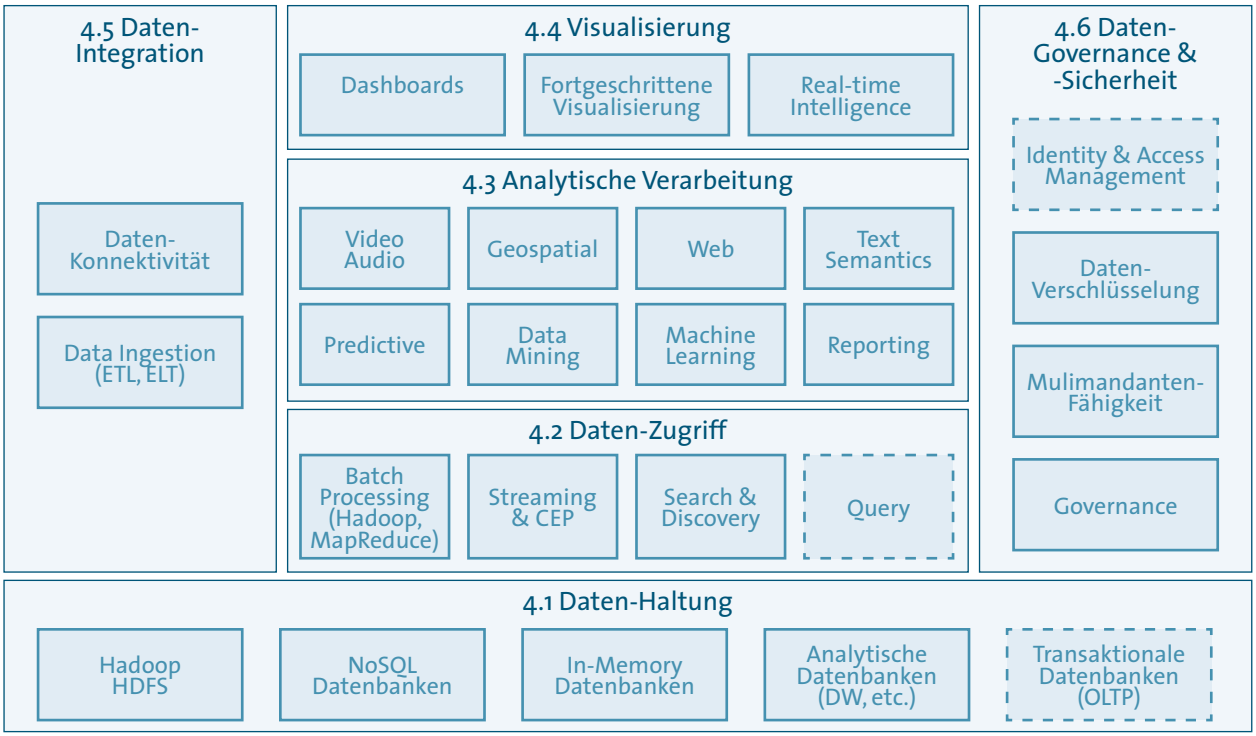


Abbildung 2: Taxonomie von Big-Data-Technologien

Somit markieren die Schichten 4.1 bis 4.4. in der Abbildung 2 den direkten Weg von von den Rohdaten hin zu geschäftsrelevanten Erkenntnissen. Dieser Weg wird flankiert von Daten-Integration (Abschnitt 4.5) und Daten-Governance sowie Daten-Sicherheit (Abschnitt 4.6). Diese flankierenden Bereiche garantieren, dass sich der Weg von den Rohdaten zur Erkenntnis in existierende Standards großer Unternehmen einbettet und sich ein harmonisches Zusammenspiel von Big Data mit existierenden Technologien, Prozessen und Compliance-Vorgaben einstellt.

Die Abbildung 2 bietet eine Taxonomie der Technologien an, die für eine Big-Data-Komplettlösung benötigt werden. Dargestellt sind kommerziell relevante Bausteine[21] für den Big-Data-Einsatz in Unternehmen – ohne Anspruch auf Vollständigkeit. Die Abbildung 2 ist somit als modularer Technologie-Baukasten zu verstehen, nicht jedoch als präskriptive Referenzarchitektur.

In der Tabelle 2 werden die Technologie-Bausteine aus Abbildung 2 kurz und im Kapitel 4 vertiefend vorgestellt.

Schicht	Komponente	Erläuterung
Daten-Haltung	Hadoop Distributed File System	Verteilte Datenspeicherung, welche sich durch Skalierbarkeit von wenigen Terabyte bis hin zu mehr als Hundert Petabyte auszeichnet. HDFS ist die Software, welche Hochverfügbarkeit und Redundanz der Daten sicherstellt. Als physisches Speicher-Medium werden meist kostengünstige Server verwendet.
	NoSQL-Datenbanken	NoSQL[22]-Datenbanken[23] stellen eine neue Art von Datenbanktechnologien dar, um Daten im Format von Dokumenten, Graphen, Key-Value-Paaren oder Spalten hochskalierbar und nicht-relational zu speichern und zu verarbeiten. Dort, wo Daten nicht einem relationalen Datenmodell entsprechen, spielen NoSQL Datenbanken mit Ihren flexiblen Datenmodellen eine wichtige Rolle.
	In-Memory-Datenbanken	In-Memory-Datenbanken[24] ermöglichen den Zugriff auf Informationen in Echtzeit. Hochwertige Daten in einer Größenordnung von bis zu mehreren Hundert Terabyte können aus dem langsamen Festplattenspeicher in den Hauptspeicher (RAM/ Memory) verlagert werden. Damit erfolgt der Zugriff um Zehnerpotenzen schneller als bei Daten auf Festplatten. Erst seit jüngster Zeit reizen Applikationen die Vorzüge von In-Memory-Datenbanken vollständig aus, was insbesondere durch fallende Kosten für RAM begründet ist. Durch ihre verteilte Systemarchitektur bieten In-Memory Data Grids eine ausfallsichere Plattform für wachsende Daten- und Verarbeitungsanforderungen im Bereich Big Data.
	Analytische Datenbanken	Analytische Datenbanken (oft als Data Warehouse bezeichnet) gehören zur Klasse der relationalen Datenbanken. Sie sind für das Einsatz-Szenario OLAP[25] optimiert, welches sich durch moderat große Datenmengen[26], umfangreiche Unterstützung der Abfragesprache SQL sowie einer kleinen bis moderaten Anzahl der Benutzer charakterisiert. Analytische Datenbanken sind nicht neu und werden für Big Data-Projekte oft zusammen mit Hadoop eingesetzt.
	Transaktionale Datenbanken	Transaktionale Datenbanken gehören zur Klasse der relationalen Datenbanken und werden hier nur der Vollständigkeit halber sowie zur Abgrenzung gegenüber den neueren Datenbank-Typen erwähnt. Sie sind für das Einsatz-Szenario OLTP[27] optimiert und ermöglichen die gleichzeitige Arbeit von Hunderttausenden von Nutzern.

[21] Nicht berücksichtigt sind Technologien, die sich im Forschungsstadium befinden und noch keine Marktreife erreicht haben.
[22] not only SQL
[23] Synonym: nicht-relationale Datenbanken.
[24] auch In-Memory Data Grids genannt.
[25] Online Analytical Processing
[26] zum Beispiel 10 Terabyte
[27] Online Transaction Processing. Wichtig sind rasche und verlässliche Operationen zur Einfügung, Löschung und Aktualisierung von Datensätzen.

Schicht	Komponente	Erläuterung
Daten-Zugriff	Batch Processing (MapReduce)	Stapelverarbeitung (Batch Processing) bezeichnet die automatische, sequentielle und vollständige Abarbeitung der in den Eingabedateien enthaltenen Daten. Das Programm läuft dabei nach dem Start vollkommen selbständig ab. Die zu verarbeitenden Daten werden ohne Eingriff des Benutzers der Reihe nach erledigt. Die Ergebnisse können zum Beispiel in Dateien oder Datenbanken abgelegt werden. Apache Hadoop MapReduce ist ein typisches Beispiel für Stapelverarbeitung bei Big Data.
	Streaming Processing und CEP	Das Streaming-Verarbeitungs-Prinzip steht für die kontinuierliche Verarbeitung von Eingangsdaten oder -signalen bei gleichzeitiger kontinuierlicher Bereitstellung von Ergebnisdaten oder -signalen. Eingangsdaten liegen oft als Datenstrom vor[28]. Ebenso werden die Ausgangsdaten oft als Datenstrom gefordert. Diese Fähigkeit wird im CEP genutzt, wo komplexe Regeln die Verarbeitung der Daten steuern.
	Search & Discovery	Diese Kategorie umfasst das Suchen und Entdecken von Informationen in meist unstrukturierten Daten. Hauptziel von Search ist es, neben der genauen Antwort auf eine Frage auch ähnliche, verwandte Antworten vorzuschlagen und damit das Entdecken neuer Informationen und Zusammenhänge zu ermöglichen. Der Kern dieser Suchmaschinen sind Algorithmen, welche Text-Matching durchführen und Indizes bauen, welche dann mit Schlagworten durchsucht werden können.
	Query	Query zielt darauf ab, strukturierte Daten aus verschiedenen Quell-Systemen und mit verschiedenen Formaten sinnvoll zu verdichten und zu befragen. Dazu wird meist die populäre Abfragesprache SQL genutzt.
Analytische Verarbeitung	Audio/Video	Hier geht es um die Analyse multimedialer Inhalte, die Extraktion von Fakten und das Erkennen von Zusammenhängen. Oft werden Audio-Daten in Text konvertiert und dann mit Text-basierten Verfahren analysiert.
	Geospatial	Bei geospatialen Analysen geht es um die Anwendung statistischer und mathematischer Verfahren auf Daten, die einen geografischen oder räumlichen Bezug haben. Diese Daten können mit exakten Global Positioning System (GPS)-Koordinaten versehen sein. Der Bezug auf Ort oder Raum kann auch implizit, indirekt oder ungefähr sein.
	Data Mining	Diese Kategorie umfasst statistische Analyseverfahren und Modelle zum Auffinden von Mustern in großen Datenmengen.
	Predictive	Anders als bei traditionellen rückwärts gerichteten Analysen geht es bei Predictive Analytics darum, Entwicklungen vorher zu sehen und so Entscheidungen möglichst zu optimieren.
	Web	Web Analytics umfasst die Messung, die Erfassung, die Analyse und das Reporting von Daten, die es erlauben, den Internet-Auftritt eines Unternehmens zu optimieren, meist mit dem Ziel mehr Umsatz zu erzielen. Social Media Analytics analysiert die Informationen welche Nutzer online Preis geben, wie zum Beispiel Vorlieben zu bestimmten Produkten, Aktivitäten oder dem Freundeskreis, basiert auf Quellen wie Facebook oder LinkedIn.
	Machine Learning	Maschinelles Lernen umfasst eine Vielzahl von Anwendungen und Methoden, in denen Computerprogramme durch iterative Verfahren ihr Wissen stetig erweitern und somit hinzu lernen – in der Regel durch statistische oder logische Analysen gegebener Daten sowie durch die Anwendung rechenintensiver Algorithmen.

[28] zum Beispiel Echtzeit-Messungen von Sensoren oder anderen Maschinen

Schicht	Komponente	Erläuterung
Analytische Verarbeitung	Text/Semantic	Diese Kategorie umfasst linguistische und semantische Verfahren, mit deren Hilfe aus Texten relevante Informationen extrahiert, Strukturen erkannt und Verknüpfungen der Daten untereinander sowie mit anderen Datenquellen hergestellt werden, um Business Intelligence auf Text zu ermöglichen.
	Reporting	Reports sind systematische (Detail-) Berichte, die eine analytische Aufbereitung, meist in tabellarischer Form, auch als Mischform aus tabellarischen und grafischen Elementen mit Textblöcken, zur Verfügung stellt. In der herkömmlichen Nutzung werden diese häufig gedruckt, bzw. intern via Mail-Anhang verteilt. Quelle sind typischerweise strukturierte Daten, welche zusammengefasst und dann zum Beispiel als Dashboard visualisiert werden. Sie werden wie Dashboards ad-hoc oder periodisch, regelmäßig oder aufgrund spezifischer Anforderung genutzt. Die Kategorie ist nicht neu, sondern hat de-fakto den Business-Intelligence-Markt begründet.
Visualisierung	Dashboards	Ein Dashboard (englisch für Instrumententafel) besteht aus einem Arrangement von mehreren visuellen Bausteinen mit dem Ziel diese zu konsolidieren, dabei nötigenfalls zu verdichten und damit relevante Informationen auf einem Bildschirm im Überblick darzustellen. Gängige Dashboards erlauben eine mehrschichtige Darstellung (Multi-Layer/Linking). Interaktiv kann zwischen einzelnen Schichten navigiert werden und im besten Falle stehen dem Betrachter Selektionen zur Verfügung, die ihn z. B. Zeiträume einschränken oder dargestellte Inhalte variieren lassen.
	Advanced Visualization	Unter Advanced Visualization (fortgeschrittene Visualisierung) versteht man interaktive visuelle Darstellungen, die komplexe Sachverhalte und Zusammenhänge zu verstehen und kommunizieren helfen. Abgrenzend zu einfachen Dashboards beinhalten fortgeschrittene Visualisierungen auch die Möglichkeit, interaktiv Veränderungen an der Darstellung vorzunehmen, um so sukzessive verschiedene Teilfragen zu analysieren. Fortgeschrittene Visualisierungen umfassen zudem fast immer koordinierte Mehrfachansichten, mittels derer zusammenhängende Teilaspekte der visualisierten Daten zeitgleich dargestellt und zueinander in Beziehung gesetzt werden können.
	Real-time Intelligence	Der Begriff bezeichnet die kontinuierliche Echtzeit-Animation bzw. visuelle Analyse auf eingehenden Streaming-Daten. Typischerweise wird dazu ein gleitendes Zeitfenster aus dem Datenstrom extrahiert und grafisch dargestellt.
Daten-Integration	Daten-Konnektivität	Konnektoren sind Technologien mit der Aufgabe, Daten aus unterschiedlichen Systemen zugänglich zu machen. Dies können sowohl klassische Datenbanken oder Anwendungen sein, als auch Middleware-Technologien. Konnektoren abstrahieren die spezifische Beschaffenheit des Quellsystems und stellen Daten und Funktionen über Standardschnittstellen (z. B. SQL, Web Services, XML, JMS Messaging Middleware) einheitlich zur Verfügung.

Schicht	Komponente	Erläuterung
Daten-Integration	Data Ingestion (ETL, ELT)	Ingestion (Aufnahme) hat das Ziel, Daten aus verschiedensten Quellen in eine Big-Data-Lösung zu importieren. Hierbei kann es sich um Echtzeit- (Real-time-) oder Stapel- (Batch-)Daten handeln. In traditionellen Data Warehouses folgt auf diesen Daten-Import ein Umwandeln der Daten in drei Schritten: Extract, Transform, Load (ETL) bezeichnet den Prozess der Aufbereitung von Rohdaten für die nachgelagerte Speicherung dieser Daten in einer analytischen Datenbank bzw. einem Data Warehouse. Rohdaten werden normalisiert, validiert und mit einer Struktur versehen, damit also in ein relationales Datenbank-Schema überführt. Mit Hadoop etabliert sich ein neuer Batch-Prozess mit den Schritten Extract, Load, Transform (ELT). Rohdaten landen ohne Struktur in Hadoop und erst zum Zeitpunkt der Analyses erfolgt eine Transformation. Im Bereich von Echtzeitdaten verarbeiten vorgelagerte Messaging- oder Event-Processing-Systeme die eingehenden Datenströme, die dann zur Weiterverarbeitung an NoSQL-Datenbanken oder Hadoop geleitet werden können.
Daten-Governance und Sicherheit	Identity & Access Management	Identity & Access Management regelt den Zugang zu Daten und beantwortet zwei Fragen: 1. Wer ist der Benutzer? Wichtig ist die Authentifizierung der Identität, welche typischerweise in einem Verzeichnis aller Mitarbeiter[29] hinterlegt ist. 2. Welche Rechte hat der Benutzer? Bei dieser Frage geht es um die Autorisierung zu bestimmten Handlungen. Das Recht, Daten zu lesen oder zu verändern kann zum Beispiel an die Rolle des Benutzers gebunden sein. Identity & Access Management ist nicht neu, wird aber auch in Big-Data-Lösungen eingesetzt.
	Multi-Mandantenfähigkeit	Big-Data-Technologie ist multi-mandantenfähig[30], wenn sie erlaubt, auf demselben Server oder Cluster mehrere Mandanten[31] zu bedienen, ohne daß diese wechselseitig Einblick in ihre Daten, Jobs, analytischen Modelle, Benutzerverwaltung etc. haben. Die Mandanten werden physisch aus einem Infrastruktur-Pool bedient, während logisch[32] eine Isolation der Ressourcen erfolgt.
	Datenverschlüsselung	Personenbezogene und andere besonders schützenswerte Daten werden in der Verschlüsselung so umgewandelt, daß bei einem Verlust der Daten[33] kein Schaden entsteht, da die Informationen ohne den Schlüssel nicht im Klartext lesbar sind. Daten können sowohl für den Transport[34] als auch für die Lagerung verschlüsselt werden.

Tabelle 2: Kurzcharakteristik der Technologie-Komponenten

[29] Active Directory
[30] Auch: mandantentauglich
[31] Kunden oder Unternehmenseinheiten
[32] also auf Software-Ebene
[33] zum Beispiel durch Diebstahl
[34] also zum Beispiel die Übertragung im Weitverkehrsnetz

3.3 Big-Data-Architekturansatz

Big-Data-Anwendungen und verteilte Systeme

Eine Big-Data-Anwendung ist ein verteiltes System, das flexibel an erforderliche Daten-Volumina und Verarbeitungskapazitäten anpassbar ist. Erfordert eine Anwendung ein wachsendes Datenvolumen oder werden Verarbeitungs- und Antwortzeiten zu lang, so kann ein Engpass durch horizontale Skalierung behoben werden, also durch erweiterten Rechnereinsatz und Verteilung. Im Vergleich zu vertikal skalierbaren Systemen kann der Aspekt der Verteilung nicht verborgen bleiben und muss bei Entwurf und Konstruktion einer Anwendung ausdrücklich modelliert werden.

Verteilung bedeutet u.a., dass Daten mittels Replikaten redundant angelegt werden, einerseits zwecks Sicherung und Verfügbarkeit, wenn ein Daten-Knoten ausfällt, andererseits um schnelle, lokale Zugriffe für verteilte Prozesse zu ermöglichen. Bei Änderungen müssen alle Replikate berücksichtigt werden; stimmen Replikate nicht überein, gilt das System als inkonsistent.

Letztendliche Konsistenz

Verfügbarkeit war von Beginn an ein wesentliches Merkmal verteilter Systeme. Verfügbarkeit wird auch dann aufrechterhalten, wenn Kommunikation unsicher wird oder Netz-Knoten ausfallen. Weil aber auch die Daten verteilt sind und zudem Replikate vorliegen können, werden Inkonsistenzen der Daten auftreten und müssen im Ablauf des Systems gehandhabt und beseitigt werden.

Ein verteiltes System kann sogenannte Eventual Consistency sicherstellen: Unter der Bedingung, dass keine Datenänderung erfolgt und Partitionierungen des Netzes sowie Ausfälle von Netz-Knoten nach endlicher Zeit behebbar oder kompensierbar sind, wird letztendlich ein konsistenter Zustand erreicht, d.h. alle Replikate sind identisch und geben die Werte der letzten Datenänderungen wieder.

Forschung und Entwicklung verteilter Systeme haben verschiedene Bedingungen der Inkonsistenz und Verfahren zur Wiederherstellung von Konsistenz definiert, die es ermöglichen, die (In-)Konsistenz eines Systems graduell im Spektrum von »letztendlich konsistent« bis »strikt konsistent« zu wählen. Unter dem Titel »Eventually Consistent«[35] hat W. Vogels die Ergebnisse kompakt zusammengefasst und aus Nutzersicht bewertet. Ein praktische Ergänzung von »eventual consistency« ist beispielsweise die »read your writes consistency«-Bedingung: Das System stellt sicher, dass ein Prozess, der ein Daten-Element ändert, anschließend diesen geänderten Wert und niemals einen vorhergehenden Wert liest.

CAP-Theorem

Für verteilte Daten-Systeme gilt das sogenannte CAP-Theorem[36], formuliert von Eric Brewer[37], bewiesen von Nancy Lynch und Seth Gilberth[38]. Es besagt, dass ein verteiltes Daten-System maximal zwei der Eigenschaften Konsistenz, Verfügbarkeit und Toleranz von Netzwerk-Trennungen erfüllen kann:

- **Konsistenz (Consistency)**
  Alle Replikate von Daten sind identisch (single copy consistency) und zu einem Zeitpunkt sind für alle Knoten die Daten im gleichen Zustand sichtbar.

- **Verfügbarkeit (Availability)**
  Alle Anfragen an das System werden stets beantwortet.

[35] Werner Vogels, »Eventually Consistent«, Communications of the ACM 2009,|voL. 52, no. 1, Doi:10.1145/1435417.1435432
[36] Die Ausführungen zum CAP-Theorem orientieren sich an http://de.wikipedia.org/wiki/CAP-Theorem
[37] Brewer, Eric: »Towards Robust Distributed Systems,« Proc. 19th Ann. ACM Symp.Principles of Distributed Computing (PODC 00), ACM, 2000, pp. 7-10; http://www.cs.berkeley.edu/~brewer/PODC2000.pdf
[38] Gilbert, Seth and Lynch, Nancy: »Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services.« ACM SIGACT News, v. 33 issue 2, 2002, p. 51-59.

- **Partitionstoleranz (Partition Tolerance)**
  Das System arbeitet auch bei Verlust von Nachrichten, bei Ausfall von Netzknoten und Trennungen des Netzes weiter.

Seit der Veröffentlichung führte das CAP-Theorem zu einer breiten Diskussion seiner Anwendbarkeit und wurde zur Klassifikation von Big-Data-Systemen in folgende Kategorien genutzt:

- CA: konsistent und verfügbar, Partitionierung darf nicht auftreten,
- AP: verfügbar auch bei Partitionierung, aber nicht immer konsistent,
- CP: konsistent auch bei Partitionierung, aber nicht immer verfügbar.

Weil Big-Data-Anwendungen als verteilte Systeme hoch verfügbar und tolerant gegenüber Partitionierungen sein müssen, bot das CAP-Theorem Anlass zur Abgrenzung traditioneller Datenbanken (CA) von Big-Data-fähigen NoSQL-Datenbanken (AP oder CP).

Die praktischen Erfahrungen von 12 Jahren nach der Formulierung des CAP-Theorems hat Eric Brewer prägnant zusammengefasst[39]: Ein Big-Data-System muss partitionstolerant sein, und deshalb ist perfekte Konsistenz und Verfügbarkeit ausgeschlossen. Partitionierungen sind aber selten und das moderne CAP-Ziel ist es, in dem durch das CAP-Dreieck (vgl. Abbildung 3) symbolisierten Entwurfsraum die Anforderungen an Konsistenz und Verfügbarkeit gegenseitig abzuwägen, und eine bestmögliche Kombination für die jeweilige Anwendung zu finden.

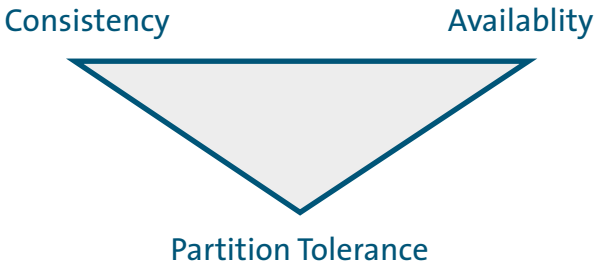


Abbildung 3: CAP-Dreieck

Verfügbarkeit und Konsistenz werden als graduelle Größen verstanden. Verfügbarkeit ist bei der Nutzung eines Systems an Hand der Reaktionszeiten bewertbar; ein Ausfall wird angenommen, wenn eine Reaktion inakzeptabel lange dauert. Schnellere Reaktion geht auf Kosten sichergestellter Konsistenz, die Garantie von strikteren Konsistenz-Anforderungen führt zu verzögerter Reaktion.

Diese Abwägung wird anwendungsspezifisch bei der Konfiguration des Systems und die Wahl der Verfahren zur Datenverwaltung und Konfliktauflösung getroffen. Ein Beispiel soll das verdeutlichen:

Für ein Bestellsystem sind Verfügbarkeit und schnelle Reaktion vorrangige Anforderungen. Deshalb wird das System auf genügend viele Knoten verteilt und die Daten werden in Replikaten gehalten. Bestellungen werden in den Warenkorb, also in die Daten übernommen. Das System antwortet unmittelbar, während unabhängig von der Systemreaktion interne Prozesse die Daten-Änderung auf die Replikate verteilen. Ist ein Benutzer mehrfach am gleichen System aktiv, können Inkonsistenzen zwischen Replikaten auftreten. Eine gängige Strategie ist, bei festgestellter Inkonsistenz die unterschiedlichen Warenkörbe zusammenzufassen. Möglicherweise erscheinen deshalb in einem Warenkorb Dinge erneut, die an anderer Stelle bereits entfernt wurden. Diese Unannehmlichkeit tritt aber nur sehr selten auf, und wird zugunsten einer schnellen Reaktion beim Füllen des Warenkorbes in Kauf genommen.

[39] Brewer, Eric: »CAP Twelve Years Later: How the »Rules« Have Changed«, InfoQ 2012, http://www.infoq.com/articles/cap-twelve-years-later-how-the-rules-have-changed

Leistungen von Big-Data-Komponenten

Big-Data-Komponenten wie Hadoop, NoSQL-Datenbanken und Storm bieten Funktionen zur horizontal skalierbaren Verarbeitung und Speicherung großer Datenbestände an. Eventual Consistency, die Wahl von Konsistenz-Bedingungen und damit verbunden die Performance-Charakteristik und Verfügbarkeit, das Erkennen von Konflikten, Verfahren zur Reparatur von Inkonsistenzen werden als Leistungen angeboten. Gute Tutorials, Beispiele und Anleitungen zur Konfiguration erleichtern die Arbeit und sind auch bei Open-Source-Produkten, insbesondere wenn sie im Portfolio kommerzieller Anbieter liegen, gut zugänglich.

Die im Abschnitt 3.3 skizzierten und durchaus umfangreichen Konzepte sollten Entwicklern von Big-Data-Anwendungen vertraut sein. Es ist auch nützlich, algorithmische Verfahren wie Konsensus-Protokolle zur Sicherung von Konsistenz oder Vector-Clocks zum Aufdecken von Inkonsistenzen zu kennen, um die erforderliche Konfigurierung und Parametrisierung der Komponenten vorzunehmen.

Konstruktion einer Big-Data-Anwendung

Im Rahmen der Entwicklung einer Big-Data-Anwendung wird eine konkrete Anwendungsarchitektur entworfen, in die Komponenten aus den Bereichen der im Abschnitt 3.2 vorgestellten Taxonomie integriert werden. Ein prägendes Merkmal von Big-Data-Anwendungen und ihren Komponenten ist, dass sie durch erweiterten Rechnereinsatz und Verteilung von Daten-Haltung und Prozessor-Leistung horizontal skalierbar sind. Dies wird unterstützt durch eine geeignete Auswahl von Big-Data-Komponenten, die Anpassung der anwendungsspezifischen Algorithmen und die Konfiguration und Abbildung auf Rechnerknoten- und Kommunikations-Infrastruktur.

Am Markt ist mittlerweile eine fast unübersehbare Vielzahl von Komponenten verfügbar (vgl. Abbildung 2 sowie Kapitel 4), die zu Big-Data-Infrastrukturen integriert werden können oder selbst schon als vollständige, konfigurierbare Infrastruktur einer Big-Data-Anwendung einsetzbar sind.

Lambda-Architektur

Eine konstruktiv nutzbare Vorlage für Konzeption und Entwurf einer Big-Data-Anwendung ist der von Nathan Marz und James Warren publizierte Ansatz der Lambda-Architektur[40]. Die in der Architektur vorgesehene Modularisierung spiegelt typische Anforderungen an Big-Data-Anwendungen wider und systematisiert sie.

Auf diese Weise ist der Architektur-Ansatz nützlich, um technische und nicht-funktionale Anforderungen an neue Anwendungen aufzudecken und zu beurteilen, unabhängig davon, in welcher Form und welchem Umfang die Module als technische Komponenten der Anwendung realisiert werden.

Anwendungs-Architektur

In eine Anwendung gehen Datenströme[41] ein, die protokolliert werden und zu einem wachsenden Datenvolumen führen. In der Lambda-Architektur werden original eingestellte Daten ohne Informationsverlust aufgezeichnet. Für neue Funktionen sowie korrigierte und erweiterte Auswertungen steht im Datenspeicher der gesamte Original-Datenbestand zur Verfügung.

[40] [MarzWarren2013] Nathan Marz and James Warren, »Big Data – Principles and best practices of scalable real-time data systems«, Manning Publications, 2013

[41] z. B. Daten aus der Nutzung und Bedienung einer Anwendung, Sensor-Daten technischer Systeme, von mobilen Anwendungen generierte Daten

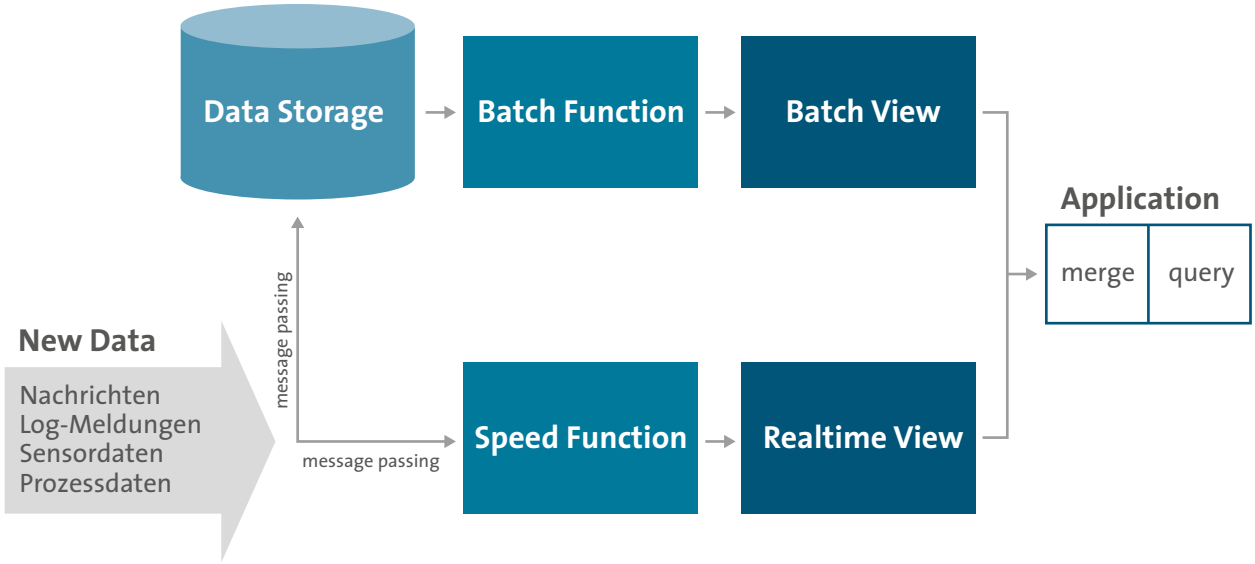


Abbildung 4: Architektur-Komponenten für Big Data

In der Architektur werden zwei Ebenen unterschieden:

- Die Batch-Ebene prozessiert alle gesammelten Original-Daten in der Batch-Funktion und bereitet sie zur Präsentation der Berechnungs-Ergebnisse im Batch View auf. Batch-Prozesse werden zyklisch wiederholt.[42]

- In der Speed-Ebene werden in einkommende Daten unmittelbar und so schnell als möglich prozessiert und für die Präsentation in der Anwendung aufbereitet. Die Speed-Ebene überbrückt die Batch-Laufzeiten und die damit verbundenen Verzögerungen zwischen dem Eintreffen von Daten und deren Berücksichtigung im Batch View.

Eine Übersicht über die Komponenten und ihre Leistungen sowie über die Anforderungen und Aspekte der Skalierung enthält die Tabelle 3.

[42] Dies ist einem ETL-Prozess vergleichbar, der ein Data Warehouse mit Daten füllt, die für Online-Analysen (OLAP) vorbereitet und optimiert sind.

Komponenten	Leistungen	Anforderungen und Aspekte der Skalierung
New data	Eintreffende Daten werden an die Batch- und die Speed-Ebene gesandt, versehen mit einem Zeitstempel und mit einem global eindeutigen Identifikationsmerkmal.	Wie hoch liegt die durchschnittliche Datenrate? Welche Spitzenwerte erreicht der Datenfluss, welche Puffer-Größen sind notwendig?
Data Storage	Die Original-Daten sind unveränderlich gespeichert, neue Daten werden der Original-Datenmenge angefügt.	Partitionierung und Verteilung der Daten zwecks Verarbeitung im MapReduce-Prozess des Batch-Layer. Skalierbarkeit bezüglich der Datenmenge und der redundanten, ausfallsicheren Speicherung.
Batch Function	Die Batch-Funktion berechnet die Informationen für die Batch-Ansicht an Hand des Original-Datenbestandes nach dem Map-Reduce-Verfahren für große Datenbestände.	Der Algorithmus muss auf Basis der verteilten Datenspeicherung parallelisierbar sein. Skalierbarkeit bezüglich der Verarbeitungsleistung.
Batch View	Die Ergebnisse der Batch-Funktion werden in Read-Only-Datenbanken gespeichert und für schnellen Zugriff indiziert. Die Daten sind auch bei verteilter, replizierter Speicherung konsistent. Inkonsistenz oder Nicht-Verfügbarkeit ist auf die Dauer des Umschaltens zwischen den Daten-Versionen beschränkt.	Schneller, möglichst über alle Knoten atomarer Übergang zu einer neuen Version aller Daten im Batch View. Schnelle Reaktionszeiten für anwendungsspezifische Queries. Skalierbarkeit bezüglich der Datenmenge und des Anfrageaufkommens.
Speed Function	Die einkommenden Daten werden so schnell verarbeitet, dass der zeitliche Verzug zwischen Verfügbarkeit der Daten und ihrer Berücksichtigung in der Anwendung vertretbar ist.	Skalierbarkeit bezüglich der ankommenden Daten-Rate und der Verarbeitungsgeschwindigkeit
Real-time View	Die von der Speed-Funktion berechneten Daten werden in einer Lese-Schreib-Datenbank gespeichert. Der notwendige Umfang ist bestimmt durch die Menge anfallender Daten im Datenstrom für die Zeitdauer von zwei Verarbeitungsläufen im Batch-Layer. Ankommende Daten sind spätestens nach dieser Zeit im Batch-View berücksichtigt und können im Real-time View entfernt werden.	Zeiträume der Inkonsistenz bei Aktualisierung der verteilten Daten (eventual consistency) sind zu minimieren. Schnelle Reaktionszeiten für anwendungsspezifische Queries. Skalierbarkeit bezüglich der Datenmenge, der Lese-Rate und bezüglich der Schreib-/Update-Rate der von der Speed Function produzierten Daten
Application	Die Anwendung wickelt Anfragen ab und kombiniert dabei die Daten des Batch View und des Real-time -View.	Zusammenführung und Abgleich der Informationen zwischen dem Batch und dem Real-time View. Skalierbarkeit bezüglich der Nutzung durch gleichzeitige, unabhängige Anfragen.

Tabelle 3: Lambda-Architektur – Komponenten, Leistungen, Anforderungen

Lambda-Architektur in der Praxis

Eine Anwendung muss nicht unter allen Aspekten skalierbar sein. Beispielsweise erfordert ein technischer Leitstand zwar ein umfangreiches, wachsendes Datenvolumen und hohe Verarbeitungsleistung zur Analyse und Aufbereitung gespeicherter und ankommender technischer Datenströme, die Anzahl unabhängiger Arbeitsplätze bleibt aber gering.

Grundsätzlich werden in jeder Iteration des Batch-Prozesses alle Daten verarbeitet. Bei Änderung der Algorithmen ist das im Allgemeinen der Fall. Anwendungsspezifisch können neue Daten auch inkrementell verarbeitet werden.

Die Gewichtung von Batch und Speed Layer und die jeweiligen Funktionen können anwendungsspezifisch variieren.

Der Mehraufwand zum Betrieb einer verteilten Big-Data-Anwendung im Vergleich zu einer nicht skalierbaren, zentralisierten Lösung muss anwendungsspezifisch mit Blick auf Datenvolumen und Verarbeitungsgeschwindigkeit abgeschätzt werden.

4 Relevante Technologie-Komponenten für Big-Data-Anwendungen

Das Kapitel 4 ist für den Leitfaden zentral; es erläutert die Komponenten des in Abbildung 2 vorgestellten Technologie-Baukastens im Detail.

- Technologien der Daten-Haltung werden im Abschnitt 4.1 (S.36 ff.) vorgestellt. Mit Mittelpunkt steht Hadoop – die bisher wichtigste Innovation im Big-Data-Umfeld. Außerdem werden Datenbanken erörtert, die speziell auf die Anforderungen in Big-Data-Situationen zugeschnitten sind.
- Der Daten-Zugriff bildet den Schwerpunkt des Abschnitts 4.2 (ab S. 51) Der Daten-Zugriff beinhaltet verschiedene Technologien, die es sehr unterschiedlichen analytischen Anwendungen ermöglichen, auf den Pool von großen Datenmengen zuzugreifen. Hierin enthalten sind – neben dem schon erwähnten MapReduce – unter anderem Hive, welches Zugang zu Daten auf HDFS über die von den relationalen Datenbanken vertraute Abfragesprache SQL bietet, aber auch Streaming – der Zugang zu Live-Datenströmen. Search- & Discovery- Technologien wiederum ermöglichen das Suchen und Entdecken von Informationen in meist unstrukturierten Daten analog zur Suchmaschine Google.
- Die analytische Verarbeitung bildet den eigentlichen Kern von Big-Data-Anwendungen. Der Abschnitt 4.3 (ab S. 61) umfasst Technologien zur Verarbeitung der verschiedenen Datentypen sowie wichtige Themen wie Predictive Analytics, Data Mining und Maschinelles Lernen.
- Fortgeschrittene Visualisierungen stellen ein mächtiges und hochgradig flexibles Werkzeug im Analyseprozess dar und werden im Abschnitt 4.4 (S. 73ff.) thematisiert.
- Die Big-Data-Denkweise impliziert auch einen neuen Umgang mit Daten und eine Neudefinition der Daten-Integration. Es findet ein Wandel vom »Extract-Transform-Load« zum »Extract-Load-Transform« statt (Abschnitt 4.5, S. 89ff.).
- Das Kapitel schließt mit Betrachtungen über Daten-Governance und Sicherheit (Abschnitt 4.6, S. 96 ff.). Es wird aufgezeichnet, was an Governance für Big Data neu ist. Gerade in diesem Bereich dürfen keine Versäumnisse zugelassen werden.

4.1 Daten-Haltung

Hadoop spielt in vielen Big-Data-Anwendungen eine zentrale Rolle. Von einem preiswerten Daten-Parkplatz hat sich Hadoop so weiter entwickelt, dass es Unternehmensentscheidungen in nahezu Echtzeit durch verschiedenste Analyseverfahren unterstützt. Diese gehen über die Möglichkeiten traditioneller Business Intelligence Tools weit hinaus. Hadoop ist ein Open-Source-Framework zur parallelen Datenverarbeitung auf sehr hoch skalierenden Server-Clustern. Zu diesem Top-Level-Projekt der Apache Software Foundation gehört eine zweistellige Anzahl von Unterprojekten. Hadoop wird im Unterabschnitt 4.1.1 in seiner Gesamtheit vorgestellt, auch wenn sich die Unterprojekte über die Bereiche Daten-Haltung, -Zugriff, -Integration, -Sicherheit und Betrieb erstrecken. Hadoop hat sich für viele Problemstellungen als sehr tragfähig und deshalb

als wesentlicher Motor der Big-Data-Entwicklung erwiesen. Ein breites und innovatives Ökosystem aus Open Source und kommerziellen Produkten liefert Verbesserungen, mit denen sich Hadoop zunehmend zu einer unternehmensweiten, gemeinsamen Daten-Plattform entwickelt – einem Shared Service.

Der Unterabschnitt 4.1.2 führt in die Welt der Big-Data-relevanten Datenbanken ein. Es wird erläutert, warum frühere Investitionen in relationale Datenbanken eine gute Grundlage für Big-Data-Projekte bilden, aber zur Erhaltung der Wettbewerbsfähigkeit nicht ausreichen, wenn man die vielen innovativen Möglichkeiten in Betracht zieht, um Daten in Geschäftswert und Gewinn umzuwandeln. Unternehmen sollten Investitionen in In-Memory-Datenbanken (zur Beschleunigung des Zugriffs auf Daten) sowie in NoSQL (für mehr Flexibilität in den Datenstrukturen und bei der Verarbeitung) prüfen und dabei die TCO über kurze und längere Zeiträume ermitteln. Beide Technologien finden mittlerweile weiten Einsatz und ergänzen die relationalen Datenbanken.

4.1.1 Hadoop

Viele Unternehmen sind von der Vielfalt, Geschwindigkeit und Menge an Daten überfordert. Ein echter Mehrwert für das Geschäft kann aus diesen Daten nur gewonnen werden wenn es gelingt,

- die Daten so billig zu speichern wie möglich,
- die Daten auf vielfältige und flexible Weise zu »befragen«, um wertvolle und umsetzbare Erkenntnisse zu gewinnen,
- diese Erkenntnisse zeitnah zur Verfügung stehen.

Diese drei Fähigkeiten bilden den Kern von Hadoop. Hadoop wird daher in diesem Unterabschnitt als Framework in seiner Gesamtheit vorgestellt, auch wenn sich die Unterprojekte über die Bereiche Daten-Haltung, -Zugriff, Integration, -Sicherheit und Betrieb erstrecken.

Herausforderungen mit herkömmlichen Lösungen

Herkömmliche Lösungen sind in mehrfacher Hinsicht extrem limitiert. In der klassischen Business Intelligence, basiert auf relationalen Datenbanken, müssen Fragen schon festgelegt werden, bevor Daten überhaupt gespeichert werden. Techniker nennen das »Schema on Write« – die Struktur und damit die Semantik der Daten werden in der Konfiguration der Datenbank definiert, welche dann die Daten speichert. Aus diesem Grund sind relationale Datenbanken nicht in der Lage, neue Datentypen zu akzeptieren, deren Schema noch nicht definiert ist. Sie versagen also dann, wenn die Big-Data-Dimension »Vielfalt« relevant ist. Erkenntnisse, die nur aus einem »Entdeckungsprozess« – einem interaktiven Lernen aus den Daten – hervorgehen, können so nur mühsam gewonnen werden.

Außerdem skalieren herkömmliche Lösungen nicht gut, weder aus technischer noch aus wirtschaftlicher Sicht. Hohe Kosten pro Terabyte an Daten machen es unwirtschaftlich, alle Daten über lange Zeiträume vorzuhalten. Enterprise Data Warehouses zum Beispiel speichern meist nur verdichtete Daten, nicht jedoch Rohdaten. Aus Kostengründen werden nur einige wenige Jahre an Daten vorgehalten. Mit existierenden, häufig proprietären Technologien ist es für die meisten Unternehmen schlicht unbezahlbar, alle Daten für lange Zeiträume zu speichern, insbesondere auch Rohdaten oder solche Daten, deren Wert noch unklar ist. Als Fazit ergibt sich: Die Datenmengen übersteigen die verfügbaren Budgets.

Kleine Budgets für große Datenmengen – Hadoop macht Skalierbarkeit bezahlbar

Die Internet-Riesen Yahoo, Google und Facebook sind als erste mit dem Problem konfrontiert worden, große Datenmengen möglichst billig zu speichern, da die Nutzer dieser Online-Dienste ihre Inhalte gratis zur Verfügung stellen und die Kosten nur über Werbung gedeckt werden. Apache Hadoop wurde ursprünglich konzipiert, um große Datenmengen für diese Internet-Riesen extrem günstig zu speichern und auf vielfältige Weise zu verarbeiten.

Mittlerweile wird die Technologie bei Unternehmen aller Branchen eingesetzt.
Hadoop macht Skalierbarkeit bezahlbar. Die Abbildung 5 zeigt, dass Hadoop pro Terabyte circa 20x günstiger ist als verfügbare Alternativen[43].

Dieser Kostenvorteil basiert darauf, dass Hadoop

- meist auf preiswerter Hardware ohne Bindung an einen bestimmten Hersteller läuft und
- Open-Source-Software[44] ist.

Fully Loaded Cost per Raw TB Deployed
US$ '000s

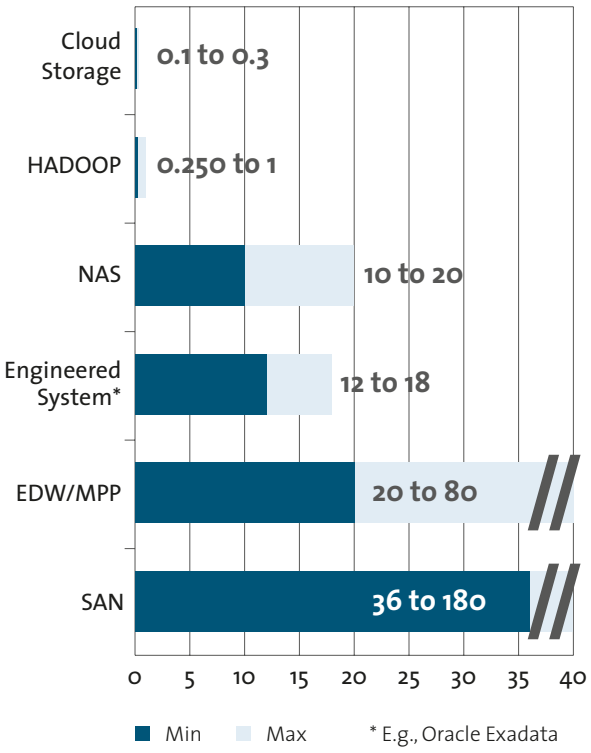


Abbildung 5: Kostenvergleich Hadoop versus Alternativen

Die Abbildung 5 zeigt die gesamten Anschaffungskosten einer Lösung, inklusive Hardware, Software, Installation und Wartung für das erste Jahr, pro Terabyte (TB) an Daten. Typische Kosten für einen Hadoop-Cluster liegen bei 250 bis 1.000 US$ pro TB. 250 US$ pro TB sind repräsentativ für eine Konfiguration, die für »viel Datenspeicher und wenig Verarbeitung« optimiert ist. 1.000 $ pro TB sind typisch für eine Konfiguration, die für »Datenspeicher mit intensiver Verarbeitung« optimiert ist, also Server mit guter RAM-Ausstattung.

Performance durch Parallelität ohne Grenzen

Die kostengünstigen, exponentiell wachsenden Kapazitäten von Online-Storage ermöglichen die Speicherung immer größerer Datenmengen als Basis für analytische Fragestellungen. Was typischerweise nicht wächst, ist die zur Verfügung stehende Zeit zur Beantwortung dieser Fragen. Die technische Herausforderung besteht also darin, massiv zunehmende Eingabedaten zu verarbeiten, ohne dabei aus der Zeit zu laufen, oder in den Big Data definierenden V-Eigenschaften ausgedrückt: Bei zunehmendem Volumen (Volume) wird die Verarbeitungsgeschwindigkeit (Velocity) so gesteigert, dass ein Ergebnis in der gleichen Zeit produziert wird; es geht hier nicht darum, die Antwortzeit selbst beliebig zu verkürzen. Ein solches System hat die Verarbeitung und den Datenfluss einer großen Datenmenge zu organisieren. Der eigentliche Datenzugriff und die Analyse der Daten werden in frei programmierbare Callback-Funktionen ausgelagert; damit wird ein weiteres V erreicht, die Variabilität.

Eine Datenmenge im Petabyte-Bereich kann natürlich nicht im Millisekundenbereich verarbeitet werden und zu einer Antwort führen, da allein die Organisation einer solchen Rechenaufgabe eine gewisse Zeit kostet. Einen Anhaltspunkt für die Laufzeit geben bestimmte Benchmarks wie z. B. Terasort. Für die Sortierung eines Petabyte an Daten werden in einem großen Cluster nur noch 33 Minuten benötigt. Es lassen sich auch noch größere Mengen bearbeiten, ohne dass die Performance einbricht[45].

[43] Beispiele: Speicherlösungen oder analytische Datenbanken wie Enterprise Data Warehouses
[44] Bei Open Source fallen keine Lizenzgebühren für die Nutzung der Software an, obwohl sich viele Unternehmen für den Wirkbetrieb entscheiden, spezialisierte Firmen – die Promotoren der verschiedenen Hadoop-Distributionen – mit der Wartung zu beauftragen.
[45] z. B. lassen sich 10 Petabyte noch in der 12fachen Zeit problemlos sortieren.

Heutige Mehrprozessor-Rechner lassen sich typischerweise mit Hauptspeicher im einstelligen Terabyte-Bereich ausstatten. Ihre IO-Leistungsgrenze liegt bei einigen GB/s[46], was eine Lesezeit von mehreren Tagen für ein Petabyte bedeutet. Für Probleme im Petabyte-Maßstab sind somit selbst leistungsstarke einzelne Rechner um den Faktor 100 unterdimensioniert. Sinnvolle Antwortzeiten lassen sich nur erreichen, indem man viele Rechner parallel an der Aufgabe arbeiten lässt. Nach dem Amdahlschen Gesetz begrenzt bei jedem Parallelisierungsansatz der Kehrwert des nicht parallelisierbaren sequentiellen Anteils die theoretisch erreichbare Performanzsteigerung, z.B. bei 5% serieller Verarbeitung liefern auch beliebig viele Rechner höchstens die 20fache Performance. Bezogen auf das Big-Data-Problem gilt es also, möglichst die gesamte Bearbeitung zu parallelisieren.

Die geschilderten Herausforderungen lassen sich mit einer Shared-Nothing-Architektur bewältigen, bei der jeder Rechner des verteilten Systems unabhängig von allen anderen seine Aufgaben erfüllt.

Die Knoten enthalten auf lokalen Platten oder direkt verbundenen dezentralen Storage-Systemen jeweils einen Teil der Daten und führen darauf eine Teilaufgabe durch. Die Verarbeitung wandert also zu den Daten und nicht umgekehrt die Daten zum Prozessor.

Es gibt Aufgabenstellungen, die sich sehr leicht auf diese Weise parallelisieren lassen. Z.B. die Spracherkennung in einer großen Menge von Audiofiles und die anschließende Umsetzung in entsprechende Textdateien können völlig unabhängig auf vielen Knoten parallel

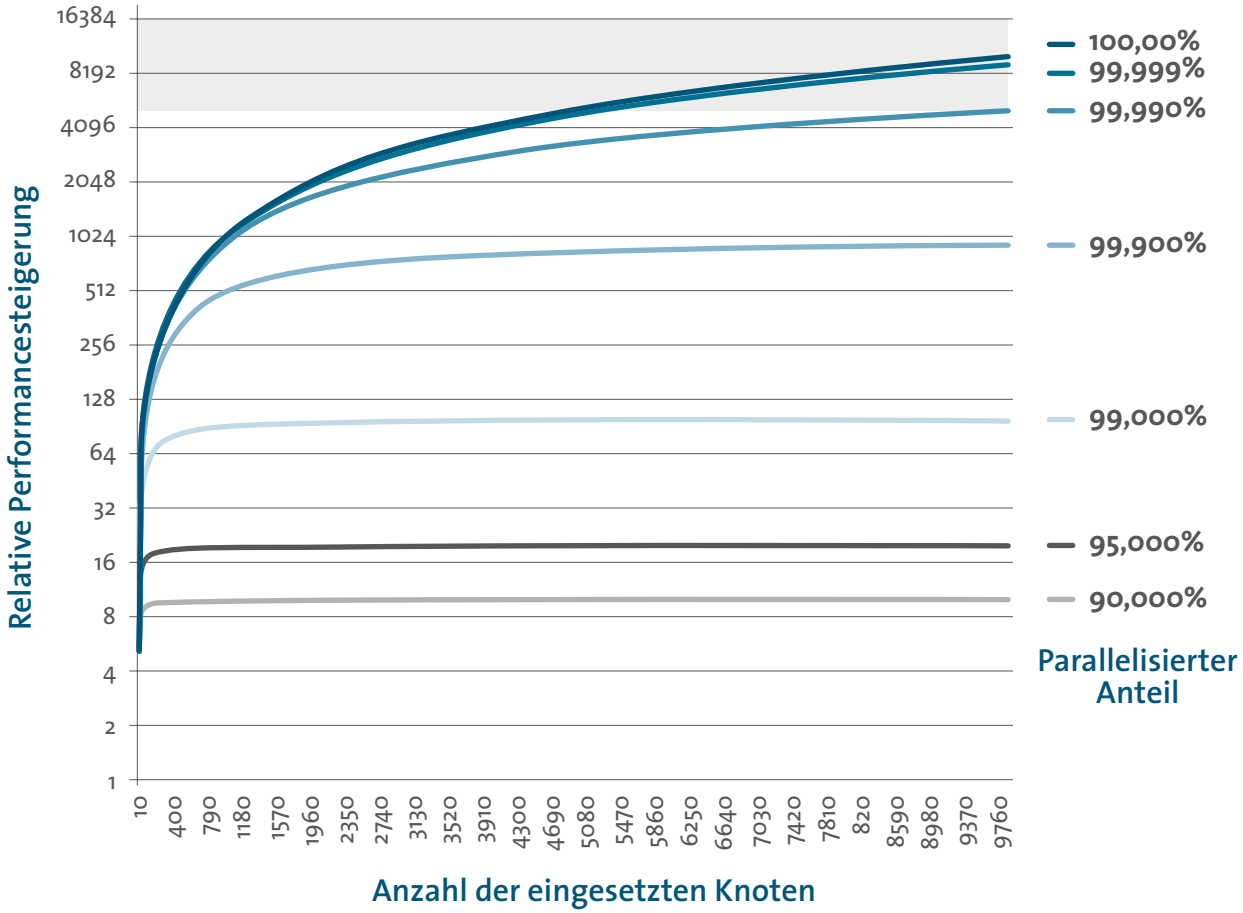


Abbildung 6: Performance-Begrenzung für unterschiedliche Parallelisierungsgrade

[46] PCI Express 3.0 liefert 985 MB/s pro Lane

durchgeführt werden. Das Ergebnis in diesem Beispiel ist allerdings wiederum ziemlich groß, und für die detailliertere Analyse ist eine weitere Verarbeitung notwendig.

Der Kern von Hadoop ist eine hochgradig parallele Architektur sowohl für Datenhaltung als auch für Datenverarbeitung:

- HDFS ist ein verteiltes Dateisystem, optimiert für serielle Verarbeitung, hohes Volumen und hohe Verfügbarkeit.
- MapReduce ist ein Parallelisierungs-Framework für die Datenverarbeitung in großen Server-Clustern.

Stapel-Verarbeitung mit Apache Hadoop MapReduce

Apache Hadoop MapReduce ist das ursprüngliche Framework zum Schreiben von Anwendungen auf Hadoop, mit dem sich große Mengen von strukturierten und unstrukturierten Daten parallel auf einem Cluster von Tausenden von Maschinen verarbeiten lassen.

Die gesamte zu untersuchende Datenmenge wird in sinnvolle Portionen aufgeteilt, und in einer ersten Map-Phase werden die Portionen unabhängig voneinander und parallel von einer Map-Funktion bearbeitet[47]. Die Ergebnisse werden jeweils mit einem Schlüssel gekennzeichnet. Nun werden alle Zwischenergebnisse, die etwas miteinander zu tun haben und deshalb mit demselben Schlüssel

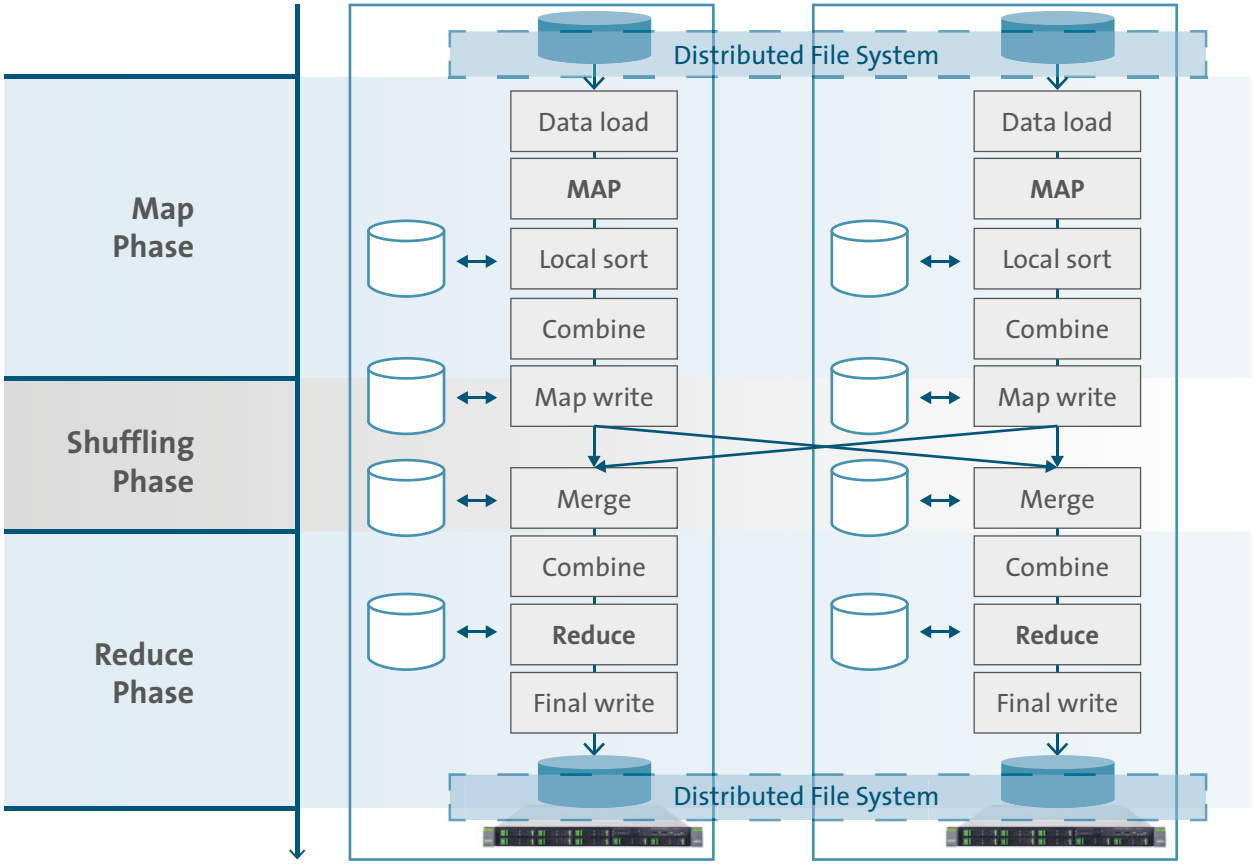


Abbildung 7: Shared-Nothing-Architektur des MapReduce-Ansatzes

[47] z. B. untersucht, entschlüsselt, konvertiert

gekennzeichnet sind, an einer Stelle zusammengezogen. In diesem als »Shuffling« bezeichneten Schritt werden Daten zwischen den Knoten ausgetauscht. Um auch in dieser Phase durch Parallelisierung zu skalieren, ist ein Switched Network ohne Überprovisionierung zwischen den Knoten notwendig. In der darauffolgenden Reduce-Phase erhält die Reduce-Funktion alle Zwischenergebnisse mit gemeinsamem Kennzeichnungsschlüssel, wird diese nun weiter auswerten[48] und dann ein Endergebnis zurückliefern.

In diesem Ablauf übernimmt das MapReduce-Framework die gesamte Ablaufsteuerung – bestehend aus der Portionierung der Datenmenge, Zuordnung der Knoten, Aufruf der Map-Funktion, Shuffling der Zwischenergebnisse, Aufruf der Reduce-Funktion, Herausschreiben der Ergebnisse in die verteilte Datenhaltung. Die fachliche Aufgabenstellung einschließlich der Interpretation der Daten wird in Form der Map- und Reduce-Funktionen eingebracht.

Die Abbildung 7 zeigt die Shared-Nothing-Architektur des MapReduce-Ansatzes: Daten aus dem verteilten Filesystem werden auf dem Rechner gelesen und verarbeitet, auf dem sie physikalisch liegen. In der Map- und Reduce-Phase arbeiten die Cluster-Knoten isoliert und zu 100% parallelisiert. Leistungsfähige Netzwerke gewährleisten in der Shuffle-Phase für den Datenaustausch zwischen den Knoten ein Minimum an Wartezeiten und Serialisierung.

Die Ablaufsteuerung und Kommunikation übernimmt das Framework. Es erreicht sehr gute lineare Skalierung bis in Größenordnungen von mehreren tausend Rechnern. Die problembezogene Programmierung ist äußerst flexibel; sie kann sich auf zwei Funktionen Map und Reduce beschränken und setzt damit kein tiefes Cluster-Know-how voraus[49]. Zur Optimierung gibt es weitere Schnittstellen um z. B. die Zwischenergebnisse auf einem Knoten zu verdichten oder das Shuffling zu beeinflussen.

Die »shared nothing« Architektur von Hadoop stellt die Verfügbarkeit des Gesamtsystems in mehrfacher Hinsicht sicher. Ein System bestehend von Hunderten oder Tausenden von Rechenknoten und Netzwerkgeräten muss mit häufigen Ausfällen von Teilkomponenten gut zurechtkommen. Bei einer Mean Time Between Failures (MTBF) von ca. einem Jahr bei Serversystemen treten in einem 1000-Knoten-Cluster statistisch 3 Serverstörungen am Tag auf. Hinzu kommen Fehlersituationen durch Bugs in der System- und Anwender-Software.

Daher ist die Datenhaltung im Hadoop Distributed File System (HDFS) redundant ausgelegt. In der Standardeinstellung ist jeder Datenblock dreimal vorhanden, neben dem Primärblock existieren eine Kopie auf einem zweiten Server innerhalb desselben Racks und eine zusätzliche Kopie in einem entfernten Rack.

Hadoop ist ein Open-Source-Framework zur Datenverarbeitung auf sehr hoch skalierenden Parallel-Clustern. Zu diesem Top-Level-Projekt der Apache Software Foundation gehören mehr als 13 Unterprojekte. Um den Kern der verteilten Datenhaltung in HDFS und der Verarbeitung in MapReduce gruppieren sich weitere Apache-Open-Source-Projekte in Bereichen wie Daten-Zugriff, -Integration, -Sicherheit und Betrieb. Einen Überblick über die zweite Generation von Hadoop gibt Abbildung 8.

Verbesserungen mit der zweiten Generation von Hadoop

Bisher ist es vielen Unternehmen schwer gefallen, das Potenzial von Big Data wirklich auszuschöpfen. Viele experimentierten zunächst mit einigen der 13 Funktionsmodule von Apache Hadoop (vgl. Abbildung 8), einem Bündel von Technologien, für deren Beherrschung Hadoop-Nutzer der ersten Stunde[50] große Teams einsetzen und mehrere Jahre investieren mussten.

[48] z. B. korrelieren, gruppieren, summieren, filtern
[49] anders als z. B. MPI-Programmierung im HPC-Umfeld
[50] darunter eBay, Facebook und Yahoo

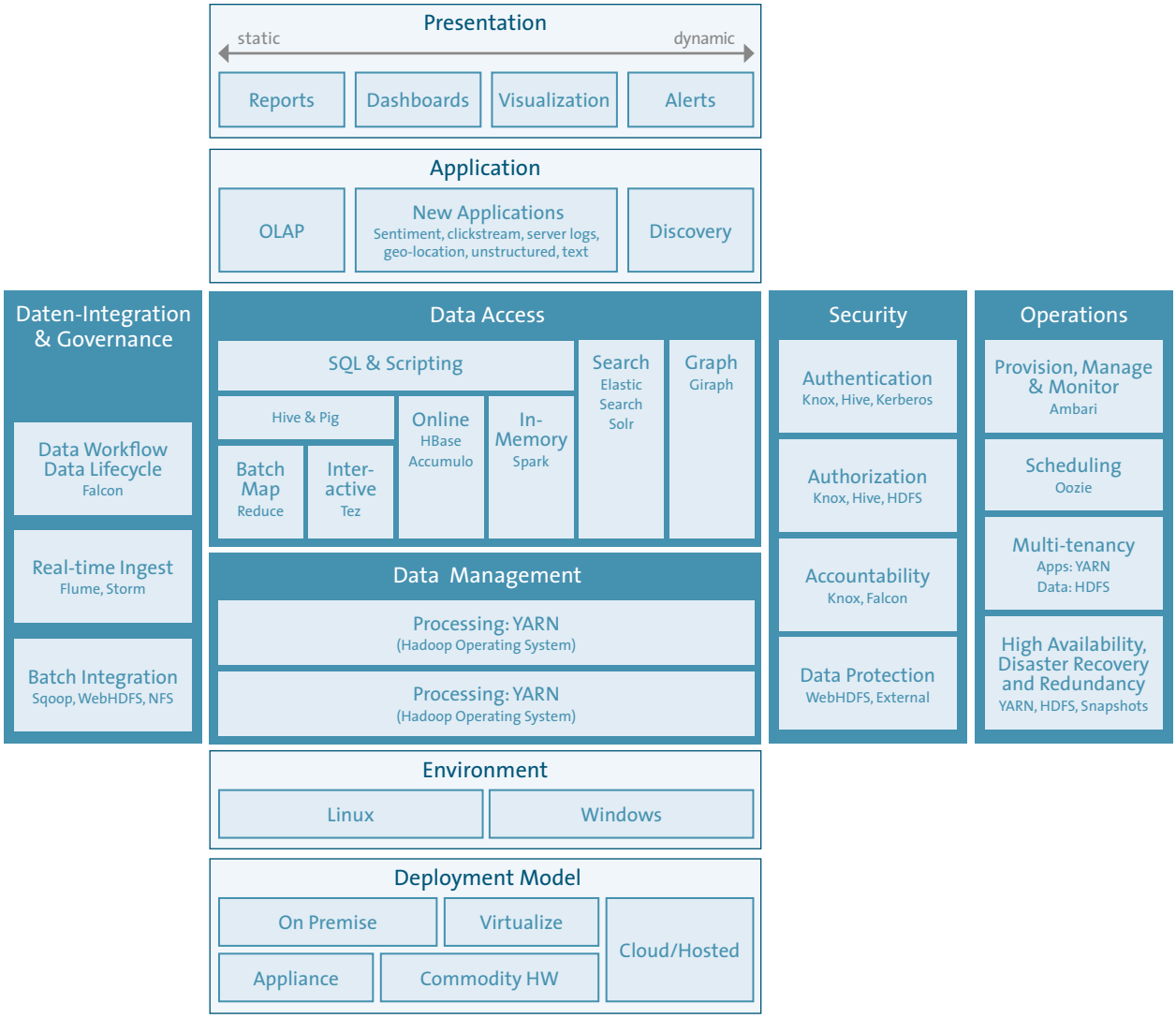


Abbildung 8: Hadoop-Gesamtarchitektur

■ Funktionsumfang einer Hadoop Distribution

Die Hadoop-Technologie der ersten Generation ließ sich weder einfach einführen noch leicht handhaben[51]. Das Ergebnis war, dass das Angebot häufiger als erwartet den Dienst versagte und viele Probleme erst bei hoher Auslastung zutage traten. Ungeachtet des breiten Trainingsangebotes der führenden Anbieter von Hadoop-Distributionen besteht ein Know-how-Defizit in Unternehmen. Viele dieser Lücken werden durch die zweite Generation von Hadoop-Tools geschlossen.

YARN als »Betriebssystem« für Hadoop

Die wohl wichtigste Neuerung in Hadoop 2 ist YARN. Mit der zunehmenden Verbreitung von Hadoop in Unternehmen hat sich gezeigt, dass das System vielfältige Verarbeitungsmodelle – auch jenseits der Batch-Verarbeitung – unterstützen muss, um typischen Unternehmen ein breiteres Anwendungsspektrum bieten zu können. Die meisten Unternehmen möchten Daten im verteilten Datensystem von Hadoop speichern und bei

[51] Neue Nutzer hatten Schwierigkeiten, die unterschiedlichen Komponenten eines Hadoop-Clusters zu konfigurieren. Scheinbar geringfügige und daher leicht übersehene Details wie z. B. Patchversionen erwiesen sich als extrem wichtig.

gleichbleibendem Service-Level unterschiedliche, gleichzeitige Zugriffsmöglichkeiten haben. Dies ermöglicht in Hadoop 2.0 das Ressourcenmanagement-Tool YARN, welches verschiedene Anwendungen voneinander trennt und neben der einfachen Stapelverarbeitung noch eine Vielzahl weiterer Anwendungsfälle unterstützt, darunter interaktive Verarbeitung, Online-Verarbeitung, Streaming und Graphenverarbeitung. Dies ist analog zu Betriebssystemen wie Microsoft Windows oder Apple IOS, die verschiedenste Anwendungen unterstützen.

Dieser neue Ansatz trennt die beiden Aufgaben des JobTrackers in Ressourcenverwaltung und Applikationssteuerung. So kann man ohne Übertreibung sagen, dass sich Hadoop vom preiswerten Daten-Parkplatz zu einem Framework entwickelt hat, das schnelle und fundierte Entscheidungen unterstützt. Das klassische MapReduce-Muster ist dann nur noch eine Ausprägung der Parallelisierung; andere Strategien können durch andere Application Master festgelegt werden und gleichzeitig in demselben Cluster angewandt werden.

Betrieb mit Ambari

Eine der zentralen Erwartungen von Kunden ist, dass das System gut zu handhaben ist. Das trifft vor allem auf die geschäftskritischen Anwendungen zu, mit denen es Service-Provider zu tun haben. Mit dem intuitiven Web-Interface Ambari hat Hadoop hier einen großen Schritt nach vorne gemacht. Über Ambari lassen sich Hadoop-Cluster sehr viel einfacher einrichten, verwalten und überwachen. Ambari ermöglicht eine automatisierte Erstinstallation ebenso wie laufende Upgrades ohne Serviceunterbrechung, eine hohe Verfügbarkeit und die Wiederherstellung im Notfall – das alles sind Faktoren, die für den effizienten IT-Betrieb unverzichtbar sind.

Verbesserte Sicherheit durch Knox

Auch die verbesserte Sicherheit und das optimierte Daten-Lebenszyklus-Management spielen eine große Rolle für Unternehmen, die eine Allzweckplattform für Big Data aufbauen möchten, mit der unterschiedliche Abteilungen, Anwendungen und Datenrichtlinien bedient werden können. Für die Sicherheit sorgt das Knox-System, das einen einzelnen, sicheren Zugang für den gesamten Apache-Hadoop-Cluster bietet. Falcon steuert das Framework für das Daten-Lebenszyklus-Management bei, und zwar über eine deklarative Programmiersprache (ähnlich XML), mit der sich Datenbewegungen steuern, Daten-Pipelines koordinieren und Richtlinien für den Lebenszyklus sowie für die Verarbeitung von Datensätzen festlegen lassen. Ferner unterstützt Hadoop 2 die Verschlüsselung der Daten sowohl während des Transports als auch während der Speicherung auf Festplatten.

Hadoop als Ökosystem

Hadoop ist frei verfügbar unter der Apache-Lizenz. Der Aufbau und Betrieb eines Hadoop-Clusters und die Integration der vielen Hardware- und Softwarekomponenten ist jedoch arbeitsaufwändig und benötigt ein spezielles Know-how. Hier kann man auf Beratung, Training und Support verschiedener Anbieter zurückgreifen, die vorintegrierte Lösungen anbieten, sogenannte Hadoop-Distributionen. Hierzu zählen rein auf Hadoop fokussierte Anbieter wie Hortonworks und Cloudera aber auch etablierte Generalisten wie z. B. IBM, EMC und Intel.

Darüber hinaus wächst das Ökosystem unabhängiger Softwarehändler, die ihre Lösungen auf Hadoop zertifizieren lassen. Eine größere Zahl von Herstellern hat Hadoop als mehr oder weniger zentralen Baustein in seine Lösungen integriert oder bietet Hadoop in der Cloud als Service an. Dies ist aus zwei Gründen wichtig.

- Erstens hängt bei der Kaufentscheidung vieles davon ab, wie sich Hadoop in die bestehende IT-Umgebung integrieren lässt, die in den meisten Fällen Business-Intelligence-Lösungen und Data Warehouses traditioneller Anbieter umfasst.

- Zweitens werden dadurch Bedenken hinsichtlich der mangelnden Kenntnisse in der Organisation ausgeräumt. So verfügt etwa die Deutsche Telekom über etwa 600 IT-Mitarbeiter mit SQL-Kenntnissen. Zwar werden einige dieser Mitarbeiter sich umfassendes Wissen über und mit Hadoop aneignen, doch können

dank der Integration auf Produktebene, wie sie z. B. Microsoft und Teradata bieten, auch solche Mitarbeiter Anfragen über Hadoop stellen, die (noch) keine Hadoop-Spezialisten sind.

Im Abschnitt 5.1 (S. 100) wird die wichtige Rolle erläutert, welche Hadoop in einer modernen unternehmensweiten Daten-Architektur im Zusammenspiel mit etablierten Technologien einnimmt.

4.1.2 Big-Data-relevante Datenbanken

Der Unterabschnitt 4.1.2 bezieht sich auf die unterste Zeile im Big-Data-Technologie-Baukasten (vgl. Abbildung 2).

Herausforderungen im Daten-Management – Bedarf zum Umgang mit Big Data

In einer branchenübergreifenden Studie[52] wurden die aktuelle Nutzung von Datenbanken und dabei auftretende Herausforderungen untersucht. Die Herausforderungen lassen sich in drei Punkten zusammenfassen:

- **Integration verschiedener Datenquellen und -formate:**
  Neue Datenquellen, aber auch zunehmende Volumina und Geschwindigkeiten gestalten die Daten-Integration wesentlich komplizierter, zumal zunehmend gefordert wird, Daten in Echtzeit zu verarbeiten. Im Ergebnis gaben 75% der Organisationen an, mit dem Problem der Daten-Integration konfrontiert zu sein. Das traditionelle Datenmanagement konzentrierte sich auf strukturierte Daten, die Unternehmen in CRM- oder in ERP-Systemen, in Data Marts und Data Warehouses oder anderen Anwendungen zumeist im eigenen Rechenzentrum speicherten. Heutzutage müssen Applikationen jedoch die Verarbeitung unterschiedlich strukturierter Daten auch aus neuen Quellen bewältigen; damit entstehen Integrationsanforderungen, die nicht selten außerhalb der typischen Einsatzbereiche relationaler Datenbanken liegen.

- **Bewältigung des hohen Wachstums der Datenvolumina:**
  Im Durchschnitt verdoppelt sich der Umfang der Daten alle 18-24 Monate; für einige Unternehmen aus den Bereichen Einzelhandel, Finanzdienstleistungen, Verarbeitendes Gewerbe oder Telekommunikation vervierfacht sich in dieser Zeit das Datenvolumen. Unternehmenszusammenschlüsse erfordern schnelle Fortschritte bei der Konsolidierung und Integration von Daten und erhöhen die Last auf den Datenbanken weiter. Die Datenbank-Technologie ist zwar weit vorangeschritten, aber die Verarbeitung von Daten im Umfang von Dutzenden oder Hunderten Terabytes in sehr großen Datenbanken bleibt anspruchsvoll, insbesondere wenn Antwortzeiten unter einer Sekunde gefordert sind.

- **Absicherung erhöhter Performanz:**
  Unternehmen mussten sich schon immer mit der Leistungsfähigkeit erfolgskritischer Applikationen auseinandersetzen – eine Herausforderung, die sich mit steigenden Volumina und Geschwindigkeiten des Datenaufkommens weiter verschärft hat. Demzufolge verbleibt die Performanz unter allen Anforderungen an vorderster Stelle. Probleme in der Performanz von Datenbanken spiegeln nicht selten Engpässe bei den Eingabe- oder Ausgabeoperationen infolge unzureichender Speicherkapazitäten bzw. fehlender Optimierung konkurrierender Prozesse oder suboptimaler Verwaltung von Zwischenpeichern und Indizierungen wider. Andere Ursachen der Probleme können in nicht ausreichender technischer Expertise oder mangelhaft programmierten Daten-Zugriffen aus den Anwendungen liegen. Außerdem beträgt der Aufwand für das Tuning und die Optimierung einer Multi-Terabyte Datenbank oft das Doppelte im Vergleich zu kleineren Datenbanken. Als Konsequenz aus den Performanz-Problemen muss man sich die Frage stellen, ob nicht alternative Architekturen und neuere Technologien für bestimmte Einsatzfälle die bessere Wahl darstellen.

[52] Diese Unterabschnitt basiert auf folgendem Report: Yuhanna, Noel (2013, June 7): The Steadily Growing Database Market Is Increasing Enterprises' Choices. Forrester Report.

Datenbank-Markt in tiefgreifender, mehrdimensionaler Transformation

Das explosionsartige Wachstum der Daten in den Dimensionen Volumen, Geschwindigkeit und Vielfalt setzt etablierte Datenmanagement-Lösungen in Unternehmen unter Druck. Die Nachfrage aus dem Business nach neuen Applikationen, die oft auch umfangreichere Datenmengen nach sich ziehen und den Einsatz fortgeschrittener Analytik voraussetzen, beflügelt die Innovationen im Markt für Datenbanken. Die weitverbreiteten relationalen Datenbanken eignen sich zwar für das etablierte OLTP sowie für Applikationen zur Entscheidungsvorbereitung, jedoch entstehen vielfältige neue Anforderungen aus dem Business[53], die Unternehmen anregen, über die Strategie für ihre Datenbanken neu nachzudenken. Die Entscheider für die Entwicklung und Bereitstellung von Applikationen haben dabei mehr Wahlmöglichkeiten als in der Vergangenheit. Die für Big-Data-Anwendungen relevanten Datenbanken werden im Unterabschnitt 4.1.2 näher beleuchtet und in Abbildung 9 vorgestellt.

Relationale Datenbanken

Relationale Datenbanken bilden in erster Linie die Basis für transaktionale Applikationen. Für den Einsatz in Bereichen wie ERP, CRM, SCM, Call Center oder Auftragseingang müssen Datenbanken folgende Anforderungen erfüllen: Unterstützung der parallelen Arbeit zahlreicher Nutzer, Performanz, Skalierbarkeit. Datenbanken aus dieser Produktklasse bewältigen die Verarbeitung von Echtzeit-Daten, sind für Operationen zur Einfügung, Löschung und Aktualisierung von Datensätzen optimiert, wie sie für anspruchsvolle Applikationen zur Transaktionsverarbeitung typisch sind, und ermöglichen die gleichzeitige Arbeit von Hunderttausenden Nutzern. Datenbanken dieser Produktklasse werden von Unternehmen wie Actian, IBM, Microsoft, MySQL, Oracle, PostgreSQL und SAP angeboten.

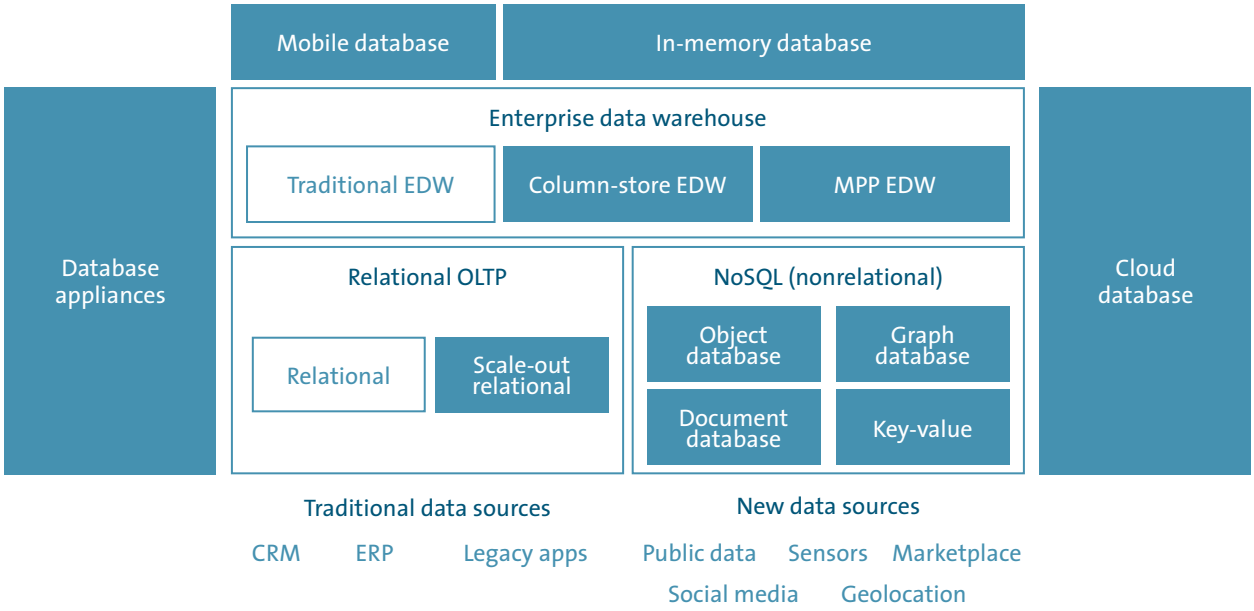


Abbildung 9: Klassifikation von Datenbanken nach Einsatzgebieten[54]

[53] darunter auch der Wunsch, Nutzen aus Big-Data-Anwendungsfällen zu ziehen

[54] Angepasst mit freundlicher Genehmigung von Forrester Research: Yuhanna, Noel; Gilpin, Mike; Brown, Vivian (2013): The Steadily Growing Database Market Is Increasing Enterprises' Choices – An Overview of the Enterprise Database Management System Market In 2013. Forrester Research, June 7, 2013

Relationale Datenbanken mit skalierbarer Architektur

Mit dem Konzept der horizontalen Fragmentierung[55] kann die Skalierbarkeit der Datenbank-Performanz verbessert werden. Zu den Datenbank-Lösungen, die relationale Datenbank-Technologie mit Shard-Architekturen verbinden, zählen Clustrix, MemSQL, ScaleArc, ScaleBase, ScaleDB, StormDB, TransLattice, VMware vFabric SQLFire sowie VoltDB.

Enterprise Data Warehouses

Ein Data Warehouse speichert Daten, die in erster Linie für Business Intelligence (BI), Analytics und andere Berichtsaufgaben genutzt werden. Unternehmen setzen ETL-Technologien ein, um in bestimmten Zeitabständen Daten aus relationalen OLTP-Datenbanken in das Data Warehouse zu übertragen; dort werden sie in »Datenwürfeln« (Data cubes) für die Datenanalyse (OLAP) verarbeitet. Das Data Warehouse bildet die Back-end-Infrastruktur für ein breites Spektrum von Technologien zum Management von Kunden, Produkten, Mitarbeitern und Geschäftsprozessen. Die meisten Unternehmen betreiben Data Warehouses und bauen deren Kapazität und Funktionalität weiter aus.

Anbieter von Datenbanken und Data Warehouses offerieren heutzutage neue, auf den Bedarf von Unternehmen ausgerichtete Appliances und zielen damit auf Anwendungsszenarien, die bisher mit traditionellen Software-Stacks bewältigt wurden. Eine Database Appliance ist eine vorkonfigurierte Gesamtheit von aufeinander abgestimmten Software- und Hardware-Komponenten zur Lösung von Aufgaben mit bestmöglicher Performanz und erlaubt es dem Nutzer, sich auf die geschäftlichen Herausforderungen zu konzentrieren, während dem technische Aspekte in den Hintergrund treten können. Heutzutage setzen 23% der Unternehmen Database Appliances ein – ein Anteil, der sich voraussichtlich in den nächsten drei Jahren verdoppeln wird. Database Appliances bieten verbesserte Performanz und Skalierbarkeit und verringern den Managementaufwand; diese Fortschritte werden durch Konsolidierung, höheren Automatisierungsgrad und zeitlich aufgeschobenen Ausbau der Hardwarekapazität erreicht.

NoSQL-Datenbanken

NoSQL umfasst ein Spektrum verschiedener nicht-relationaler Datenbank-Managementsysteme, die mit Blick auf bestimmte Anwendungsszenarien optimiert sind; dazu zählen Social Media, Predictive Analytics, Web-Applikationen für extreme Einsatzanforderungen, umfangreiche Business Analytics sowie Echtzeit-Applikationen. NoSQL unterstützt in der Regel flexible Schemata, skalierbare Architekturen, unstrukturierte Daten sowie Datenspeicher, die für stark vernetzte Daten optimiert sind. Gegenwärtig setzen 20% der Unternehmen bereits NoSQL ein, wobei MongoDB (7% der Unternehmen) und Cassandra (6%) an der Spitze stehen. NoSQL wird weiter wachsen: So haben weitere 26% der Befragten Planungen, bis 2016 NoSQL-Datenbanken einzusetzen. Applikationsarchitekten und Entwicklungsleiter sollten die Nutzung von NoSQL-Datenbanken für Anwendungen dann ernsthaft in Erwägung ziehen, wenn die spezifischen Anforderungen mit relationalen Datenbanken nicht effektiv abgebildet werden können.

Zur Kategorie NoSQL gehören Key-Value-Datenbanken (Key-Value Stores), dokumentenorientierte Datenbanken (Document Stores), Graph-Datenbanken, sowie Objekt-Datenbanken:

- **Key-Value-Datenbanken (Key-Value Stores) bieten schnellen Zugriff auf verteilte Daten.**
  NoSQL Key-Value Stores sind in der Lage, mit den Dimensionen des Internets umzugehen – Tausende von Servern, Millionen von Nutzern – und das mit extremer Geschwindigkeit, optimiertem Speicher und Retrieval. Zu solchen Leistungsparametern sind Key-Value Stores in der Lage, weil sie auf zahlreiche Funktionen relationaler Datenbanken verzichten und nur solche implementiert haben, die für extreme Web-Anwendungen benötigt werden. In dieser

[55] Englisch: Sharding

Kategorie gehören Aerospike, Amazon DynamoDB, Amazon SimpleDB, Apache Cassandra, Basho Riak, Couchbase, DataStax Cassandra, IBM Informix C-ISAM, Keyspace, memcached, Oracle NoSQL sowie Redis zu den bekanntesten Lösungsangeboten.

- **Dokumentenorientierte Datenbanken (Document Stores) bieten ein flexibles Datenmodel.**
  Eine dokumentenorientierte Datenbank speichert jede Zeile als ein Dokument und bietet dabei die Flexibilität einer beliebig großen Zahl von Spalten und Feldern jeder Größe und jeden Typs. Die Struktur oder das Schema jedes Dokuments ist so flexibel, wie es für die Applikation benötigt wird; sollten neue Anforderungen auftreten, so ist eine Anpassung möglich. Twitter setzt einen Document Store zur Verwaltung der Nutzerprofile unter Einbeziehung der Follower und Kurznachrichten (Tweets) ein. Im Prinzip könnte auch eine relationale Datenbank für diesen Zweck genutzt werden. Jedoch eignen sich dokumentenorientierte Datenbanken für die spezifischen Anforderungen besser – insbesondere dann, wenn die Anwendung ein flexibles Schema erfordert oder unvorhersehbare Verknüpfungen zwischen den Datenelementen. In der Kategorie der dokumentenorientierten Datenbanken haben Apache CouchDB, Couchbase, eXist-dbx, MarkLogic Server, MongoDB, OrientDB sowie terrastore einen hohen Bekanntheitsgrad erreicht.

- **Graph-Datenbanken beschleunigen den Zugriff auf vernetzte Daten.**
  Graph-Datenbanken vereinfachen und beschleunigen den Zugriff auf Daten mit vielen Relationen. Graphen bestehen aus Knotenpunkten (Dingen), Kanten (Verbindungen) und Eigenschaften (Key Values); sie eignen sich für die Speicherung und den Zugriff auf komplexe Beziehungen zwischen Daten. Im Unterschied zu Key-Value-Datenbanken unterstützen Graph-Datenbanken Verbindungen direkt und ermöglichen den schnellen Zugriff auf starke vernetzte Daten. Zu den Einsatzgebieten von Graph-Datenbanken zählen Applikationen für sozialen Netzwerke wie Facebook, Twitter oder LinkedIn, aber auch Empfehlungsdienste (Recommendation Engines), Mustererkennung (Pattern Analysis) zur Betrugsprävention sowie zur Analyse des Konsumentenverhaltens, Analyse von Kommunikationsnetzen zwecks Lastverteilung und Routing, Identitäts- und Berechtigungs-Management oder Predictive Analytics. Eine der weitverbreiteten Graph-Datenbanken ist Neo4J. Weitere stehen zur Auswahl – beispielsweise AllegroGraph, FlockDB, GraphBase, IBM DB2 NoSQL Graph Store, Objectivity InfiniteGraph sowie OrientDB und Giraph.

- **Objekt-Datenbanken eignen sich für den Zugriff auf Objekte und deren Skalierung.**
  Objekt-Datenbanken werden seit Jahrzehnten eingesetzt, sind eng mit den Objekt-orientierten Programmiersprachen verbunden und stehen Applikationsentwicklern auch in einer NoSQL-Ausprägung zur Verfügung. Objekt-Datenbanken haben Vorzüge bei der Steuerung der Gleichzeitigkeit vieler Nutzeraktivitäten sowie bei der Objekt-Navigation und eignen sich für verteilte Architekturen. Diese Datenbanken setzen auf dem gleichen Modell auf wie die Objekt-orientierten Programmiersprachen, zu denen u.a. C#, Java, Objective-C, Perl, Python, Ruby oder Smalltalk zählen. Bekannte Vertreter der NoSQL-Objekt-Datenbanken sind InterSystems Caché, Objectivity/DB, Progress Software ObjectStore, Versant db4, Object Database, Versant Object Database sowie VMware GemStone/S.

In-Memory-Datenbanken

Neben OLTP, Data Warehouses und NoSQL bilden In-Memory-Datenbanken eine besondere Kategorie. In-Memory-Datenbanken ermöglichen den Zugriff auf Informationen in Echtzeit. Auf Daten kann in mehreren Zehnerpotenzen schneller zugegriffen werden, wenn sie sich im Hauptspeicher und nicht auf Plattenspeichern befinden. In-Memory-Datenbanken sind nicht neu, aber erst seit jüngster reizen Applikationen die Vorzüge dieser Architektur vollständig aus und Unternehmen setzen sie verstärkt ein, weil die Speicherkosten fallen und der Komfort zunimmt. In-Memory-Datenbanken bilden die Basis für viele Einsatzbeispiele in Bereichen wie Predictive Modeling, Echtzeit-Daten-Zugriff und Big Data. Die

In-Memory-Speicherung und -Verarbeitung von Kundendaten ermöglicht das Upselling und Cross-Selling von neuen Produkten auf der Basis von Kundenpräferenzen, Freundeskreisen, Kaufmustern und früheren Bestellungen. Zu den Anbietern in diesem Bereich zählen Aerospike, Altibase, IBM, MemSQL, Oracle, SAP, Starcounter sowie VoltDB.

In-Memory Data Grid

Einen Typ von In-Memory-Daten-Haltungstechnologien stellen die als In-Memory Data Grid (IMDG) bezeichneten Plattformen mit verteiltem oder elastischem Caching dar. Lösungen dieser Art vereinen mehrere Server in einem logischen Caching-Cluster, in dem die Daten ausfallsicher und schnell Anwendungen zur Verfügung stehen. Diese Systeme halten die gesamten Daten im Hauptspeicher vor und nutzen den gegenüber der Festplatte erheblich schneller zugreifbaren Arbeitsspeicher des Computers zur Datenspeicherung und -auswertung. Das ist vor allem dort gefragt, wo kurze Antwortzeiten benötigt werden – zum Beispiel im Finanzwesen, wenn Transaktionen eintreffen und Entscheidungen innerhalb von Sekunden getroffen werden müssen.

In-Memory Data Grids (z.B. Terracotta BigMemory, webSphere, GemFire) bieten einen verteilten, zuverlässigen, skalierbaren und konsistenten In-Memory-Datenspeicher. Die Architektur (vgl. Abbildung 10) erlaubt zum einen die gesamte Nutzung des Hauptspeichers eines Servers (scale-up), als auch die Erweiterung von Hauptspeichern über mehrere Server hinweg (scale-out) zu einem einzigen In-Memory-Server-Array. Die Daten können über mehrere Rechner bzw. Rechenzentren verteilt werden, das Grid kümmert sich um den Zugriff auf die Daten, ohne dass die Anwendungen wissen müssen, wo sich die Daten befinden. Dadurch entfällt der zentrale Flaschenhals einer Datenbank, die Anwendungen werden leichter skalierbar. Datenvolumina, die bis in den dreistelligen Terabyte-Bereich gehen können, werden somit in schnellen Hauptspeicher-Technologien auf Basis von Standard-Hardware und -Software verarbeitet.

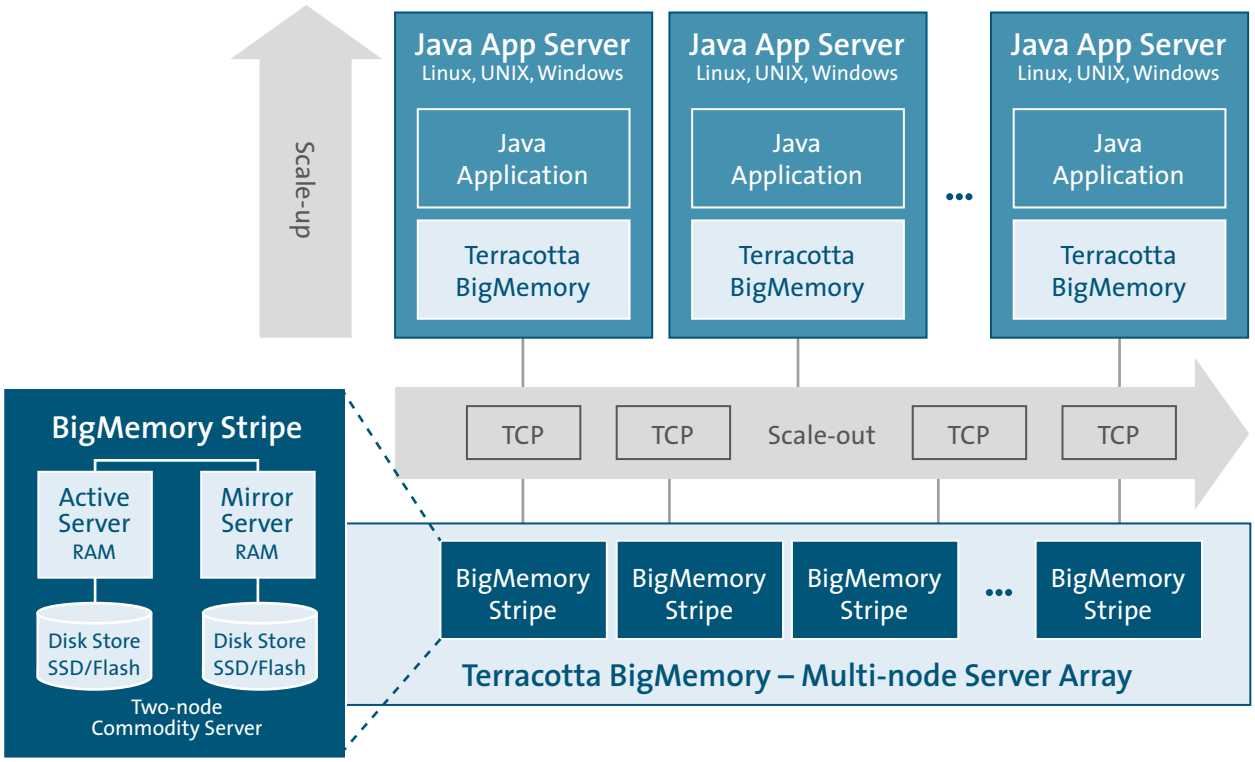


Abbildung 10: In-Memory-Data-Grid-Architektur am Beispiel Terracotta BigMemory

Im Abschnitt 6.2 wird vertiefend auf die Bedeutung und Einsatzszenarien von In-Memory Computing eingegangen.

Schlussfolgerungen

Frühere Investitionen in relationale Datenbanken bilden eine gute Grundlage, reichen aber zur Erhaltung der Wettbewerbsfähigkeit nicht aus, wenn man die vielen innovativen Möglichkeiten in Betracht zieht, um Daten in Geschäftswert und Gewinn umzuwandeln. Unternehmen sollten folgende Gesichtspunkte und Optionen prüfen:

- **Investitionen in In-Memory-Datenbanken zur Beschleunigung des Zugriffs auf Daten:**
  Ohne In-Memory-Technologie könnten Analytics, Personalisierung und Next-Best-Offer-Analyse viel zu lange dauern und dazu führen, dass Unternehmen Chancen vorbeiziehen lassen müssen, Kunden zum richtigen Zeitpunkt Produkte und Dienstleistungen zu verkaufen. In-Memory-Datenbanken könnten den Hebel für die Beschleunigung der am meisten zeitkritischen Verarbeitung von Informationen bilden und durchschlagende Einkaufserfahrungen vermitteln.

- **Investitionen in NoSQL für mehr Flexibilität in den Datenstrukturen und bei der Verarbeitung:**
  NoSQL-Graph-Datenbanken können Social Networks und andere vielfach verflochtene Daten sehr schnell durchforsten und eine erstaunliche Vorausschau und Personalisierung unterstützen. Andere NoSQL-Datenbanken wie Key-Value Stores können dabei helfen, mit flexiblen Datenmodellen umfangreiche Bestände von Kundendaten zu speichern und zu verarbeiten. Wenn diese Flexibilität erfolgskritisch ist, sollten Unternehmen NoSQL umsetzen.

- **Verständnis für die Konsequenzen auf der Kostenseite schärfen:**
  Bei der Bewertung von NoSQL oder anderen spezialisierten Datenbanken sollten Unternehmen die Kosten über die gesamte Lebenszeit im Auge haben und nicht etwa nur die geringen Vorkosten dieser Open-Source-Technologien. Eingesparte Lizenzkosten für eine relationale Datenbank können durch erhöhten Aufwand für Anwendungsprogrammierung und Betrieb bei einer Open-Source-Technologie überkompensiert werden. Das gilt insbesondere für NoSQL-Ausprägungen, bei denen die vollständige Ausnutzung der Flexibilität umfangreiche Entwicklungsarbeiten in Hardware-nahen Programmiersprachen wie Java voraussetzt. Es ist bei NoSQL zu einer sorgfältigen Kostenanalyse und Ermittlung der TCO über kurze und längere Zeiträume zu raten.

- **Vorsicht vor Migrationen im großen Stil weg von RDBMS:**
  Die Migration vorhandener Applikationen auf der Basis von relationalen Datenbanken zu NoSQL bedeutet oft ihre weitgehende Neuentwicklung und die Definition neuer Schemata. Entwickler, deren Erfahrungen auf das relationale Model beschränkt sind, benötigen Weiterbildung bei der vollen Ausnutzung der Vorzüge flexibler Schemata in Applikationen.

4.2 Daten-Zugriff

Der Kernbaustein MapReduce der aktuellen Big-Data-Kerntechnologie Hadoop wird zu Recht als Batch Processing-Komponente eingeordnet (vgl. Tabelle 2). Tatsächlich ist das MapReduce-Framework ja ein System, das die parallele Ausführung von Jobs auf den Datenknoten eines Hadoop-Clusters plant, die Ausführung überwacht und die Berechnungsergebnisse zusammenführt (vgl. Unterabschnitt 4.1.1). Im Unterabschnitt 4.2.1 werden praktische Aspekte der Anwendung der MapReduce-Komponente betrachtet und zwei populäre Bausteine erläutert, welche den Einsatz im IT-Alltag deutlich vereinfachen: Pig und Hive.
Der traditionelle Ansatz für Data Warehouse und Big Data analysiert ruhende Daten. Die Überwachung und Steuerung dynamischer Prozesse bedarf eines anderen Ansatzes. Hierbei werden zeitlich geordnete Ereignisse aus heterogenen Quellen überwacht, verdichtet, gefiltert und korreliert. Das ist das Feld von Streaming und Complex Event Processing (Unterabschnitt 4.2.2).
Ausführungen über Search & Discovery sowie Query ergänzen den Abschnitt.

4.2.1 Batch Processing

Bei der Batch-Verarbeitung werden Geschäftsvorfälle gesammelt und in – häufig nächtlichen, von Online-Betrieb freien – Batch-Läufen[56] verarbeitet. So werden durch Batch-Skripte bzw. ETL-Werkzeuge immer wiederkehrend die neu angefallenen Daten aus den operativen Systemen abgezogen und für entsprechende analytische Zielsysteme aufbereitet. Die neu berechnete Datenbasis bzw. Scores stehen den entsprechenden (Geschäfts-) Prozessen bzw. Analysenwerkzeugen mit entsprechender zeitlicher Verspätung zur Verfügung. Zur Verminderung dieser ETL-Verzögerungen beim Laden von Daten müssen die Batch-Läufe immer weiter parallelisiert und optimiert werden.

Barrieren einer traditionellen Batch-Verarbeitung

In einem Big-Data-Szenario wird man dementsprechend schnell an vier Barrieren einer traditionellen Batch-Verarbeitung stoßen:

Barriere	Erläuterung
Limitierte Sichtbarkeit	Es gibt im Unternehmen zu viele (Alt-) Applikationen, aus denen noch keine Daten über geeignete Schnittstellen abgezogen werden können.
Limitierte Skalierbarkeit	Die Quellapplikationen bzw. deren eingerichtete Schnittstellen sind eventuell nicht auf immer wiederkehrende Abfragen bzw. auf Massendatenexport ausgerichtet.
Limitierte Agilität	Die Ursprungsformate sind oft rigide in relationalen Schemata gespeichert. Batch-Austauschformate wie CSV speichern aus Performancegründen evtl. nur Auschnitte der Originaldaten.
Eingeschränkte Historisierung	Eventuell hält das operative System nur einen kleinen Ausschnitt der Transaktionsdaten produktiv vor, z. B. Daten eines Jahres. Ältere Transaktionsdaten stehen dann nur noch in einem Archiv oder voraggregiert in einem Data Warehouse zur Verfügung.

Tabelle 4: Barrieren einer traditionellen Batch-Verarbeitung

Ein zentraler Hadoop-basierter Enterprise Data Lake hingegen vereinigt die Zwischenspeicherung der Originaldaten aus den Originalsystemen im HDFS und deren Tranformationen auf dem Wege einer hochperformanten parallelen Batch-Verarbeitung.

[56] Die Bezeichnung stammt aus den 60er Jahren des 20. Jahrhunderts, denn die Daten (und oft auch Programme) lagen dabei als Lochkarten vor und wurden als Stapel eingelesen und verarbeitet.

MapReduce-Programmierung als Anwendungs-Hemmnis

Als problematischer Blocker für die Anwendung von Hadoop hat sich bereits in den Kinderjahren dieser Technologie die Notwendigkeit gezeigt, die Verarbeitungs-Jobs in Java programmieren zu müssen. Dies setzt neben guten Entwicklerkenntnissen ein tiefes Verständnis der Hadoop-Arbeitsweise voraus.

In der Praxis stellt sich immer wieder die Frage, ob ein bestimmtes Problem in einem Hadoop-Cluster effektiv gelöst werden kann. Ohne auf theoretische Aspekte der Berechenbarkeit im MapReduce-Paradigma einzugehen, lässt sich die allgemeine Antwort umgangssprachlich so formulieren: »Prinzipiell lassen sich alle (berechenbaren) Aufgaben in Hadoop lösen – man muss nur den passenden Java-Programmcode dazu finden!«

Die Algorithmen müssen also in einem speziellen MapReduce-Stil formuliert werden, was für bestimmte Aufgabenstellungen eine durchaus anspruchsvolle Aufgabe darstellt. Ein typisches Beispiel dafür ist die Implementierung einer INNER JOIN-Operation, wie man sie aus relationalen Datenbanksystemen kennt und wie sie in der Hadoop-Batchverarbeitung sehr häufig vorkommt. Was im deklarativen Datenbank-Standard SQL ohne nachzudenken direkt ausgedrückt werden kann, muss in einem MapReduce-Programm durch diverse Zeilen Java-Code beschrieben werden – dies beinhaltet die Möglichkeit von Fehlern oder eines ineffizienten Jobs. Der Entwickler ist vollständig selbst für die Optimierung zuständig.

Sicherlich stellt allein schon die Verwendung einer komplexen Hochsprache, wie Java oder C#, in einer Programmierumgebung wie Eclipse oder Visual Studio für Geschäftsanwender (wie z. B. Business Analysten) eine zu hohe Hürde dar. Um die Schwelle für die Adaption von Hadoop zu senken, war es also notwendig, die Komplexität der Erstellung der MapReduce-Jobs deutlich zu vermindern. Genau aus diesem Grund wurden die Apache-Projekte Pig und Hive ins Leben gerufen. Beide Komponenten verfolgen den Ansatz, durch einfache, schnell erlernbare Sprachen den Zugang zu Hadoop zu erleichtern und problemorientiert – befreit von den Details des MapReduce-Frameworks – arbeiten zu können. Die Jobs müssen zwar weiterhin in einer Programmiersprache formuliert werden, der Erfolg der Statistiksprache R und der Office-Makrosprache VBA zeigen aber, dass das kein Hinderungsgrund für eine weite Verbreitung sein muss.

Pig als prozedurale Hadoop-Erweiterung

Pig wurde ursprünglich im Jahre 2006 bei Yahoo! entwickelt, um auch Nicht-Programmierern die Arbeit mit Hadoop zu ermöglichen. Daten-Analysten ist es mittels der mit Pig verbundenen Sprache Pig Latin möglich, eigene Skripte zu entwickeln, ohne die Low-Level-Java-Programmierung zu verwenden. Pigs Anwendungsbereiche sind die Datenintegration und Manipulation. Pig Latin ist eine Datenfluss-Programmiersprache, mit der sich Verarbeitungs-Pipelines beschreiben lassen, die dann vom Pig Framework in MapReduce-Jobs überführt werden. Der Begriff Pig ist übrigens kein Acronym, sondern eher eine Metapher. Die folgenden Kernsätze aus der »Pig Philosophy« machen klar, worum es geht:

- »Pigs eat anything« – mit Pig lassen sich beliebige Daten verarbeiten;
- »Pigs live anywhere« – Pig kann innerhalb und außerhalb von Hadoop Clustern verwendet werden;
- »Pigs are domestic Animals« – Pig ist einfach in der Anwendung;
- »Pigs fly« – Pig kann Daten schnell verarbeiten.

Dass die Arbeit mit Pig Latin tatsächlich recht einfach ist, soll das kurze Pig Latin-Skript in Abbildung 11 verdeutlichen[57].

Vermutlich wird jemand, auch wenn er nie zuvor mit der Pig Latin-Programmiersprache zu tun hatte, mehr oder weniger auf Anhieb verstehen, worum es in diesem Beispiel geht: Rohdaten werden aus einem HDFS-Verzeichnis

[57] Vgl.: Wikipedia-Eintrag zu Pig: http://en.wikipedia.org/wiki/Pig_(programming_tool)

```
input_lines = LOAD ‚/tmp/my-copy-of-all-pages-on-internet‘ AS (line:chararray);
words = FOREACH input_lines GENERATE FLATTEN(TOKENIZE(line)) AS word;
filtered_words = FILTER words BY word MATCHES ‚\\w+‘;
word_groups = GROUP filtered_words BY word;
word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS count, group AS word;
ordered_word_count = ORDER word_count BY count DESC;
STORE ordered_word_count INTO ‚/tmp/number-of-words-on-internet‘;
```

Abbildung 11: Pig-Latin-Illustration – Umsetzung des legendären Hadoop Wordcount-Beispiels

geladen (LOAD), es werden Token und dann »richtige« Worte gebildet (TOKENIZE und FILTER), identische Worte werden zu Gruppen zusammengefasst und dann gezählt (GROUP und GENERATE COUNT) und schließlich werden die Worte nach der Häufigkeit des Vorkommens angeordnet (ORDER BY), sowie das Ergebnis in eine Datei geschrieben. Pig Latin-Skripte erinnern in ihrer Ausführlichkeit an die Formulierung in einer Pseudocode-Sprache. Die Verwendung von SQL-Sprachelementen erhöht die Lesbarkeit weiter. In Pig sind typische Datenflussoperationen, wie Join, Sort, Filter verfügbar – zusätzlich lassen sich Bibliotheken und benutzerdefinierte Funktionen einbinden.

Einen ausreichend großen Hadoop-Cluster vorausgesetzt, können Pig-Jobs im Brute-Force-Stil große Datenmengen prinzipiell schnell verarbeiten. Pig führt dabei nur einfache regelbasierte Optimierungen des Datenflusses durch. Der Anwender ist im Wesentlichen selbst dafür verantwortlich, eine sinnvolle Verarbeitungsreihenfolge vorzugeben. In einem Hadoop-Cluster mit mehreren hundert Knoten wird das keine Rolle spielen. In einem Zehn-Knoten-Cluster kann ein ungünstig formuliertes Pig Latin-Skript aber zu frustrierenden Erlebnissen führen.

Flexibilität ist eine ausgesprochene Stärke von Pig. Es werden zwar Datentypen verwendet, aber es wird nicht streng typisiert. Pig ist sehr nachsichtig gegenüber unpassenden Daten und versucht diese bestmöglich zu interpretieren. Außerdem kennt Pig einige recht allgemeine Datentypen, die es gestatten, die Daten zunächst einfach nur zu laden und dann die Struktur der Daten im Skript nach und nach zu verfeinern.

Die Produktivität, die beim Einsatz von Pig erreicht werden kann, ist naturgemäß davon abhängig, ob es Mitarbeiter gibt, die sich bereits gut mit Hadoop auskennen und die Pig Latin-Programmierung beherrschen. Trotz der Einfachheit dieser Spezialsprache bleibt die Gestaltung von Datentransformations-Prozessen hier eben eine Programmieraufgabe, die zudem in einer (derzeit noch) recht kargen Entwicklungsumgebung ausgeführt werden muss. Es stellen sich also die üblichen Fragen nach Lernaufwand, Codewartung und Dokumentation.

Hive als deklaratives Hadoop »Data Warehouse«

Hive ist die zweite Hadoop-Technologie, die häufig im Zusammenhang mit Daten-Integrationsaufgaben genannt wird. Das Hive-Projekt wurde bei Facebook gestartet, um Anwendern eine relationale Sichtweise auf in Hadoop gespeicherte Daten zu ermöglichen. Die verwendete Abfragesprache HiveQL ist stark an SQL angelehnt, wenn auch vom Umfang her sehr deutlich reduziert und spezialisiert. In erster Linie wurde Hive entwickelt, um Anwendern zu gestatten, mit ihren gewohnten Abfrage- und Business-Intelligence-Werkzeugen Auswertungen in Hadoop-Datenbeständen durchzuführen. Man spricht von einem Hive-Warehouse: Daten werden in der traditionellen Tabellenform präsentiert und können über standardisierte Datenbank-Schnittstellen, wie ODBC oder JDBC abgefragt werden. Aufgrund der teils sehr hohen Latenzzeiten bei der Ausführung von HiveQL-Abfragen sind die interaktiven Möglichkeiten allerdings eingeschränkt – dagegen stellt Hive stellt einen sehr praktischen Weg dar, Hadoop-Daten in einem traditionellen Data Warehouse verfügbar zu machen.

Obwohl es nicht die ursprüngliche Aufgabenstellung von Hive war, hat es doch als Werkzeug für ETL-Aufgaben Popularität erlangt. Das hat verschiedene Gründe:

- Für das Erschließen spezieller Datenformate, wie JSON, lassen sich in Hive spezielle Storagehandler und Serializer/Deserialisierer verwenden.
- Es gibt außerdem eine Menge eingebauter HiveQL-Funktionen, die für die Datenanalyse eingesetzt werden können.
- Der wichtigste Grund für die Verwendung von Hive für Datenintegrationsjobs ist sicherlich im Verbreitungsgrad begründet, den SQL in der Datenbankwelt genießt.

HiveQL vereinfacht den Umstieg auf Hadoop, und viele ETL-Aufgaben lassen sich durch SQL-Ausdrücke einfach lösen. Ein kurzes HiveQL-Beispiel dient der Illustration (vgl. Abbildung 12):

Für jemanden, der sich schon einmal mit SQL beschäftigt hat, ist dieses Skript sofort verständlich.

Die Herausforderung beim Einsatz von Hive für ETL-Aufgaben besteht in der Regel in der Überführung der Eingangsdaten in Tabellenform. Ist dies mit Hive-Mitteln nicht möglich, so muss eventuell noch ein Pig-Skript vorgeschaltet werden. Sobald sich die Daten aber in der regulären Form einer Hive-Tabelle befinden, dann ist die Formulierung von ETL-Jobs eine überschaubare Aufgabenstellung.

Die Vor- und Nachteile bei der Arbeit mit Hive ähneln denen mit Pig: Nach der automatischen Übersetzung von HiveQL in MapReduce-Jobs können sehr große Datenmengen verarbeitet werden; Hive ist einigermaßen tolerant gegenüber varianten Schemata (wenn auch nicht so flexibel wie Pig) und die HiveQL-Sprache erschließt sich schnell.

Die Nachteile der Formulierung von Transformations-Prozessen in Skriptform sind bei Hive noch etwas spürbarer als bei Pig, da die Verarbeitungs-Pipelines komplexerer

```
CREATE TABLE logdata(
 logdate string,
 logtime string,
 ...
 cs_Referrer string) ROW FORMAT DELIMITED FIELDS TERMINATED BY , ,;

LOAD DATA INPATH ,/w3c/input/‘ OVERWRITE INTO TABLE logdata;

SELECT
 logdate, c_ip, COUNT(c_ip)
FROM
 WordsInTexts
GROUP BY
 logdate
ORDER BY
 logdate, COUNT(c_ip)
LIMIT 100
```

Abbildung 12: Illustrationsbeispiel für HiveQL

Aufgaben durch die Verwendung von verschachtelten HiveQL-Unterabfragen oder durch temporäre Zwischentabellen gelöst werden müssen. Das trägt nicht gerade zur Klarheit bei und kann zu Performance-Fallen führen.

Werkzeugunterstützung für die Joberstellung

Es ist zu berücksichtigen, dass man bei wachsenden Integrationsprojekten mit manuellem »Hand-coding« ohne Werkzeugunterstützung, wie z. B. einem Team-Repository, regelmäßig an organisatorische Effizienzgrenzen stößt. Integrationswerkzeuge (z. B. ETL, ESB) unterstützen hierbei typischerweise in den Projektphasen Design, Dokumentation, Deployment, Betrieb und Monitoring. Mittlerweile bieten eigentlich alle Integrationssoftware-Hersteller zumindest Konnektoren für persistente Big-Data-Datenhaltungssysteme an (z. B. Hadoop, MongoDB) – ähnlich wie für etablierte Datenbanken oder Dateisysteme, die lesend bzw. schreibend angesprochen werden können.

Um die manuelle Hadoop/MapReduce-Programmierung zu vereinfachen und den Datenzugriff zu erleichtern, sind ETL-Technologien entsprechend erweitert worden. Im Kontext eines Daten- und Transformationsflusses können MapReduce-Jobs konfiguriert, generiert und ausgeführt werden. Hierbei werden Technologien wie Hive und Pig genutzt, aber aus Vereinfachungsgründen gekapselt und abstrahiert. Ähnlich wie in ETL-Verfahren mit Java als Transformationslogik kann nun Hadoop komplementär oder alternativ verwendet werden, um ELT-Prozesse zu automatisieren (vgl. Abbildung 13).

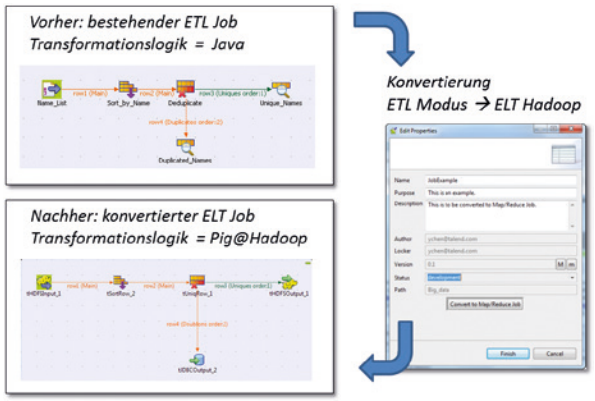


Abbildung 13: Werkzeuge zum Umbau eines vorhandenen ETL-Jobs in einen MapReduce-Job

4.2.2 Streaming und Complex Event Processing

Der traditionelle Ansatz für Data Warehouse und Big Data analysiert ruhende Daten, in denen der Anwender mit verschiedenen Techniken nach dem Gold der Erkenntnis gräbt. Die Überwachung und Steuerung dynamischer Prozesse bedarf eines anderen Ansatzes. Hierbei werden zeitlich geordnete Ereignisse aus heterogenen Quellen überwacht, verdichtet, gefiltert und korreliert. Im Bild des Goldsuchers entspricht dies dem Goldwäscher, der mit seinem Sieb den Datenstrom nach Goldkörnern filtert. Muster (Goldkorn) und Regel (Größe) sind hier konstant, die Daten dagegen variabel und in schneller Bewegung, daher auch der Begriff Streaming.

Für dieses Szenario haben sich zwei grundlegende Ansätze etabliert:

- Distributed Stream Computing Platforms (DSCP) und
- Complex-Event Processing (CEP).

DSCP-Lösungen verwenden ganze Serverfarmen (Grids), um Abfragen und Algorithmen zur Mustererkennung in Echtzeit auf kontinuierliche Datenströme (zum Beispiel Text, Video, Audio) anzuwenden. Im Vordergrund steht hier die massive Skalierung durch isolierte parallele Verarbeitung.

Complex Event Processing

Complex Event Processing ist eine Technologie zur Verarbeitung von Informationsströmen, die Daten von verschiedenen Quellen kombiniert, um Meßwerte zu aggregieren, wichtige Ereignisse zu identifizieren und zeitnah auf sie zu reagieren.

Typische Konzepte enthält die Tabelle 5.

Konzept	Erläuterung
Projektion	Projektionen berechnen Werte aus den Parametern eines Ereignisses oder erzeugen daraus Folgeereignisse.
Filter	Filter prüfen Bedingungen gegen ein oder mehrere Ereignisparameter. Sie propagieren das Ereignis, wenn die Bedingungen erfüllt sind.
Gruppierung	Gruppierungen partitionieren die Ausgangsereignisse. Sie ermöglichern so eine Auswertung nach Gruppen. Sie reichen die Ereignisse dann typischerweise an andere Operation weiter.
Aggregierung	Aggregierungen berechnen Summen, Anzahl, Durchschnitt, Maximum, Minimum von Ereignisparametern.
Join	Joins vergleichen und verknüpfen Ereignisse verschiedener Ströme. Dabei werden die entsprechenden Werte und Timestamps herangezogen.

Tabelle 5: Typische Konzepte in CEP-Anwendungen

Datenströme haben offensichtlich einen Zeitkontext. CEP-Systeme unterstützen daher Zeitfensterabstraktionen. Zeitfenster bewegen sich dabei abhängig von Konditionen der Abfrage bzw. des Filters.

- **Springende Fenster** kumulieren Ereignisse über Zeitabschnitte. Wenn alle Ereignisse für den Zeitabschnitt erfasst wurden, werden sie als Ereignismenge zur Weiterverarbeitung gegeben. Das Fenster »springt« dann um einen fixierten Abschnitt weiter.

- **Ereignisbasierte Fenster** produzieren nur dann Output, wenn während des Zeitfensters Ereignisse auftreten. Die Ergebnismenge ist die Gesamtheit der während der Fenstergröße aufgetretenen Ereignisse.

- **Zählerbasierte Fenster** geben für einen Zähler n jeweils Eventfolgen der Länge n zurück.

CEP erfreut sich hoher Nachfrage im Finanzsektor wie bei Versorgungs- und Fertigungsunternehmen weltweit, ebenso im Internet of Things und im Social Web. Statt traditioneller SQL-Abfragen historischer Daten ermöglicht CEP hochfrequente Analysen gegen Datenströme mit niedrigsten Latenzzeiten.

Die bekannteste Anwendung für CEP ist das hochfrequente algorithmische Trading (HFAT) über regulierte Börsen wie in unregulierten Over-the-Counter (OTC) Swaps in sogenannten Dark Pools. HFAT-Exzesse trugen ebenso 2008 zum Ausbruch der Finanzkrise bei wie zu Kurseinbrüchen in jüngerer Zeit.

Etwas weniger kontroverse Anwendungen von CEP umfassen die Echtzeitüberwachung von Anlagen und Objekten sowie Smart Meters für Gas, Wasser und Elektrizität, ebenso wie Produkt- und Sentiment-Analysen von Social Daten, z. B. Twitterstreams. Anbieter im Gesundheitswesen sehen künftige Anwendungen in Meßgeräten für Blutdruck, Herzfrequenz und andere physiologischer Daten, die über die Cloud Daten an CEP-Systeme liefern, um gesundheitsgefährdende Ausschläge rechtzeitig zu erkennen. Ein weiterer großer Markt für CEP ist die Überwachung seismischer und kosmischer Aktivität in Echtzeit, um laufende und künftige Umweltrisiken zu erkennen.

Verschiedene weitere Anwendungsfälle für CEP und DSCP werden im Abschnitt 5.5 näher beleuchtet.

Streaming und CEP ergänzen Data Warehouses, Dashboards und Hadoop-Datensenken um die Fähigkeit, die Quelldaten zu verarbeiten und gegebenenfalls anzureichern, bevor sie zu Analyse gelangen. Diese Zwischenergebnisse, sogenannte Rollups, können dann in HDFS oder In-Memory-Datenbanken gespeichert und von analytischen Dashboards oder interaktiver Visualisierung genutzt werden.

4.2.3 Search und Discovery

2004 wurde das Verb »Googeln« in den Duden aufgenommen. Spätestens seit dieser Zeit haben sich Unternehmen und Privatpersonen daran gewöhnt, bei der Suche nach Informationen im Internet auf Suchmaschinen zu vertrauen. Eine Suchmaschine bedient ein Kontinuum an Bedürfnissen, von gezielter Suche nach einem sehr speziellen Begriff über ein exploratives, iteratives Browsing bis hin zu Text-Mining und -Analyse.

Relevanz für Big Data – warum gerade jetzt?

Big Data bringt frischen Wind in die Marktkategorie Search and Discovery.
Auf der Anwenderseite ist festzustellen, dass viele Big-Data-Projekte eine Search-Komponente beinhalten. So ist es in E-Commerce-Shops wichtig, über sogenannte Recommendation Engines andere, meist ähnliche Produkte zu empfehlen, die eine größere Wahrscheinlichkeit aufweisen, gekauft zu werden. Auch im Kundendienst oder bei Wissens-Portalen geht es darum, möglichst schnell zu einer Antwort zu kommen.

Letztlich bietet Search ein Fenster[58] zum Einblick in die Datenflut. Ferner ist Text der ursprünglichste und noch am häufigsten verwendete Datentypus für Big Data.
Auf der Anbieterseite kann festgehalten werden, dass die meisten der Best-of-Breed Enterprise-Search-Produkte mittlerweile in größeren Lösungen und Firmen aufgegangen sind[59]. Und auch die Google Search Appliance für Unternehmen gehört offensichtlich nicht zum Kerngeschäft von Google. In der entstandenen Marktlücke sind die Open-Source-Projekte ElasticSearch und Lucene auf großes Kundeninteresse gestoßen – sie gelten als kostengünstig und einfach im Einsatz.

Was bedeutet Search für Big Data?

Mit ElasticSearch bzw. Lucene haben sich im Big-Data-Ökosystem auch Open-Source-Lösungen etabliert. Der »Motor« von ElasticSearch ist ein Java-basierter Datenbank-Server, der Daten in einem Format speichert, welches für Text-basierte Abfragen und Analysen optimiert ist. Diese Funktionalität steht den Anwendungen[60] über HTTP zur Verfügung. Den Kern von ElasticSearch bilden sogenannte Java Libraries in Lucene. Dieser Kern enthält die Algorithmen, die Text Matching durchführen und Indizes aufbauen, welche dann mit Schlagworten durchsucht werden können. ElasticSearch bietet das funktionale Umfeld, um diese Indizes in einer Applikationslandschaft einzusetzen: Application Programming Interfaces (API)[61], Skalierbarkeit, Betreibbarkeit.

Bei Search geht es darum, schnell relevante Antworten zu geben, die verschiedene Sprachen unterstützen und gut organisiert sein müssen.

Wie unterscheidet sich Search von Datenbanken?

Hauptziel von Search ist es, neben der genauen Antwort auf eine Frage auch ähnliche, verwandte Antworten zu suggerieren. Für den Benutzer geht es also um das Entdecken[62] neuer Informationen und Zusammenhänge. Technologien wie ElasticSearch sind darauf ausgerichtet, ungefähre Antworten aus großen Datenmengen zu extrahieren.

---

58 Andere Fenster sind zum Beispiel SQL-Abfragesprachen.

59 Beispiele in dieser Kategorie sind Verity (von Autonomy akquiriert, jetzt HP), Endeca (von Oracle akquiriert und auf eCommerce fokussiert), Vivisimo (von IBM akquiriert) und FAST (von Microsoft akquiriert und in Sharepoint integriert).

60 z. B. einer E-Commerce-Web-Seite

61 mit denen andere Anwendungen wie zum Beispiel ein E-Commerce-Shop die Search-Funktionalität aufrufen

62 Engl. Discovery

Diese Fähigkeit zur Approximation – zur Findung ungefährer Antworten – unterscheidet ElasticSearch von traditionellen Datenbanken. Da Big Data mit unstrukturierten und polystrukturierten Daten zu tun hat, kommt dieser Fähigkeit große Bedeutung zu. Daher besteht die Grundidee darin, statistische Analysen und Algorithmen auf Textdaten anzuwenden. Mit einer möglichst einfachen Benutzeroberfläche wird diese Komplexität vor dem Benutzer weitestgehend verborgen.
Technologisch gesehen weist ElasticSearch gewisse Ähnlichkeiten mit NoSQL-Dokumenten-Datenbanken wie Couchbase oder Cassandra auf, da in beiden Fällen textbasierte Dokumente gespeichert werden, auch wenn NoSQL-Dokumenten-Datenbanken nicht über die Möglichkeit verfügen, ähnliche Ergebnisse zurück zu senden, da keine Indizes erstellt werden.

Search- und Discovery-Technologien wie ElasticSearch eignen sich gut für folgende Fragestellungen:

- Durchsuchen einer großen Anzahl von Produktbeschreibungen für die beste Übereinstimmung mit einem bestimmten Suchbegriff (z. B. »Schirm«) und Rücksendung der besten Ergebnisse,
- Aufzeigen der verschiedenen Abteilungen bzw. Orte, in denen der Suchbegriff (»Schirm«) erscheint (»Facettierung«, z. B. Regenschirm, Sonnenschirm, Abschirmung, unter möglichem Ausschluss anderer Sprachen die auch das Wort Schirm enthalten),
- Suche nach Wörtern, die ähnlich klingen,
- Automatisches Ausfüllen eines Suchfelds auf der Basis eines nur teilweise getippten Wortes, basierend auf zuvor eingegebenen Suchbegriffen, unter Berücksichtigung von falschen Schreibweisen,
- Speichern einer großen Menge von semi-strukturierten (JSON-)Daten verteilt über einen Server Cluster, mit einem bestimmten Niveau an Redundanz.

Bei traditionellen relationalen Datenbanken stehen Präzision und Integrität der Daten im Vordergrund. Antworten sind schwarz oder weiß, nicht Schattierungen von Grautönen. Daher eignet sich Search nicht für Fragestellungen, für die relationale Datenbanken optimiert sind, wie beispielsweise:

- Berechnung des noch verbleibenden Lagerbestands
- Summierung aller ausstehenden Forderungen im laufenden Monat
- Ausführung und Bestätigung von Transaktionen mit Rollback-Unterstützung
- Vergabe von Identifikatoren bzw. Schlüsseldaten, die garantiert nur einmal vorkommen dürfen[63].

4.2.4 Query

Query ist die englische Bezeichnung für Abfrage. In der Welt der traditionellen Business Intelligence ist SQL die Standard-Sprache für Abfragen. Im allgemeinen Sprachgebrauch steht SQL als Abkürzung für Structured Query Language. SQL ist eine Datenbanksprache zur Definition von Datenstrukturen in relationalen Datenbanken sowie zum Bearbeiten (Einfügen, Verändern, Löschen) und Abfragen von darauf basierenden Datenbeständen.

Die Sprache basiert auf der relationalen Algebra. Ihre Syntax ist relativ einfach aufgebaut und semantisch an die englische Umgangssprache angelehnt. Fast alle gängigen relationalen Datenbank-Systeme unterstützen SQL – allerdings in unterschiedlichem Umfang und leicht voneinander abweichenden »Dialekten«.

Durch den Einsatz von SQL strebt man die Unabhängigkeit der Anwendungen vom eingesetzten Datenbank-Management-System an. Mit SQL können Benutzer zum Beispiel Datenbestände auswählen, filtern, sortieren, verknüpfen und umbenennen. Das Ergebnis einer Abfrage

[63] z. B. eine Telefonnummer pro Kunde

sieht aus wie eine Tabelle und kann meist auch wie eine Tabelle angezeigt, bearbeitet und weiterverwendet werden.

SQL-Befehle lassen sich in drei Kategorien unterteilen:

- Befehle zur Definition des Datenbank-Schemas, mit anderen Worten das Festlegen einer Struktur für die Daten,
- Befehle zur Datenmanipulation (Ändern, Einfügen, Löschen) und zum lesenden Zugriff,
- Befehle für die Rechteverwaltung und Transaktionskontrolle.

In der Welt der traditionellen relationalen Datenbanken zeichnet sich SQL unter anderem durch die Fähigkeit aus, die Integrität einer Transaktion durch Commits und Rollbacks zu unterstützen.

Eine Transaktion bezeichnet eine Menge von Datenbankänderungen, die zusammen ausgeführt werden müssen.[64] Der Begriff Commit bezeichnet die Ausführung einer Transaktion.[65] Eine Transaktion wird mit der SQL-Anweisung Commit beendet. Alle Änderungen der Transaktion werden persistent gemacht, und die relationale Datenbank stellt durch geeignete Mittel wie z. B. Logging sicher, dass diese Änderungen nicht verloren gehen.

Kann die vollständige Abarbeitung der elementaren Datenbank-Operationen der Transaktion nicht durchgeführt werden[66], müssen alle durchgeführten Änderungen an dem Datenbestand auf den Ausgangszustand zurückgesetzt werden. Der Vorgang, der alle Änderungen einer Transaktion zurücksetzt, wird als Rollback bezeichnet. Mit dem Befehl Rollback wird eine Transaktion ebenfalls beendet, es werden jedoch alle Änderungen seit Beginn der Transaktion rückgängig gemacht. Das heißt, der Zustand des Systems ist der gleiche wie vor der Transaktion.

In der Welt von Big Data bleibt SQL relevant, weil Apache Hadoop Hive (siehe Unterabschnitt 4.1.1) im Zusammenspiel mit HCatalog es ermöglichen, SQL-Abfragen auch auf Daten in HDFS laufen zu lassen. Mit anderen Worten, Millionen von Benutzern, die mit SQL vertraut sind, können auch Datensätze in Hadoop befragen, ohne selber genauere Kenntnis von Hadoop zu haben.

[64] So ist beispielsweise die Transaktion »Buchung eines Geldbetrags« durch zwei atomare Datenbank-Operationen gekennzeichnet, das »Abbuchen des Geldbetrages von Konto A« und die »Buchung des Geldbetrages auf Konto B«.

[65] Im Beispiel der doppelten Kontenführung wird durch das Verhindern von ungültigen Teilbuchungen eine ausgeglichene Kontobilanz gewährleistet.

[66] z. B. aufgrund eines Fehlers

4.3 Analytische Verarbeitung

Die beiden vorhergehenden Abschnitte des Kapitels 4 haben Technologien vorgestellt, mit denen Daten erfasst, gespeichert und für die Verarbeitung vorbereitet werden. Der Abschnitt 4.3 befasst sich mit der Kernaufgabe von Big Data – der Gewinnung geschäftsrelevanter Erkenntnisse. Dafür werden wichtige Werkzeuge beschrieben, die sich zum großen Teil direkt am Einsatzszenario sowie am Datentyp orientieren.

4.3.1 Orts- und raumbezogene Datenanalyse

Viele Daten weisen einen Ortsbezug auf. Neue Sichten auf Daten erlauben auch Geo-Informationssysteme (GIS). Die schnell steigenden Nutzerzahlen von Smartphones und Digitalkameras bieten die Grundlage, dass heutzutage viele Informationen zusammen mit GPS-Koordinaten erhoben werden. So sind z.B. im Handel Analysen unter Einbeziehung der räumlichen Dimension seit vielen Jahren Standard zur Bewertung neuer Standorte. In die Absatzprognosen für einzelne Artikel gehen Wettermodelle ebenso ein wie die Einwohnerverteilungen bestimmter Einkommensklassen. Versicherungen ermitteln anhand von Geolokationen zusammen mit Überflutungsmodellen die Schadensrisiken von Immobilien.

Mit Big-Data-Methoden werden nun auch anderen Daten, die lediglich implizit über einen Ortsbezug verfügen, mit Geoinformationen verknüpft. So werden Texte nach Orten, Straßen, Restaurants usw. durchsucht. Dieses Beispiel verdeutlicht, wie mit Big-Data-Ansätzen unterschiedliche externe Daten mit internen Daten eines Unternehmens über die Dimension »Ort« in Beziehung gesetzt werden, um neuartige Datenprodukte zu schaffen.

Durch die Verfügbarkeit von Geodaten im großen Maßstab sind Big-Data-Technologien mit effizienten Algorithmen gefragt, die auch große Datenmengen in kurzer Zeit analysieren können. Hier bringen analytische, relationale Datenbanksysteme bereits Funktionen mit, die in Standard-SQL eingebettet sind. So lassen sich komplexe geo-basierte Anfragen[67] mit Standardwerkzeugen auf großen Datenmengen durchführen. Gerade die Möglichkeit, einen Index in einer relationalen Datenbank auch auf der Ortsdimension anzulegen, erlaubt einen effizienten Umgang mit großen Datenmengen.

4.3.2 Web Analytics

Web Analytics umfasst die Messung, die Erfassung, die Analyse und das Reporting von Daten, die es erlauben, eine Webpräsenz zu optimieren. Führende Online-Händler experimentieren täglich auf ihren Seiten: Sie wollen herausfinden, ob mehr oder weniger große Änderungen[68] zu gewünschten Kundenreaktionen führen.

Viele Unternehmen analysieren routinemäßig die »Conversion Rate« einer Seite, die Anzahl der Klicks auf einer Seite, die Suchbegriffe, die zum Besuch der Seite geführt haben, sowie weitere Kennzahlen.[69] Mit Big-Data-Methoden lassen sich aus diesen Daten Erkenntnisse ziehen, die den Wert einer Darstellung von Kennzahlen in Form von Reports bei weitem übertreffen.

Marktführer analysieren komplette Pfade von Benutzern durch die verschiedenen Seiten und optimieren damit die Webpräsenz als Ganzes. Da jeden Tag neue Artikel in einen Webshop eingestellt werden, können sich die Fragestellungen von einem Tag auf den anderen ändern und erweitern. Jeder neue Artikel kann auch neue Eigenschaften erzeugen, die vom Shopsystem protokolliert werden. Es ist nahzu unmöglich, eine solche Dynamik in einem Datenmodell zu erfassen. Vielmehr sollten durch Methoden des Late Bindings (vgl. Abschnitt 5.9) den Daten erst zur Laufzeit einer Analyse eine gewisse Struktur

[67] Umkreissuche, Ermittlung der Schnittmenge von Objekten usw.
[68] Darstellungsgröße von Produktbildern, Anordnung und Beschriftung von Navigationselementen innerhalb der Seite etc.
[69] Solche Services werden von vielen Anbietern als Cloud-Lösung offeriert.

aufgeprägt werden, die für diese aktuelle Fragestellung der Dimension Variety die Komplexität nimmt. Erweist sich eine Analyse als gewinnbringend, dann sollte über eine Operationalisierung hinsichtlich des Datenmodells nachgedacht werden.

Die Vielzahl an Erkenntnissen aus Web Analytics hinsichtlich einzelner Benutzergruppen werden dazu genutzt, um in Echtzeit den Inhalt einer Webseite anhand der Zuordnung eines Benutzers zu einer gewissen Benutzergruppe dynamisch anzupassen.

Graphenbasierte Methoden eignen sich zur Modellierung der aus den Web Analytics gewonnenen Daten, um mehr als die offensichtlichen Abhängigkeiten von z. B. Produkten im Webshop zu erkennen und so Benutzer über die Platzierung von Angeboten[70] gezielter zu steuern.

Des Weiteren wird Web Analytics immer mehr mit anderen Analysetechniken wie z. B. Predictive Analytics verknüpft, um nicht nur das Verhalten eines Kunden zu analysieren und damit Webseiten zu optimieren, sondern auch Voraussagen über seine nächsten Schritte zu treffen und ihm z. B. speziell zugeschnittene Angebote zu präsentieren.

4.3.3 Text- und Semantische Analyse

Inhaltliche Erschließung mittels Semantik

Neben Datenströmen, die z. B. aus der Vernetzung von Geräten stammen oder von Sensoren generiert werden, stellt die Auswertung von Daten, die in textueller Form vorliegen, ein großes wirtschaftliches Potential für Unternehmen dar. So lassen sich beispielsweise durch die Analyse von Social-Media-Daten[71] Produkttrends erkennen oder Dokumente mit Hilfe von Textanalyse und semantischen Technologien durch Zusatzinformation anreichern, so dass die angereicherten Daten ein ganz neues Potential für Anwendungen und Analysen bieten (vgl. Abschnitt 5.7).

Diese Daten sind jedoch ihrer Natur nach unstrukturiert und basieren in wesentlichen Teilen auf natürlicher Sprache. Eine direkte Analyse über klassische Verfahren, wie Data Mining oder Business Intelligence, ist hierfür nicht möglich. Vielmehr kommen linguistische und semantische Verfahren zum Einsatz, mit deren Hilfe aus den unstrukturierten Datenströmen und Texten relevante Informationen extrahiert, Strukturen erkannt und Verknüpfungen der Daten untereinander sowie mit anderen Datenquellen hergestellt werden. In gewisser Weise ist das Ziel, »BI auf Text« zu ermöglichen – dafür sind jedoch innovative Techniken notwendig, wie die folgenden Beispiele verdeutlichen:

- Bei der Analyse von Social-Media-Daten gilt es, die Texte der Nutzer zu analysieren und zu strukturieren, dabei ggf. einen spezifischen Jargon oder Slang zu berücksichtigen sowie eventuell ein Stimmungsbild abzuleiten (Sentiment-Analyse, vgl. Abschnitt 5.3).

- Die Beiträge auf Blogs und Foren werden inhaltlich erschlossen, Problembeschreibungen und Symptome analysiert, Produktbezeichnungen und Komponenten extrahiert (vgl. Abschnitt 5.3).

- In Dokumenten werden Sinnzusammenhänge erkannt und Bezüge zu anderen Informationen wie CRM-Systemen oder Produkt-Katalogen hergestellt.

Die beispielhaft aufgeführten Szenarien erfordern den Einsatz von semantischen und Sprach-Technologien im Big-Data-Umfeld.

[70] »Wer Produkt A gekauft hat, hat auch die Produkte B, C, D gekauft.«
[71] http://www.computerwoche.de/g/big-business-dank-big-data,103429

Verarbeitungsschritte der Sprachtechnologie

Da viele Informationen in Form von Texten vorliegen, gehört die Sprachtechnologie zu den Schlüsseltechnologien für die Gewinnung strukturierter Informationen.

Die Verarbeitung von Dokumenten kann in drei Bereiche eingeteilt werden:

- die dokumentenspezifische Verarbeitung,
- die sprachspezifische Verarbeitung und
- die domänenspezifische Verarbeitung.

Für diese drei Bereiche sind in Tabelle 6 bis Tabelle 8 Sprachtechnologie-Komponenten kurz beschrieben, die häufig für die Informationsextraktion (IE) eingesetzt werden.

Teilaufgaben	Erläuterung
Bereinigung und Normalisierung der Daten	Nachdem die Daten z. B. durch Webcrawls zur Verarbeitung bereit stehen, ist in der Regel eine Bereinigung der Daten für die weitere Textanalyse notwendig, z. B. durch das Entfernen von Markup Tags[72]. Vielfach sind Dokumente auch in unterschiedlichen Zeichenkodierungen gespeichert – zur Anzeige und zur Verarbeitung der Texte ist dann ggf. eine Konvertierung notwendig.
Anreicherung der Dokumente mit Metadaten	Metadaten, wie z. B. das Crawl-Datum, die Versionsnummer, etc. können bei der Textanalyse ebenfalls eine zum Inhalt zusätzliche Informationsquelle dienen, um z. B. die Auswahl der für die Textanalyse relevanten Dokumente einzuschränken.

Tabelle 6: Teilaufgaben bei der dokumentenspezifischen Verarbeitung

Teilaufgaben	Erläuterung
Sprachenerkennung	Die Sprachenerkennung dient dazu die Sprache des Textes zu ermitteln, um in den nachfolgenden Schritten Komponenten mit den für die jeweilige Sprache trainierten Sprachmodellen auszuwählen bzw. Texte in nicht gewünschten Sprachen von der weiteren Verarbeitung auszuschließen.
Satzsegmentierung	Die Satzsegmentierung strukturiert den Text in Sätze. Eine Satzsegmentierung ist in der Regel als Vorstufe für das Parsing notwendig. Zudem hilft die Erkennung von Satzgrenzen auch bei der Extraktion von Relationen zwischen Entitäten, da die Auswahl von möglichen Entitätenkandidaten auf den Satz begrenzt wird.
Tokenisierung	Die Tokenisierung teilt den Text in Worteinheiten. Im einfachsten Falle geschieht dies dadurch, dass Leerzeichen als Wortgrenzen aufgefasst werden. In der Regel werden zusätzlich Satzzeichen wie Punkt, Komma, Semikolon, etc. vom Wort getrennt und als separates Token aufgefasst, selbst wenn sie mit dem Wort verbunden sind. Die Tokenisierung dient als Vorverarbeitung für die weiteren Verarbeitungsschritte.
Wortstammreduktion (engl. stemming)	Die Wortstammreduktion bezeichnet das Verfahren bei der Wörter auf ihren Wortstamm reduziert werden (z. B. werden die Wörter Autos und Automobil zum Wortstamm Auto reduziert). Eine solche Komponente dient u.a. dazu bei einer späteren Suche auch flektierte Wörter zu finden.
Lemmatisierung	Die Lemmatisierung bezeichnet den Vorgang, die Grundform eines Wortes (Lemma) zu bilden. Bei Verben ist dies der Infinitiv, so dass z. B. eine Lemmatisierungskomponente das Wort »lief« in die Grundform »laufen« umformt.

[72] auch als Boilerplate removal bezeichnet

Teilaufgaben	Erläuterung
Part-of-Speech Tagging	Part-of-Speech (POS) Tagging bezeichnet das Verfahren der Zuordnung von Wortarten zu Wörtern. Hierbei ist der Kontext relevant. So kann die Zeichenfolge »langen« je nach Kontext ein Verb oder ein Adjektiv sein, was sich in verschiedenen POS-Tags äußert. Die POS-Tagging-Komponente dient häufig als Vorverarbeitungsschritt für das Parsing.
Parsing	Das Parsing dient dazu, den Text in eine syntaktische Struktur zu überführen. Parsing kann z. B. eingesetzt werden, um über komplexe grammatische Strukturen verbundene Entitäten zu erkennen und somit die Extraktion von Fakten oder Relationen zu ermöglichen bzw. zu verbessern.
Koreferenzauflösung	Koreferenzauflösung hat das Ziel zu ermitteln, ob verschiedene linguistische Objekte im Text auf die gleiche Instanz verweisen. So sollte eine Korefenzkomponente beispielsweise Pronomen[73] mit der im Text vorhergehenden Nennung des Objekts in Beziehung setzen können. Die Zusammenhänge können jedoch auch komplexerer Natur sein, z. B. insofern, als es sich bei »USA«, »den Vereinigten Staaten von Amerika«, »United States« um das gleiche Land handelt.
Erkennung genereller Eigennamen	Die Eigennamenerkennung bezeichnet das Verfahren, Wörtern bzw. Begriffen Typkategorien[74] zuzuordnen. Während die Erkennung von Eigennamen in der Regel domänenspezifisch ist[75], sind verschiedene Eigennamentypen domänenübergreifend[76].

Tabelle 7: Teilaufgaben bei der sprachspezifischen, aber domänenübergreifenden Verarbeitung

Teilaufgaben	Erläuterung
Domänenspezifische Eigennamenextraktion	Für viele BI-Anwendungen, wie z. B. Trendanalyse oder Produkt-Monitoring, ist die Erkennung von spezifischen Entitäten, z. B. die Erkennung von Produkten eine Voraussetzung. In der Regel ist dazu ein Training von Eigennamen-Extraktionskomponenten für die spezifische Domäne auf Basis eines manuell annotierten Datensatzes notwendig.
Stoppwortlisten	Eine Stoppwortliste ist eine Liste mit Wörtern oder Begriffen, die von der Verarbeitung ausgeschlossen werden sollen.
Topic-Modell	Das Ziel der Topic-Modellierung ist die automatische Zuordnung von Begriffen zu Themen, um somit die thematische Gruppierung von Dokumenten zu ermöglichen. Ein Topic-Modell modelliert diese Zuordnung auf Basis von Worteigenschaften, Kontextinformation und anderen aus dem Text zu extrahierenden Informationen.
Faktenextraktion	Die Faktenextraktion hat das Ziel, vorher festgelegte Arten von Fakten in einem Text zu identifizieren und zu extrahieren. Die Bedeutung von Worten und Phrasen erschließt sich besonders gut aus der Sequenz der Worte in einem Satz. Daher modellieren die meisten Verfahren die Abfolge der unterschiedlichen Worte eines Satzes und deren Eigenschaften.
Relationsextraktion	Die Relationsextraktion dient dazu, Relationen zwischen Entitäten im Text zu erkennen und zu extrahieren, z. B. die Erkennung von Protein-Protein-Interaktionen in biomedizinischen Texten. Häufig basiert die Relationsextraktion auf einer vorhergehenden Eigennamenerkennung.

Tabelle 8: Teilaufgaben für spezifische Domänen bzw. Anwendungen

[73] z. B. »er«, »sie«, »es«
[74] z. B. Person, Ort, Organisation
[75] Das Wort »Schwefel« kann beispielsweise für ein Chemieunternehmen ein »Produkt« sein, während es für eine Geologen sinnvollerweise als »Mineral« annotiert werden sollte.
[76] z. B. Zeitausdrücke oder Währungen

4.3.4 Video and Audio Analytics

Neben den Datenströmen, welche unmittelbar auf Text-Dokumenten basieren (vgl. Unterabschnitt 4.3.3), wachsen insbesondere Audio- und Video-Inhalte explosionsartig an:

- Von Unternehmen veröffentlichte Marketing-Videos enthalten detaillierte Beschreibungen zu den Produkten und Dienstleistungen.
- In Trainings-Videos und aufgezeichneten Webinaren finden sich zahlreiche Details für Schulungs-Zwecke.
- Nachrichten-Sendungen bieten reiche Informations-Schätze, wobei hier sowohl offizielle Nachrichten-Kanäle, unternehmens-interne Archive oder auch Internet-Quellen (YouTube) zum Tragen kommen können.

Beispielweise verfügt allein die National Library in Washington über Zettabyte an Audio- und Videomaterial – woraus unmittelbar ersichtlich wird, dass eine manuelle Erschließung ausgeschlossen ist.

Unter Nutzung der im Unterabschnitt 4.3.3 beschriebenen Techniken lassen sich aber auch die Schätze in diesen Daten erheben und somit Audio- und Video-Archive erschließen:

- Zunächst müssen die enormen Datenmengen an sich überhaupt gespeichert und für die weitere Verarbeitung zugreifbar gemacht werden, wofür Big-Data-Techniken aus dem Bereich Daten-Haltung (vgl. Abschnitt 3.2) zum Tragen kommen. Insbesondere die kostengünstige Speicherung sowie die Möglichkeit zur parallelen Verarbeitung bieten hier enorme Vorteile.
- Anschließend werden aus den gespeicherten Audio- und Video-Daten die Text-Informationen extrahiert, wofür sogenannte Transkriptions-Algorithmen genutzt werden können, die je nach Qualität des Ausgangsmaterials sehr gute Text-Protokolle erzeugen.
- Schließlich kommen die im Unterabschnitt 4.3.3 dargestellten linguistischen und semantischen Verfahren zum Einsatz, um die Transkriptionen inhaltlich zu erschließen und Sinnzusammenhänge herzustellen.

4.3.5 Predictive Analytics

Predictive Analytics ist ein Gebiet des Data Mining, mit dem aus einem Datensatz Trends und Verhaltensmuster abgeleitet und vorhergesagt werden können. Hierbei kommen je nach Anwendungszenario verschiedene statistische Algorithmen und Modellierungs-Techniken zum Einsatz, die darauf abzielen, Muster in aktuellen oder historischen Daten zu erkennen und ein System[77] richtig zu beschreiben und daraus Ableitungen für das zu künftige Verhalten dieses Systems treffen zu können.
In der Wirtschaft wird Predictive Analytics z. B. eingesetzt, um transaktionale Daten zu analysieren und daraus Geschäftsrisiken und -opportunitäten frühzeitig zu erkennen.

Üblicherweise beschreibt man die Vorgehensweise von Predictive Analytics in drei Schritten:

- Descriptive (Beschreiben),
- Predictive (Vorhersagen),
- Prescriptive (Empfehlen).

Im ersten Schritt müssen alle relevanten Daten für das entsprechende System gesammelt werden, um daraus die Muster zu erkennen die zu einem bestimmen Verhalten[78] führen können.

Im zweiten Schritt wird ein passendes statistisches Modell entwickelt, welches das Verhalten des untersuchten Systems hinreichend gut beschreibt, um daraus Vorhersagen über sein Verhalten in der Zukunft ableiten zu können.

---

[77] z. B. der Zustand einer Pumpe, das Wetter oder Finanzdaten
[78] z. B. dem Ausfall einer Pumpe oder den Absturz eines Aktienwerts

Im dritten Schritt müssen Empfehlungen ausgearbeitet werden, die das System bei einem bestimmten Trend in eine gewünschte Richting beeinflussen oder ein vorhergesagtes Ereigniss verhindern[79].
Der Einsatz von Predictive Analytics ist ein kontinuierlicher, iterativer Prozess. Durch den fortschreitenden Einsatz werden die eingesetzten Modelle immer weiter verbessert und angepasst, und damit werden auch die Vorhersagen immer präziser.

In-Database Analytics

Müssen größere Datenmengen verarbeitet werden, so kann es auch sinnvoll sein, die Analysen direkt in der Datenbank auszuführen. Das bedeutet, dass die Berechnungen dort durchgeführt werden, wo die Daten gespeichert sind und nicht, wie sonst üblich, die Daten erst zu einem Berechnungsserver transferiert werden, der dann die Analysen berechnet. Das hat den Vorteil, dass weniger Daten über das Netzwerk transportiert werden müssen und die Last des Berechnungsservers reduziert wird.
Dieses Vorgehen nennt sich In-Database Analytics. Dabei wird der für die Berechnung notwendige Programmcode in der Datenbank abgelegt und durch SQL oder eine andere Datenbankanweisungen ausgeführt. Gerade im Big-Data-Umfeld werden durch diese Herangehensweise Datenmengen handhabbar, die vorher nicht verarbeitet werden konnten. Die Hersteller analytischer Datenbanksysteme bieten daher ein breites Spektrum unterschiedlicher Werkzeuge und Methoden an, die es erlauben, komplexe statistische Berechnungen in der Datenbank auszuführen.

4.3.6 Data Mining und R

Data Mining – Extraktion wirtschaftlich nutzbarer Muster

Der Begriff Data Mining ist ein sehr bildlicher Oberbegriff für eine Vielzahl von verschiedenen Methoden, Verfahren und Techniken, der die Intention zusammenfasst – geradezu im Sinne eines »Daten-Bergbaus« – Schätze, also verwertbares Wissen, aus den Daten des Unternehmens zu fördern. Insbesondere bezeichnet Data Mining im Kontext dieses Leitfadens das intelligente, größtenteils automatisierte Aufspüren und die Extraktion von interessanten, d.h. wirtschaftlich nutzbaren Mustern und Zusammenhängen in großen Datenbeständen. Dabei sind die eingesetzten Methoden, Verfahren und Techniken interdisziplinär und stammen aus klassischen Bereichen der Mathematik, Statistik und Informatik sowie der Biologie und Physik.

Da es keine einzige Methode gibt, die für alle möglichen Problemstellungen geeignet ist bzw. alle anderen Methoden dominiert, hängt damit die Entscheidung bezüglich der zu verwendenden Methodik von der jeweiligen Problemstellung sowie – auch dies ist wichtig – von dem Erfahrungshorizont des Data-Mining-Experten ab. Insbesondere wichtig ist, dass die Arbeit von einer mächtigen und leistungsfähigen Plattform unterstützt wird, die zudem noch eine große Verbreitung haben soll, um nicht an dieser Stelle in Engpässe zu laufen.

Da die Behandlung aller am Markt verfügbaren möglichen Plattformen diesen Leitfaden bei Weitem sprengen würde, soll nur eine mögliche Plattform hier einmal näher demonstriert werden. Damit das durchgängige Beispiel dieses Abschnitts bei Interesse auch praktisch nachvollziehbar ist, wurde die kostenlose und frei verfügbare Plattform R gewählt.

Plattform R – De-facto-Standard-Tool für Data Mining

R ist eine freie Programmiersprache für statistisches Rechnen und statistische Grafiken. R ist Teil des GNU-Projekts, auf vielen Plattformen verfügbar[80] und gilt zunehmend als die statistische Standardsprache sowohl im kommerziellen als auch im wissenschaftlichen Bereich[81].

---

[79] z. B. eine Reparatur an einer Pumpe bevor diese ausfällt
[80] http://www.r-project.org/
[81] http://r4stats.com/articles/popularity/

Der Funktionsumfang von R kann durch eine Vielzahl von Paketen erweitert und an spezifische Problemstellungen angepasst werden. Viele Pakete können dabei direkt aus einer über die R-Console abrufbaren Liste ausgewählt und automatisch installiert werden.

Zentrales Archiv für diese Pakete ist das Comprehensive R Archive Network (CRAN). Aktuell stehen über 5.000 Pakete auf CRAN[82] zur Verfügung. R läuft in einer Kommandozeilenumgebung. Darüber hinaus hat der Nutzer die Auswahl unter mehreren grafischen Benutzeroberflächen (GUI), beispielsweise RStudio[83] (vgl. Abbildung 14):

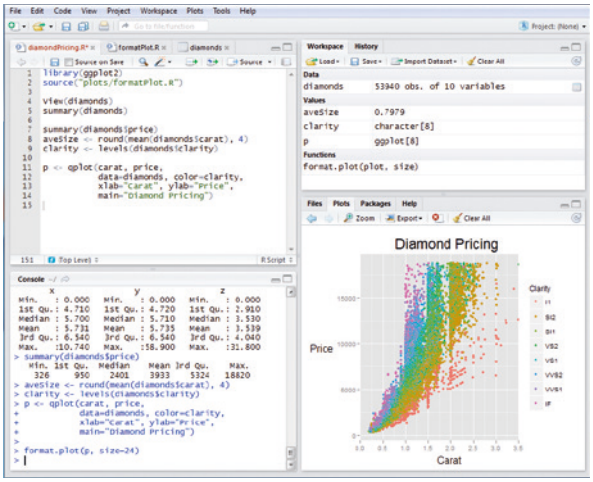


Abbildung 14: RStudio – freie grafische Benutzeroberflächen für R

Im Bereich Data Mining gibt es ebenfalls sehr viele frei verfügbare Pakete[84] sowie GUIs.

Als Beispiel soll das Paket Rattle[85] (vgl. Abbildung 15) dienen:

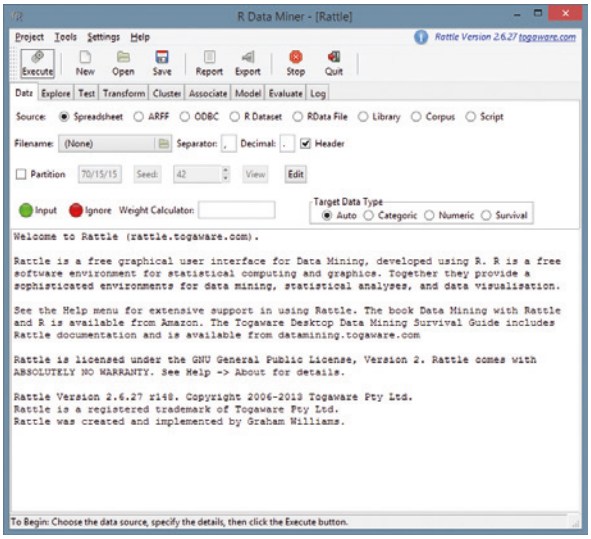


Abbildung 15: Rattle – freie grafische Benutzeroberfläche für Data Mining

Als Illustration dient ein vollständiges, einfaches und bewusst neutrales Beispiel[86], welches einen möglichen Ablauf eines Data Mining-Projektes skizziert: Der Vorhersage der Regenwahrscheinlichkeit auf Basis verschiedener Wetterdaten.

Die Gliederung der Benutzeroberfläche in den Reitern ist dem Data Mining-Prozess nachempfunden. Zuerst wird der mitgelieferte Beispieldatensatz weather geladen (im Data-Tab auf Execute und dann auf Yes klicken) (vgl. Abbildung 16).

82 http://cran.r-project.org/web/packages/index.html
83 http://www.rstudio.com/
84 http://cran.r-project.org/web/views/MachineLearning.html
85 http://rattle.togaware.com/
86 Hier noch einmal detaillierter nachlesbar:
http://www.springer.com/cda/content/document/cda_downloaddocument/9781441998897-c1.pdf?SGWID=0-0-45-1277951-p174110667
Die Benutzeroberfläche ist auch in Deutsch verfügbar, die Screenshots wurden jedoch aus Konsistenzgründen zum besseren Abgleich mit dieser Quelle auf Englisch belassen.

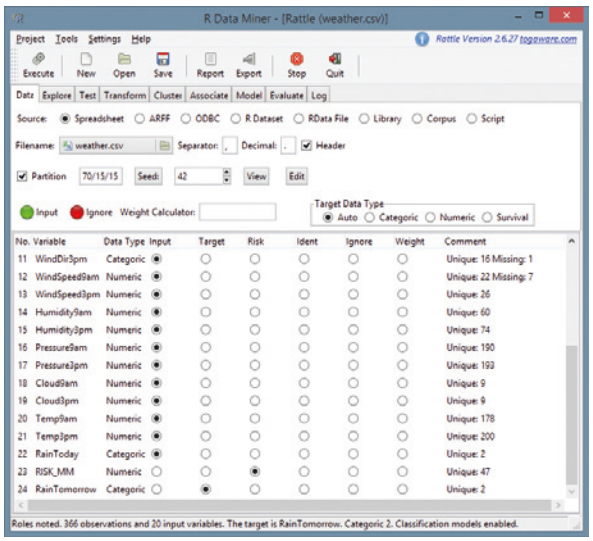


Abbildung 16: Schritt 1 – Laden des Beispieldatensatzes

Die Variable RainTomorrow ist als Zielvariable gekennzeichnet, da in den Daten nach Mustern gefahndet werden soll, ob es morgen regnet oder nicht. Zuerst verschafft sich der Anwender einen Überblick über die Daten, in dem er sich die Verteilung in Bezug auf die Tagestemperatur und Sonnenscheindauer anschaut. Er geht hierzu auf den Reiter Explore und markieren dort unter Distributions für MinTemp und Sunshine jeweils Box Plot und Histogram (vorher deaktiviert er noch Advanced Graphics unter Settings) (vgl. Abbildung 17).

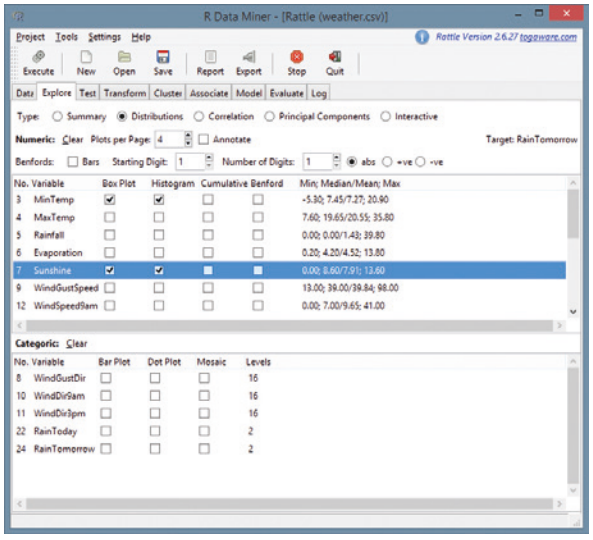


Abbildung 17: Schritt 2 – Gewinnung des Überblicks über die Daten

Danach klickt der Anwender wieder auf Execute, das Ergebnis sieht folgendermaßen aus:

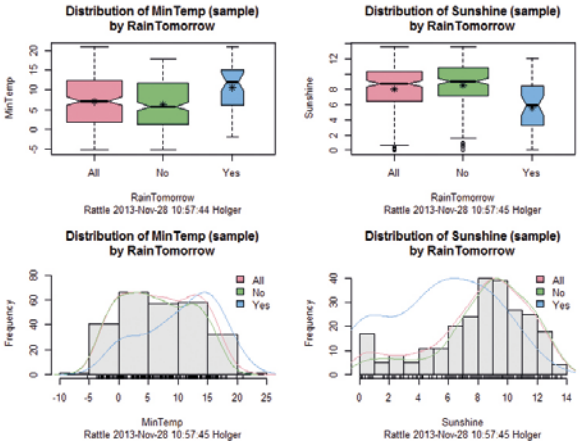


Abbildung 18: Schritt 3 – erste grafische Analyse von zwei Variablen

In diesen Grafiken ist bereits sehr viel Information enthalten; es zeigt sich bereits auf den ersten Blick, dass beide Variablen eine gewisse Trennschärfe in Bezug auf die Prognose von Regen am nächsten Tag haben könnten. So scheinen z.B. eine höhere Temperatur und kürzere Sonnenscheindauer eine höhere Regenwahrscheinlichkeit am nächsten Tag anzukündigen. Der Anwender probiert noch einmal eine andere Variable, die morgendliche Windrichtung, welche dedizierte Kategorien hat, d.h. er markiert Bar Plot, Dot Plot und Mosaic für WindDir9am im selben Fenster und klicken wieder Execute (vgl. Abbildung 19).

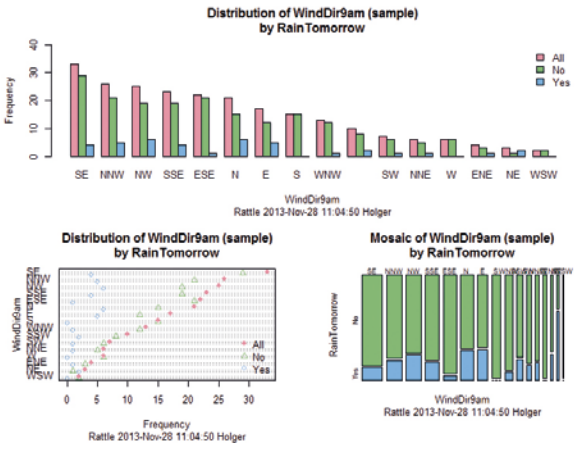


Abbildung 19: Schritt 4 – grafische Analyse einer weiteren Variablen

Ohne auch hier ins Detail zu gehen, ließe sich z. B. die These formulieren, dass nördlicher Wind zu einer erhöhten Regenwahrscheinlichkeit führen könnte, was genauer zu untersuchen wäre. Als nächstes will der Anwender ein Modell bauen, welches die verschiedenen Variablen in einen Zusammenhang stellt. Dafür eignet sich z. B. ein Baum, an dem sich ablesen lässt, welches die geeignetsten Variablen zur Prognose sind. Dafür geht der Anwender auf Reiter Model und klickt Execute (vgl. Abbildung 20).

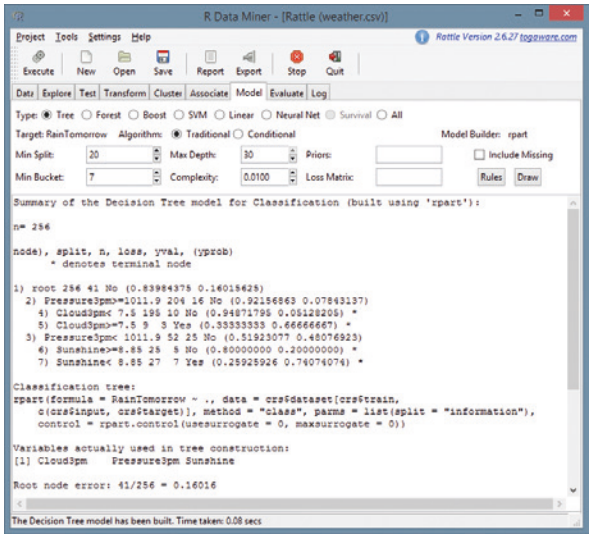


Abbildung 20: Schritt 5 – Untersuchung der verschiedenen Variablen im Zusammenhang

Zur Veranschaulichung klickt der Anwender auf Draw (vgl. Abbildung 21).

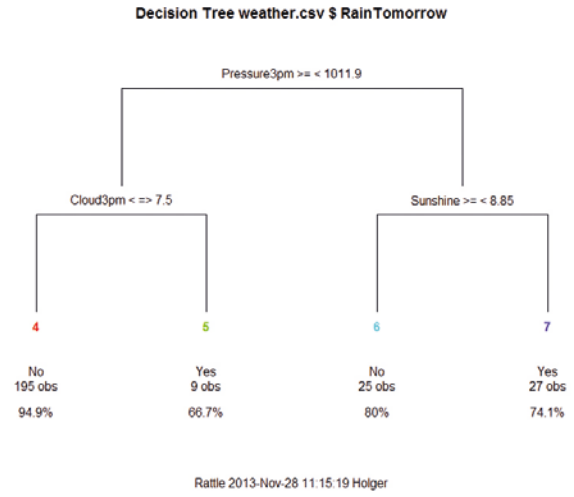


Abbildung 21: Schritt 6 – Generierung eines Entscheidungsbaums

Der Anwender sieht hier z. B. dass nach der Regel Nr. 7 eine 74%ige Regenwahrscheinlichkeit am nächsten Tag vorliegt, wenn der Luftdruck weniger als 1.012 Hektopascal beträgt und die Sonnenscheindauer geringer als 8,9 Stunden ist. Diese Regeln kann über die Schaltfläche Rules auch noch einmal explizit ausgelesen werden (vgl. Abbildung 22):

```
Tree as rules:
Rule number: 7 [RainTomorrow=Yes cover=27 (11%) prob=0.74]
Pressure3pm< 1012
Sunshine< 8.85
Rule number: 5 [RainTomorrow=Yes cover=9 (4%) prob=0.67]
Pressure3pm>=1012
Cloud3pm>=7.5
Rule number: 6 [RainTomorrow=No cover=25 (10%) prob=0.20]
Pressure3pm< 1012
Sunshine>=8.85
Rule number: 4 [RainTomorrow=No cover=195 (76%) prob=0.05]
Pressure3pm>=1012
Cloud3pm< 7.5
```

Abbildung 22: Schritt 7 – Auslesen der Regeln des Entscheidungsbaums

Zum Abschluss dieses Beispiels soll noch die Güte des Modells überprüft werden (beim Reiter Evaluate auf Testing und wieder auf Execute klicken) (vgl. Abbildung 23):

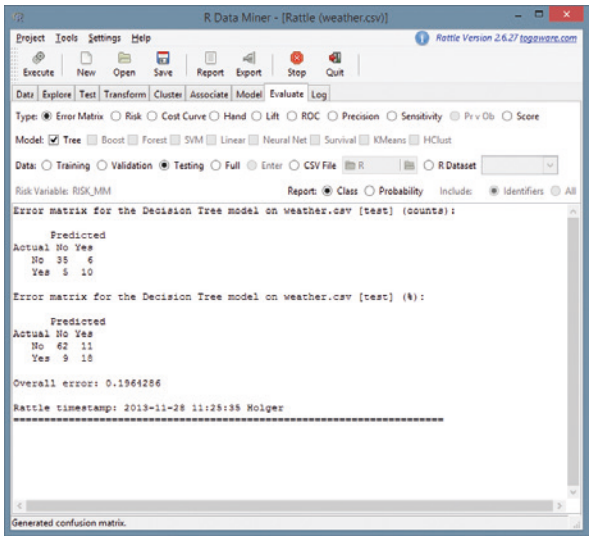


Abbildung 23: Schritt 8 – Überprüfung der Modellgüte

Anhand der sog. Error matrix ist sichtbar, dass das Modell bereits ohne weitere Optimierung in 62% der Fälle »kein Regen« und in 18% der Fälle »Regen« richtig vorhersagt, d.h. in über 80% richtig liegt. In 11% der Fälle macht das Modell eine falsch-positive Vorhersage, d.h. es sagt Regen voraus, es bleibt aber trocken. In der Praxis problematischer dürften die 9% falsch-negativen Fälle sein, in denen man ohne Regenschirm aus dem Haus geht und dann doch nass wird. An dieser Stelle würde in der Praxis eine weitere Verfeinerung des Modells ansetzen müssen.

Das Beispiel kann jedoch nur einen winzigen Teil der Möglichkeiten von Data Mining bzw. von Rattle und R aufzeigen[87].

4.3.7 Machine Learning

Begriffsbestimmung

Der Begriff Machine Learning beschreibt eine Vielzahl von Anwendungen und Methoden, in denen Computerprogramme selbstständig neues Wissen erwerben. Üblicherweise geschieht dies durch statistische oder logische Analysen gegebener Daten sowie durch die Anwendung rechenintensiver Algorithmen, um optimale Lösungen zu gestellten Aufgaben zu finden. Im Unterschied zu einfachen statistischen Auswertungen oder den generell ergebnisoffeneren Zielsetzungen des Data Mining[88], z.B. neue Muster zu finden, ist im Machine Learning meist das Wie im Lösen konkreter Probleme die zentrale Fragestellung. Zum Beispiel können in komplexen Planungsaufgaben zwar alle nötigen Daten explizit vorliegen, aber eine optimale Strategie nur rechnergestützt ermittelbar sein, da für eine manuelle Auswertung zu viele Optionen existieren (kombinatorische Explosion).

Typischerweise wird durch Machine Learning gewonnenes Wissen für die Analyse historischer Daten aufbereitet (vgl. Data Mining) oder in automatischen Prozessen direkt zur Anwendung gebracht. Insbesondere bei der automatischen Anwendung ist die fehlerfreie Erfassung von Informationen, eine korrekte Interpretation der Daten in ihrem Kontext sowie die Generalisierung des daraus erworbenen Wissens von zentraler Bedeutung, da manuelle Validierung und Korrekturen meist entfallen müssen. Zum Beispiel können zeitliche Merkmale und sich ändernde Trends eine automatische Anpassung des Wissens aus historischen Daten an die zukünftige Verwendung erfordern. Daraus ergibt sich, dass Machine-Learning-Anwendungen meist als mehrschichtige Systeme anstatt isolierter Komponenten betrachtet werden und an ihrem Endergebnis, üblicherweise quantitativ, zu messen sind. Praktische Beispiele maschinellen Lernens finden sich vor allem in der modellbasierten Datenanalyse[89], der

[87] Weitere Informationen im Buch »Data Mining with Rattle and R« von Graham Williams (2011), aus dem auch das Beispiel und das oben verlinkte 2. Kapitel stammen (weitere Auszüge: http://www.amazon.de/exec/obidos/ASIN/1441998896/348-62-21)

[88] Data Mining verwendet meist Machine-Learning-Methoden zur Mustererkennung.

[89] In verschiedenen Clustering-Verfahren werden Parameter gesucht, unter denen das Modell die Daten bestmöglich nach vorgegebenen Qualitätskriterien segmentiert.

Vorhersagen-Modellierung[90] sowie der automatischen Interaktion von Maschinen mit ihrer Umwelt[91]. Gemeinsam haben alle Machine-Learning-Anwendungen, dass

- eine konkrete Aufgabendefinition vorliegt,
- Wissen und Erfahrungen aus einer algorithmischen Anwendung gewonnen werden können und
- der Erfolg einer Methode direkt oder in Nachbetrachtung messbar ist.

Strikte Lösungswege zur Herbeiführung eines erwünschten Ergebnisses, bzw. deren Parameter, sind in der Praxis selten im Vorfeld manuell definierbar. Die Probleme sind dafür üblicherweise zu komplex[92], zu stark von den oft zuvor unbekannten Daten einer Anwendung abhängig[93] oder sie unterliegen unkontrollierbaren Umwelteinflüssen und erfordern somit automatische Anpassungen – oftmals sogar in Echtzeit[94]. Machine Learning beschäftigt sich daher mit Verfahren, um günstige Lösungsansätze für Probleme, die manuell nicht oder nur unter hohem Kostenaufwand lösbar sind, automatisch zu erlernen und in der Anwendung weiterzuentwickeln.

Übliche Ansätze beruhen darauf, erhobene Daten statistisch auszuwerten, um Relationen zwischen beobachteten Situationen, den Auswirkungen von ausgeführten Aktionen und der eigenen Leistung aufzudecken. Besonders hervorzuheben ist hierbei der statistische Zusammenhang von Beobachtungen mit vorherzusagenden Eigenschaften, zum Beispiel welche aktuellen meteorologischen Messungen die besten Indikatoren für das morgige Wetter bieten. Daher finden sich im maschinellen Lernen viele Überschneidungen mit klassischer Statistik, Data Mining (Mustererkennung) und auch Visualisierungsaufgaben zur Modellbewertung. Bei Daten- und rechenintensiven Verfahren sind insbesondere auch Lösungen aus verschiedenen Big-Data-Themenbereichen, wie der effizienten Speicherung großer Datenmengen oder der Verteilung von Rechenlast notwendig.

Supervised Learning – das Beispiel E-Mail-Klassifizierung

In sogenanntem Supervised Learning[95] besteht die Herausforderung meist darin, von beobachtbaren Objekteigenschaften über Statistiken und Mustern auf zu erlernende Zusammenhänge mit vorgegebenen Informationen oder zukünftigen Ereignissen zu schließen. Verdeutlicht wird dies am Beispiel der E-Mail-Klassifizierung. Aus Worthäufigkeitsstatistiken können mittels Data-Mining-Methoden zunächst Muster, wie die gemeinsame Verwendung von Wörtern in bestimmten Kontexten, erkannt werden. Ist dann für Trainingstexte bekannt, ob diese z. B. Spam-Mails sind oder nicht, kann von den gefundenen Muster-Instanzen innerhalb einer E-Mail (sog. Features) auf die Wahrscheinlichkeiten der jeweiligen E-Mail-Klassen geschlossen werden. Dieser Lernschritt beinhaltet meist

- die Anwendung von manuell entwickelten mathematischen Modellen,
- die automatische Anpassung der Modellparameter, um die Daten darauf abzubilden, sowie
- die Bewertung, wie exakt die erlernten Modelle die vorgegebene Klassifizierung der Trainingstexte nachvollziehen.

[90] Nach dem Lernen eines Modells aus bekannten Daten wird dieses zur Vorhersage angewendet; die beobachtete Genauigkeit bietet dabei Rückmeldung für das Lernverfahren.

[91] Komplexe Robotersysteme bekommen oft keine genaue Handlungsweise a priori vorgegeben. Stattdessen studieren sie während der Anwendung ihre Umwelt und lernen durch Rückmeldungen über ihre falschen Aktionen.

[92] Für komplexe Themen wie Wettervorhersagen existieren meiste keine exakten Erklärungsmodelle. Deshalb müssen rechnergestützte Näherungen gefunden werden.

[93] Staubsaugroboter können nicht ab Werk feste Wege einprogrammiert bekommen; sie müssen vielmehr selbst ihren Einsatzort kennenlernen und effiziente Strategien planen.

[94] Automatischer Börsenhandel muss sich ständig an aus Umweltbedingungen resultierenden Situations- und Verhaltensänderungen anpassen.

[95] Überwachtes Lernen. Dieser Begriff grenzt Aufgaben, in denen korrekte Lösungen oder Feedback von »außen« – zum Beispiel manuell – zur Verfügung gestellt werden von denjenigen Aufgaben des »unsupervised« oder »semi-supervised« Learning ab, in denen sich Algorithmen auf eigene, fest integrierte Qualitätsmerkmale – wie zum Beispiel der Konsistenz von Clustern – verlassen müssen.

Für den Erfolg des Lernverfahrens ist letztendlich entscheidend, wie gut das erlernte Wissen auf neue E-Mails übertragbar ist – zum Beispiel wenn neue Arten von Spam-Mails versendet werden, um unflexible Klassifikatoren zu umgehen.

Machine Learning und Big Data

Im Bereich großer Datenmengen bedingen sich Machine Learning und Big-Data-Lösungen oft gegenseitig. So sind Analyseverfahren, wie die Klassifizierung vieler Daten, ohne Strategien für die Integration verschiedener Datenquellen, für die Speicherung dieser Daten und deren Zugriff sowie ohne die Verteilung der Rechenlast auf parallele Rechnerknoten kaum möglich.

Wiederum stellen insbesondere Echtzeitanwendungen im Big-Data-Umfeld oft Anforderungen, welche menschliche Kapazitäten übersteigen. So sind zum Beispiel in Hochsicherheitsbereichen Fehler oder Verstöße anhand von Mustern in Sensordaten zu erkennen und über Reaktionen, wie der Abschaltung von Ressourcen, innerhalb von Sekunden zu entscheiden. Andere datenintensive Dienstleistungen, wie die Empfehlungsdienste großer Internetseiten (sog. Recommender-Systeme), sind auf Grund der anfallenden Datenmengen ohne intelligente automatische Verfahren, und somit insbesondere auch maschinellem Lernen, undenkbar.

Existierende Machine-Learning-Lösungen

Auf Grund der Heterogenität der Anforderungen an Machine-Learning-Verfahren sind vorkonfigurierte Komplettlösungen zum Direkteinsatz im Big-Data-Umfeld zurzeit noch selten. Stattdessen kann ein üblicher Arbeitsablauf festgestellt werden, der sich mit vielen anderen Big-Data-Prozessen deckt und dabei auf eine Vielzahl verfügbarer Systemkomponenten zurückgreifen kann. Dieser Standardaufbau von Machine-Learning-Anwendungen ist in Abbildung 24 illustriert.

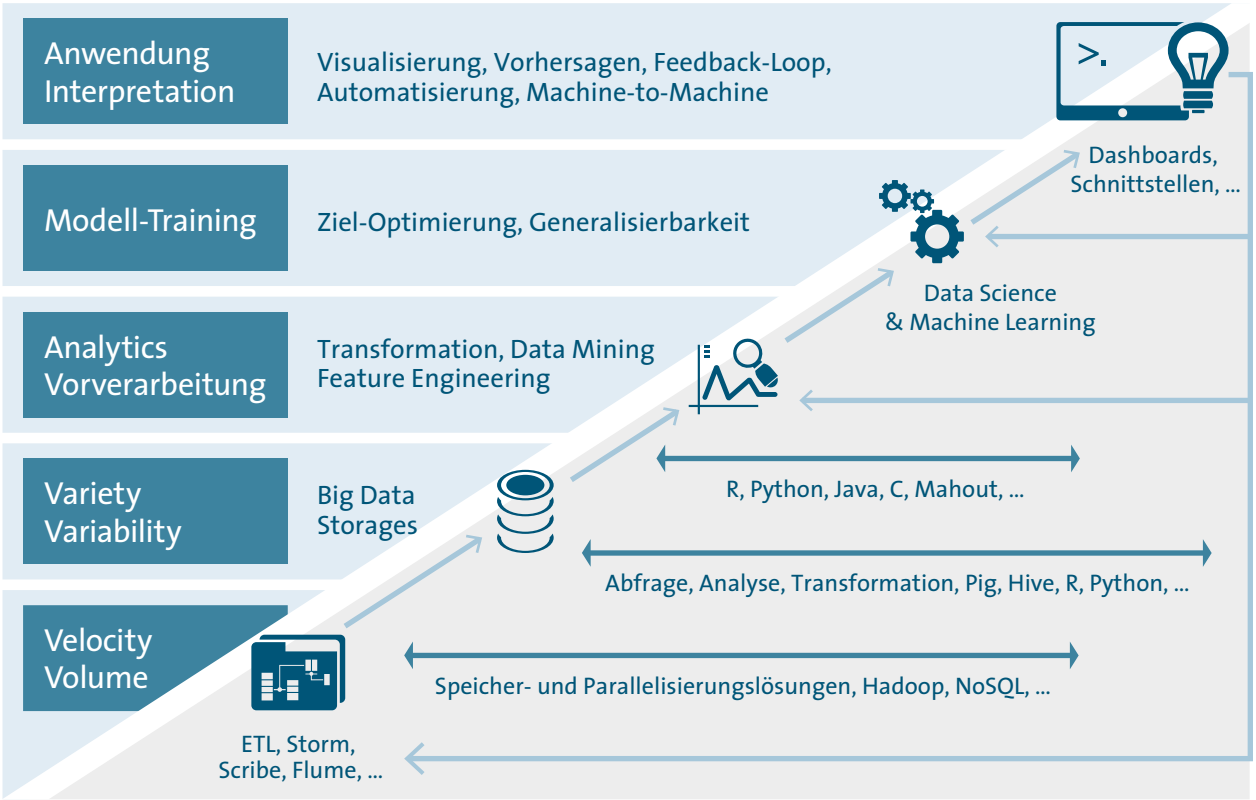


Abbildung 24: Machine-Learning-Pipeline

Zunächst müssen die für die Anwendung notwendigen Daten gewonnen, gespeichert und bereinigt werden. Anschließend sind sie für die konkrete Aufgabe und die eingesetzten Lernverfahren zu transformieren (z. B. Normalisierung) und zu analysieren (z. B. Mustererkennung). Schließlich werden Machine-Learning- bzw. Optimierungs Methoden angewendet, um aus diesen Daten und vorbereiteten Merkmalen und Mustern deren Zusammenhänge mit den Zielvorgaben zu lernen sowie das Wissen auf weitere Anwendungen zu übertragen. Optional können die erlernten Modelle zur Überprüfung visualisiert werden. Üblicherweise erfordert die vorbereitende Datenanalyse, wie zum Beispiel das Erkennen von Mustern in E-Mails, die meiste intellektuelle Arbeit sowie starke Domänenkenntnisse seitens der Entwickler. Für die Optimierung oder das Trainieren der gewählten Modelle und Systeme stehen oft schon Standardlösungen zur Verfügung.

Insbesondere im Bereich des Supervised Learning – der Herstellung des Zusammenhangs zwischen Beobachtungen und Zuweisungen – haben sich verschiedene abstrakte Verfahren entwickelt, die auf die (semi-) manuelle Datenanalyse und deren resultierende Kennzahlen aufbauen, indem sie diese statistisch auswerten und in Korrelation zu vorhandenem Wissen setzen. Dazu gehören u.a.

- lineare Regression,
- neuronale Netze,
- Entscheidungsbaumverfahren,
- Bayes'sche Modelle,
- Nachbarschaftsklassifizierer und
- Support Vector Machines.

Da diese jeweils nur für bestimmte Probleme geeignet sind, erfordert deren Wahl und Parametrisierung Rücksicht auf die zu lernende Relation (z. B. Regression oder Klassifikation), Beachtung der Art der zuvor gefundenen statistischen Merkmale und Muster (z. B. kategorische oder kontinuierliche Variablen), Experimente und Simulationen, sowie die nötige Detailkenntnis für die Ergebnisbewertung.

Bewertung von Verfahren unter Big-Data-Gesichtspunkten

Viele Machine-Learning-Verfahren wurden bereits vor dem Aufkommen der spezifischen Big-Data-Probleme entwickelt. Dementsprechend beinhalten viele Algorithmen Annahmen und Methoden, die großen Datenmengen nicht standhalten, wie zum Beispiel dem paarweisen Vergleich aller Datenpunkte untereinander bei akzeptierbarem Zeit- und Rechenaufwand. Dieser kann auf Grund des exponentiellen Wachstums der Arbeitsschritte auf großen Datenbanken selbst mit höchster Rechenleistung nicht vervollständigt werden. Daher sind Algorithmen zu bevorzugen, welche eine lineare Speicher- oder Rechenkomplexität aufweisen.

Desweiteren sind viele heute verfügbare Softwareimplementierungen unter der Prämisse moderater Datenmengen entstanden und daher oftmals allein auf Grund technischer Realisierungen nicht im Big-Data-Betrieb einsetzbar. Dazu zählen insbesondere Programme, die versuchen, Daten komplett in den Arbeitsspeicher zu laden, ohne eine Verteilung der Speicherung oder Rechenleistung vorzusehen.

Machine-Learning-Komplettpakete, die auch Strategien wie MapReduce und Speichersysteme wie Hadoop unterstützen, sind zurzeit noch rar. Beispiele solcher Lösungen und Anbieter dieser sind das Open-Source-Projekt Apache Mahout, das Unternehmen Skytree, sowie RapidMiner/ RapidAnalytics.

Des Weiteren zeichnen sich insbesondere die Programmiersprachen R, Python, Matlab, Java und C samt Derivate jeweils durch eine Vielzahl Frameworks und Programmbibliotheken aus, die in der Praxis häufig für die Realisierung von Machine-Learning-Anwendungen, auch im Big-Data-Umfeld, verwendet werden.

4.3.8 Reporting

Big Data dient letztendlich der Erkenntnis von Zusammenhängen. Die Bedeutung multidimensionaler Datenmodelle für hypothesengestützte Analysemethoden ist hinlänglich bekannt[96] und soll hier nicht vertieft werden. Der Abschnitt wiederholt kurz das Bekannte, betrachtet neue Aspekte durch Technologien wie In-Memory und bietet abschließend ein praktisches Anwendungsbeispiel sowie Empfehlungen.
Die Basistechnologie für Reporting bildet das OLAP.

OLAP-Formen

Die Basis für OLAP bietet der OLAP-Würfel (Cube), mit dem Daten multidimensional analysiert werden (vgl. Abbildung 25[97]). Der Würfel ordnet dabei die Daten logisch nach verschiedenen Dimensionen an, wie zum Beispiel Zeit, Region oder Produkt.

Nach Art und Zeitpunkt des Zugriffs auf die Daten unterscheidet man klassisch:

- MOLAP (multidimensionales OLAP) speichert Zahlen in Form von Datenpunkten. Zur Laufzeit steht damit ein performanter Cube zur Verfügung, der allerdings häufig in langwierigen Rechenoperationen berechnet werden muss, oft in Servicefenstern über Nacht.
- ROLAP (relationales OLAP) greift auf eine relationale Datenbank zur Laufzeit zu. Hierbei wird auf die Vorausberechnung verzichtet. Die langsameren Zugriffszeiten werden durch den Wegfall der Berechnungen und die Möglichkeit zu Ad-hoc-Analysen aufgewogen.
- HOLAP (hybrides OLAP) bietet eine Mischform zwischen MOLAP und ROLAP.
- Durch neuere Technologien ist es möglich, die Würfel zu größeren Teilen oder vollständig in memory zu halten und damit die Berechnungsoperation zu beschleunigen. DOLAP (Desktop OLAP) ist eine zusätzliche Form, bei der der Würfel im Arbeitsspeicher des Clients entsteht, im einfachsten Fall in Excel[98].

Weiterhin kann zwischen multidimensionalem und tabellarischem OLAP unterschieden werden. Im letzteren Falle werden Faktentabellen ganz oder teilweise in den Arbeitsspeicher geladen. Hierbei helfen neue In-Memory- und Kompressionstechnologien fast aller Datenbankanbieter.

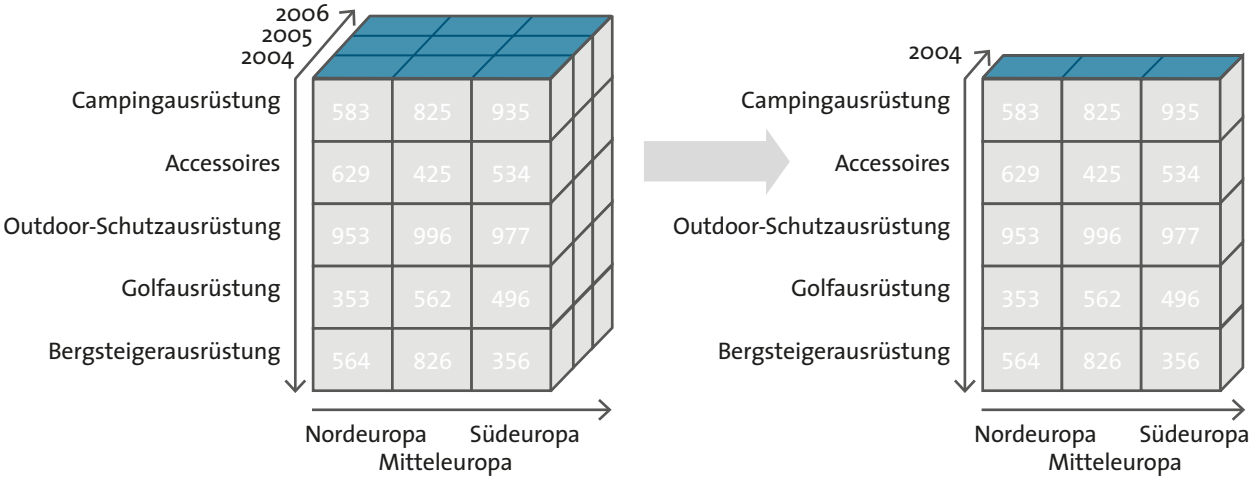


Abbildung 25: OLAP-Würfel zur multidimensionalen Datenanalyse

[96] zum Beispiel de.Wikipedia.org/Wiki/OLAP
[97] Quelle: Wikimedia.org
[98] zum Beispiel als Power Pivot Würfel, siehe www.powerpivot.com

Anwendungsbeispiel

Klout[99] ist ein führender Anbieter für Social Network Analytics und early adopter für Hadoop. Klout berechnet den Einfluss von Millionen von Menschen in sozialen Medien mithilfe einer Lösung für Big Data Analytics. Die Herausforderung besteht darin, ein 800 Terabyte Data Warehouse mit über 1 Billion Datenzeilen für die Ad-hoc-Analyse bereitzustellen. Obwohl Hive Ad-hoc-Abfragen von Hadoop über HiveQL unterstützt, ist die Antwortgeschwindigkeit für die meisten BI-Szenarien unzureichend. Hive bietet ein exzellentes und skalierbares Data Warehouse Framework auf Basis Hadoop. Es ist allerdings nicht die beste Wahl für Ad-hoc-Abfragen großer Daten. Eine bessere Lösung besteht darin, die relevanten Hadoop-Daten mittels HiveQL in einen relationalen Cube zu bringen. Dieser kann die Daten dann für Analysen und Berichte bereitstellen. Auf diese Weise erreicht Klout mittlere Antwortzeiten von weniger als 10 Sekunden für 1 Billion Datenzeilen (vgl. Abbildung 26[100]).

Leider gibt es keine Möglichkeit, eine multidimensionale Analysedatenbank (MOLAP) direkt an eine Hive-Datenquelle zu verbinden. Eine häufige Lösung, initial auch für Klout, ist die Nutzung einer Staging-Datenbank über Sqoop. Der Cube kann dann die Daten von dort importieren. Aber dieser Ansatz hat Nachteile: Er erzeugt zusätzliche Latenz und Komplexität, die die Verwaltbarkeit der Systeme und ihre Kosten beeinflusst.
Mit einem Interface für die direkte Abfrage[101] können die Staging-Datenbank und teure Kopieroperationen vermieden werden. Darüber hinaus kann der Cube quasi

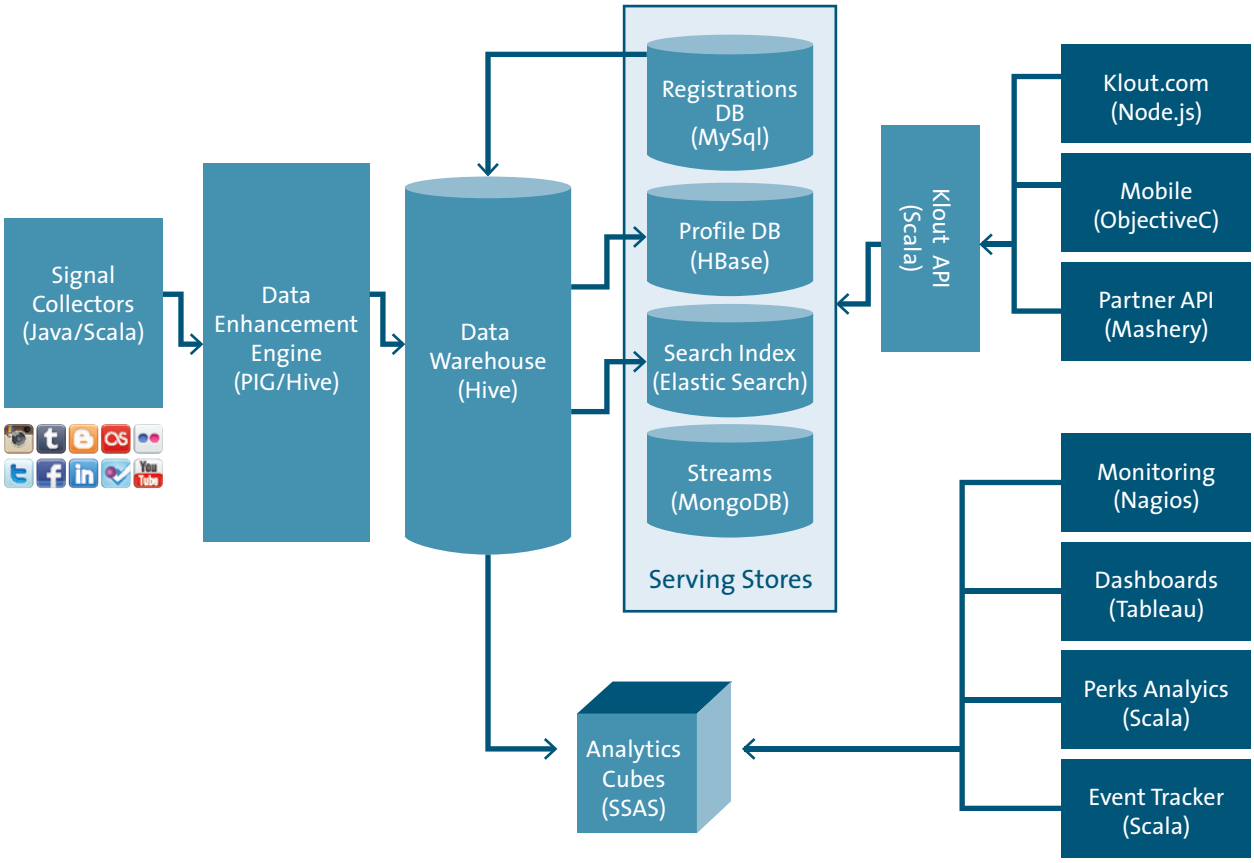


Abbildung 26: Klout-Architektur

[99] www.klout.com, siehe auch http://download.microsoft.com/download/D/2/0/D20E1C5F-72EA-4505-9F26-FEF9550EFD44/MOLAP2HIVE_KLOUT.docx
[100] Quelle: Microsoft
[101] zum Beispiel Microsoft ODBC Treiber für Hive

direkt an Hive verbunden werden, indem Hive-Tabellen als Pseudotabellen in der relationalen Datenbank angezeigt und In-Memory verwendet werden.
Die Integration von relationalem OLAP mit Hive bietet die folgenden Vorteile:

- kosteneffiziente Funktionalität für OLAP und Data Mining für viele Abfrage-Werkzeuge und BI-Anwendungen,
- Nutzung bestehenden Know-hows beim Aufbau großer Cubes mit Milliarden von Datenzeilen,
- Unterstützung für Ad-hoc-Abfragen zum Beispiel aus Excel zur Untersuchung der Klout-Algorithmen,
- optimale Leistung für große Datenmengen, weniger als 10 Sekunden Antwortzeit für 1 Billion Datenzeilen,
- nutzerfreundliche Darstellung des Cube mit Metrik und Dimensionen. Der Cube versteckt die Komplexität sowohl von SQL als auch von Hadoop für den Fachanwender.

Folgende Empfehlungen lassen sich aus der Praxis ableiten:

- Vermeiden Sie traditionelle Datenbanken für Staging-Zwecke. Schreiben Sie die Daten stattdessen in Hadoop, und benutzen Sie dann Hive in Verbindung mit direkten Abfragen und Sichten, um die Daten der Analyse zugänglich zu machen. Dieser Ansatz minimiert Latenzen und vermeidet Redundanz im Warehouse. Er kombiniert die Vorteile von Hadoop für Speicherung und Schreiboperationen, ohne den Komfort des relationalen Daten-Zugriffs aufzugeben.
- Nutzen Sie das Interface für Direktabfragen Ihrer Datenbank für heterogene Joints. Damit können Abfragen an verbundene Server weitergereicht werden und HiveQL Abfragen als Pseudo-Tabellen in ihrer relationalen Datenbank dargestellt werden. Pseudotabellen aus verschiedenen Datenquellen können dann zu Sichten kombiniert werden.
- Nutzen Sie benutzerdefinierte Funktionen (UDF) in Hive, um komplexe Datentypen wie zum Beispiel JSON in Zeilen und Spalten zu konvertieren, die SQL versteht. Durch Hive UDF kann fast jeder unstrukturierte Datentyp in HiveQL gewandelt und der Analyse bereitgestellt werden.
- Spezifizieren Sie die Hive UDF als permanent, um sie von verschiedenen relationalen Abfragen nutzen zu lassen.
- Verwalten Sie große Dimensionen in Hive-Sichten. Um Speicherplatz und Rechenzeit zu sparen, nutzen Sie nur die für die Analyse benötigten Attribute. Wenn Sie über Hive-Sichten die Fakten in den Faktentabellen verlinken, können Sie Dimensionen effizienter limitieren und gleichzeitig Faktentabellen effizienter partitionieren.
- Belassen Sie Hive-Objekte im Standardschema. Sie behalten damit die größtmögliche Flexibilität bei der Auswahl der Abfrage Werkzeuge.

4.4 Visualisierung

Das Ziel einer Datenanalyse ist stets, aus einer Menge an Rohdaten strukturierte Informationen und entscheidungsrelevante Erkenntnisse im gegebenen Anwendungsumfeld bzw. Geschäftsfall zu extrahieren.

Visualisierungen spielen nicht nur eine Schlüsselrolle bei der effizienten Kommunikation relevanter Informationen zu bekannten Sachverhalten im Rahmen des BI-Reportings (vgl. Unterabschnitt 4.3.7).

Vielmehr stellen fortgeschrittene Visualisierungen ein mächtiges und hochgradig flexibles Werkzeug im Analyseprozess dar, das die bisher diskutierten algorithmischen Verfahren der Datenanalyse (vgl. Abschnitt 4.3) im Sinne von »Unsicherheit minimieren durch visuellen Check« entscheidend ergänzt.

Genutzte Techniken abhängig von Aufgabenstellung und Konsument

Beim Einsatz visueller Analyse- und Kommunikationstechniken kommen unterschiedliche Techniken zum Einsatz. Sie lassen sich grob unterscheiden nach:

- Visualisierung zur Informationsbereitstellung versus visuelle Analyse zur Wissensaufbereitung (Erkenntnisgewinn aus Daten),
- Visualisierungsnutzung durch den Fachbereich (Entscheidungsträger) versus durch den Analysten (Domäne-Experten), sowie
- Informationskonsum oder Analyse ortsgebunden (am Arbeitsplatz) versus mobil (Meetings, Dienstreisen).

Die Abbildung 27 skizziert die Zusammenhänge der wichtigsten Schlagworte in diesem Kontext.

Anscombe´s Quartett

Als motivierendes Beispiel für die Relevanz und den Nutzen visueller Repräsentationen im Allgemeinen und visueller Analyse im Besonderen mag Anscombe´s Quartett dienen. Dabei handelt es sich um vier synthetische Punktmengen, bei denen trotz stark unterschiedlicher Verteilungen wesentliche statistische Kennzahlen identisch sind – eine irreführende Eigenschaft, die selbst bei diesen extrem kleinen Datensätzen nur sehr schwer aus der tabellarischen Darstellung abzulesen ist (Abbildung 28, links), während sie in der Visualisierung sofort evident wird (Abbildung 28, rechts).

Big Data: Neue Datentypen benötigen neue Formen der Visualisierung

Im Kontext von Big-Data-Anwendungen steht auch für die Visualisierung zunächst die Herausforderung der

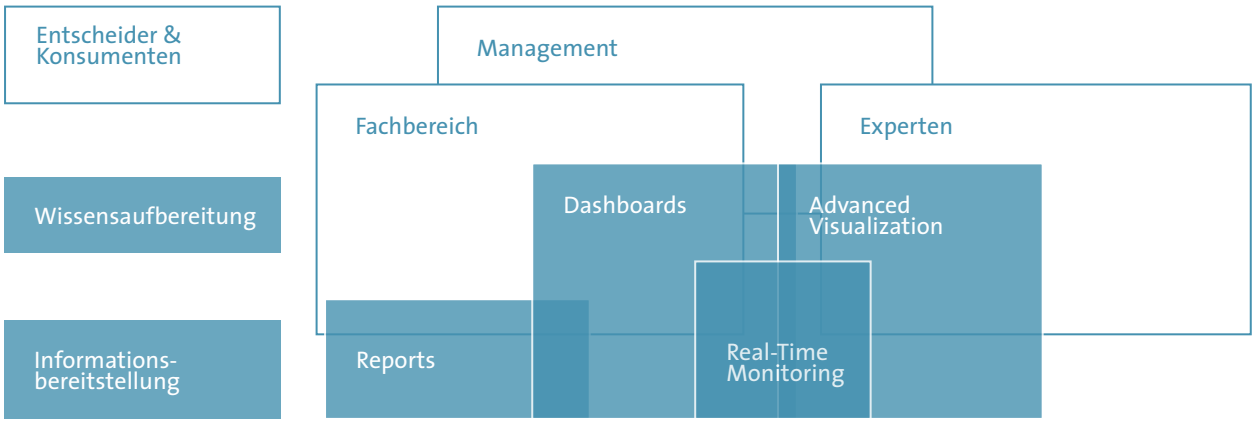


Abbildung 27: Rollen, Ziele und Visualisierungstechnologien im Überblick

I		II		III		IV	
X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10,00	8,04	10,00	9,14	10,00	7,46	8,00	6,58
8,00	6,95	8,00	8,14	8,00	6,77	8,00	5,76
13,00	7,58	13,00	8,74	13,00	12,74	8,00	7,71
9,00	8,81	9,00	8,77	9,00	7,11	8,00	8,84
11,00	8,33	11,00	9,26	11,00	7,81	8,00	8,47
14,00	9,96	14,00	8,10	14,00	8,84	8,00	7,04
6,00	7,24	6,00	6,13	6,00	6,08	8,00	5,25
4,00	4,26	4,00	3,10	4,00	5,39	19,00	12,50
12,00	10,84	12,00	9,13	12,00	8,15	8,00	5,56
7,00	4,82	7,00	7,26	7,00	6,42	8,00	7,91
5,00	5,68	5,00	4,74	5,00	5,73	8,00	6,89

Statistische Eigenschaft	Wert
Mittelwert von X in jedem Fall	9,00
Varianz von X in jedem Fall	11,00
Mittelwert von Y in jedem Fall	7,50
Varianz von Y in jedem Fall	4,122 oder 4,127
Korrelation zwischen X und Y in jedem Fall	0,816
Lineare Regressionslinie in jedem Fall	y = 3,00 + 0,500x

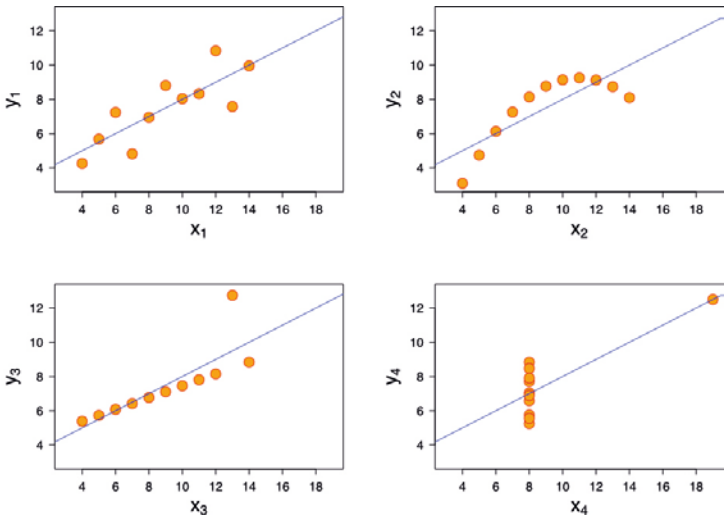


Abbildung 28: Anscombe's Quartett[102]

Skalierbarkeit eingesetzter Technologien hinsichtlich der Dimensionen Volume, Variety und Velocity:

- Die Handhabung großer Datenmengen bei der Erzeugung interaktiver visueller Darstellungen erfordert eine effiziente Integration von Visualisierungsanwendungen mit analytischen Anwendungen sowie leistungsstarke Schnittstellen zu Datenmanagement-Systemen; in sich geschlossene, inselhafte Visualisierungslösungen sind dagegen weitestgehend ungeeignet.

- Die Datenlage setzt sich zunehmend aus einer Vielzahl unterschiedlicher Datenströmen zusammen, die zudem typischerweise aus verteilten Quellen zusammenlaufen[103]. Diese Aggregation und Verdichtung muss auch auf der Präsentationsschicht sichtbar werden.

- Der im Big-Data-Umfeld verstärkt in den Fokus rückende Echtzeitaspekt[104] bedingt gegenüber klassischen statischen Datensätzen neue Ansätze sowohl für Analysen als auch Präsentation.

- Schließlich bergen umfassende Datensammlungen personalisierter oder (re-)personalisierbarer Daten eine nicht zu vernachlässigende Bedrohung für den Datenschutz. Entsprechenden Privacy-Preserving Analysis-Methoden (vgl. Abschnitt 8.1) kommt daher zunehmende Bedeutung zu; visuelle Analysemethoden, welche prinzipbedingt frühzeitig von personenbezogenen Einzeldatensätzen abstrahieren können, stellen gegenüber analytischen Verfahren oftmals einen geeigneteren Ansatz dar, Datenschutzvorgaben umzusetzen.

Überblick

Der Abschnitt 4.4 wird aus den dargelegten Gründen in drei Unterabschnitte eingeteilt (vgl. Abbildung 29).

[102] Abgebildet sind vier Mengen von Datenpunkten, die identische statistische Eigenschaften (Mittelwert, Varianz und Korrelationskoeffizienten), aber dennoch sehr verschiedene Verteilungen aufweisen [Quelle: Wikipedia],
Vgl.: Anscombe, F. J. (1973). »Graphs in Statistical Analysis«. American Statistician 27 (1): 17–21.

[103] Internet der Dinge

[104] wie z. B. dem Monitoring von Echtzeitströmen

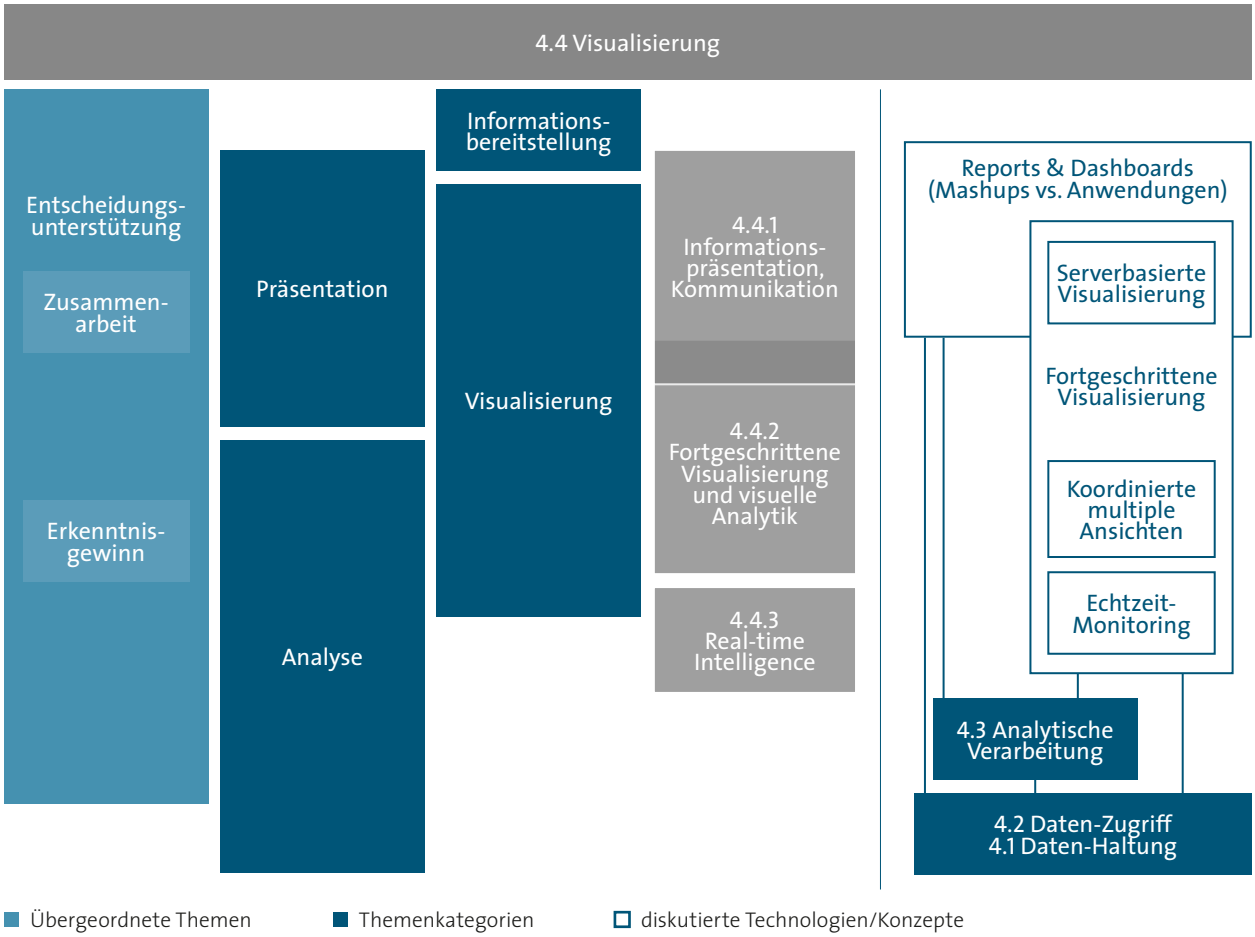


Abbildung 29: Struktur des Abschnitts 4.4

4.4.1 Dashboards

Zielführende Visualisierung

Die zielführende visuelle Aufbereitung von Inhalten hat in den letzten Jahren überdurchschnittlich an Bedeutung gewonnen. Im Fokus von Big-Data-Projekten stehen zeitnahe visuelle Entscheidungsunterstützung und visuelle Analyse. Trotz facettenreicher Vielzahl an verfügbaren visuellen Elementen liegt der angebotene wesentliche Gehalt aktueller Visualisierungstechnologien nicht im einzelnen Element oder einer Anhäufung von Elementen, sondern in der Einbettung von Forschungsergebnissen aus der Entscheidungsfindung und Wahrnehmungspsychologie.

Fehlentscheidung im Cockpit

50 % aller Unfälle lassen sich (lt. Lufthansa 1993) auf schlechte Entscheidungsfindung (Poor Airmanship) zurückführen. Das National Transportation Safety Board (NTSB) ermittelte 1994 bei Unfällen mit Turbojetflugzeugen in den USA:

- Bei 47 % der untersuchten Totalverluste war »falsche oder fehlerhafte Entscheidungsfindung« die Hauptursache.
- Bei 67 % aller Unfälle wurden falsche taktische Entscheidungen gefällt[105].

[105] z.B. unterlassene Entscheidungen trotz klarer Handlungssignale und das Nichtbefolgen von Warnings oder Alerts

Eine reine Signal-Anhäufung im Cockpit genügt nicht. Der Flugzeugführer steht als »Stratege« und »Entscheidungsträger« im Zentrum des Flugablaufs. Damit wird seine Entscheidungsfindung immer wichtiger.[106]

Somit wird ein kleines Dilemma sichtbar. Der technisch versierte Leser kennt die Mittel (Reports, Mashups, Dashboards) seit Jahren. Um ihn dabei nicht mit zeitlicher Einordnung und bekannten Trends alleine stehen zu lassen, werden Visualisierungselemente (vgl. »Technische Sicht: Visualisierung im Zeitablauf«, S. 76), dann wesentliche Anforderungen der Treiber für eine bessere Entscheidungsfindung und die Konsequenzen für die Architektur vor (vgl. »Anforderersicht: Eine Frage des Fokus«, S. 85) vorgestellt. Abschließend folgt die konkrete visuelle Datenanalyse am Beispiel (vgl. »Shneidermann-Mantra: Interaktiv-iterative Visualisierung«, S. 79).

Technische Sicht: Visualisierung im Zeitablauf

Zwei Grundtypen sind am Markt:

- Typen, die proprietäre Visualisierungen gekapselt als Anwendung ausliefern und
- Typen, deren Visualisierungsschnittstellen den offenen (W3C) Standards folgen.

Diese zweite Gruppe erlaubt zusätzlich zur vollwertigen Dashboard-Nutzung die Einbettung der Visualisierung in individuelle Mashups[107].

Das Reporting systembasierter Informationen begleitete die Entwicklung von proprietärer Software seit ihren Anfängen. Im Jahr 2003 wurde ein wesentlicher technologischer Bruch unter dem Schlagwort Web 2.0 publik: Er steht als zeitlicher Eckpfeiler für die stärkere Entkopplung der Visualisierung vom lokalen Arbeitsplatz bzw. lokalen Ressourcen. Im letzten Jahrzehnt entstand eine Vielzahl an interaktiven Elementen. Dabei zeigte sich eine Wiederholung der von monolithischen Lösungen bekannten Abfolge: Nach Listen folgte die Implementierung von Reports, grafischer Bausteine und Dashboards.

Begriffsklärung: Report versus Dashboard
Für beide gilt, dass sie direkt, ad-hoc oder periodisch, regelmäßig oder aufgrund spezifischer Anforderung genutzt werden:

- Reports sind systematische (Detail-) Berichte, die eine analytische Aufbereitung, meist in tabellarischer Form, auch als Mischform aus tabellarischen und grafischen Elementen mit Textblöcken, zur Verfügung stellt. In der herkömmlichen Nutzung werden diese häufig gedruckt bzw. intern via Mail-Anhang verteilt.
- Ein Dashboard (englisch für Instrumententafel) besteht aus einem Arrangement von mehreren visuellen Bausteinen mit dem Ziel diese zu konsolidieren, dabei nötigenfalls zu verdichten und damit relevante Informationen auf einem Schirm im Überblick darzustellen. Gängige Dashboards erlauben eine mehrschichtige Darstellung (Multi-Layer/Linking). Interaktiv kann zwischen einzelnen Schichten navigiert werden und im besten Falle stehen dem Betrachter Selektionen zur Verfügung, die ihn z. B. Zeiträume einschränken oder dargestellte Inhalte variieren lassen.

Die visuellen Bausteine in Dashboards lassen sich in zwei Gruppen einteilen (vgl. Abbildung 30):

- Report-Bausteine, die tabellarische Darstellung von Zahlen bzw. Textsequenzen als Ausschnitt einbetten,
- Visuelle Dashboard-Bausteine, die Informationen verdichtet darstellen und häufig als Navigationspunkt für eine interaktive Detailauswahl dienen.

[106] Nach: Human Factors im Cockpit, Joachim Scheiderer, Hans-Joachim Ebermann
[107] z. B. als Dashboard-Baustein in einem Unternehmensportal

Abbildung 30: Interaktives Dashboard mit sukzessiven Detailsichten in Tableau Software

Wege zum Maßanzug: Mashups ...
Seit 2003 steht das Schlagwort Mashup[108] für einzelne Web-Seiten oder Web-Anwendungen, die aus verschiedenen Quellen relevante Inhalte als Bausteine gemeinsam auf einer (Portal-)Fläche darstellen. Dies geschieht, indem jeder einzelne Bausteine über offene APIs[109] auf eine Datenquelle zugreift, den jeweiligen Inhalt aufbereitet und neben anderen anzeigt.

Als Beispiel für leichtgewichtige Mashups mag www.gelbeseiten.de gelten. Die Ergebnisliste inserierender Firmen wird neben einem Overlay Mashup, der geographischen Karte mit übergelagerten (= overlay) Referenznummern, dargestellt.
Mashups für visuelle Darstellungen nutzen spezielle Entwicklungsumgebungen, um herstellerabhängige Plugin-Technologien[110] oder alternativ offene Standards[111] einzubetten.

... und komfortable Dashboard Software
Anstelle individueller Programmierung erlauben Desktop-Softwareumgebungen die Gestaltung von Dashboards via Drag & Drop. Die Basis liefern Konnektoren, die dafür sorgen, dass Programmierung entfallen kann und direkt visuelle Elemente bedarfsgerecht angezeigt werden. Die Steigerung liefert visuelle Analyse-Software, die diese Entkopplung realisiert und zusätzlich das Shneidermann-Mantra abbildet, wie das auf S. 80 aufgezeigte Beispiel.

Anhand der Anforderungen werden im nächsten Absatz im Bereich serverbasierte Visualisierungen näher betrachtet. Technisch dienen sie entweder als zentraler Auslieferungspool programmierter Mashups oder stellen alternativ vollständige Analysesichten, die in Dashboard-Software erstellt wurden, zur Verfügung.

Anforderersicht: Eine Frage des Fokus

Erkenntnisgewinn zur Entscheidungsunterstützung ist die dominierende Herausforderung. Der Charakter der benötigten visuellen Analyse definiert wesentlich die Architektur. Hier ist zu unterscheiden zwischen

- dem »Erkenntnisgewinn durch Spezialisten (Szenario A)« und
- dem am Markt zu beobachtenden breiten Trend zum »Einsatz im betrieblichen Umfeld (Szenario B)«.

[108] to mash – vermischen
[109] d.h. Programmierschnittstellen, mittels JSON, Ajax, REST,...
[110] wie z.B. Silverlight, AdobeFlash etc.
[111] z.B. HTML5

Szenario A: Fokus Erkenntnisgewinn durch Spezialisten
Komplexe Visualisierungen bereiten spezifische Fragestellungen[112] auf. Eingesetzte fortgeschrittene Visualisierungen werden im Unterabschnitt 4.4.2 näher betrachtet. Gewonnene Erkenntnisse werden Nutzern[113] in der Regel in stark reduzierter Form[114] zur operativen Entscheidungsunterstützung zur Verfügung gestellt. Falls nötig, kann die visuelle Aufbereitung technologisch leicht als Mashup erfolgen.

Szenario B: Einsatz im betrieblichen Umfeld
Bisherige Architekturen zeigen ein großes Defizit: Werden Ad-hoc-Aussagen benötigt, reagieren vorhandene Architekturen unflexibel. (vgl. Tabelle 9[115])
Hauptanforderungen des Fachbereichs zur Verbesserung des Entscheidungsprozesses zielen daher auf größere Handlungs- und Kompetenzrahmen. Der Anforderungskatalog der Self-Service-BI enthält:

- (Empowerment): die Fähigkeit, eigene Analysen durchzuführen und dabei
  - die Entkoppelung vom Zeit- und Ressourcenengpass IT-Entwicklung und Analysespezialist,
  - der flexible breite Zugang zu internen und externen (strukturierten und unstrukturierten) Big-Data-Datenbeständen.
- (Sharing): den Diskurs mit Kollegen durch Teilen der eigenen Analysesichten, darin als Aspekte
  - Erkenntnispotentiale schaffen durch den Austausch von Analysesichten mit der Fähigkeit, dass die Kollegen die Sichten ändern, erweitern, kommentieren und damit anreichern,
  - Absicherung der Erkenntnisse und Entscheidungen mit Kollegen,
  - Persönliche Arbeitsbereiche mit eigenverantwortlich gestaltbarem personenabhängigen Zugang.
- (Communication): die breite Kommunikation der Erkenntnisse (Analysesichten) an Entscheider und Involvierte.

Der Mitarbeiter im Fachbereich wird zentraler Orientierungspunkt der Datenlieferanten (via IT), der Entscheidungsfindungsprozesse (Diskurs mit Kollegen) und der Freigabe an involvierte Dritte (vgl. Abbildung 31).

	Reifeklasse des Unternehmens		
	Best-in-Class (obere 20 %)	Durchschnitt (mittlere 50%)	Nachzügler (untere 30%)
Informationen sind innerhalb der benötigten Zeit vorhanden bei	36%	30%	14%
Häufigkeit die Information innerhalb einer Stunde zu benötigen	53%	21%	32%
Eine Spalte im vorhandenen Report ergänzen (in Stunden)	4,3	5,7	6,9
Einen neues Chart aufbauen (in Tagen)	3,8	5,8	51

Tabelle 9: Kategorisierung von Unternehmen bezüglich Reaktionsgeschwindigkeit im Reporting

[112] z. B. Genomuntersuchungen
[113] z. B. Ärzten
[114] z. B. Medikamentenliste
[115] Quelle: Aberdeen Group, März 2011

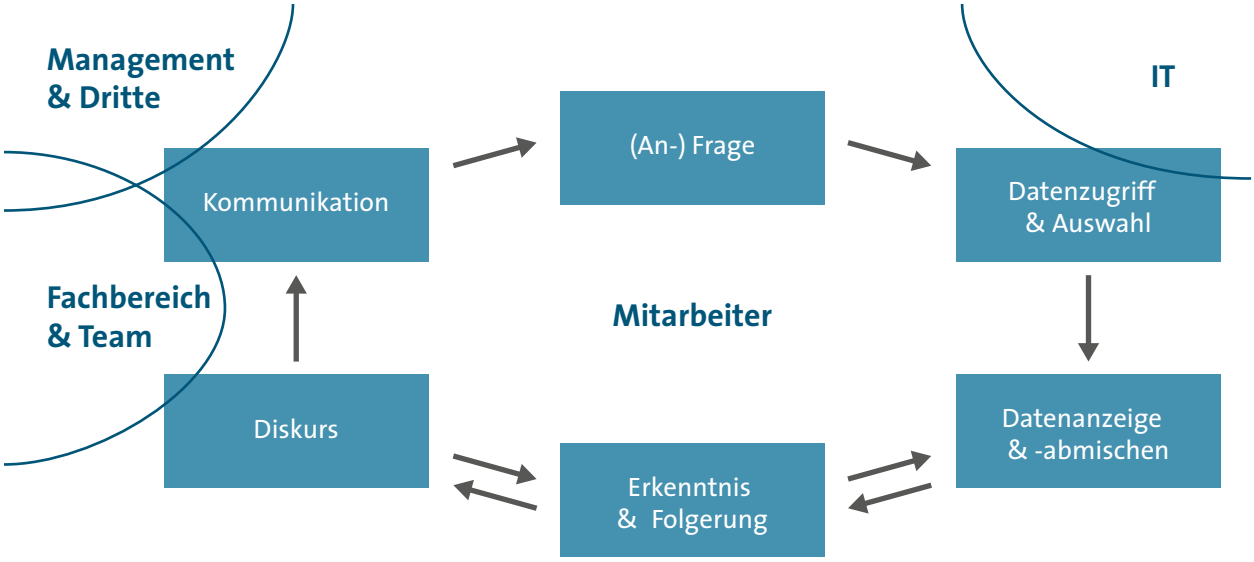


Abbildung 31: Mitarbeiterbezogener Datenanalyseprozess

Serverbasierte Visualisierung als Antwort für den Fachbereich

Diese Anforderungen decken serverbasierte Visualisierung ab, wenn die Architektur die Entkoppelung der Datenabfrage von Visualisierung und Datenquelle realisiert. Dem Fachbereich genügen dann programmierfreie Konnektoren zu Big-Data-Beständen, um die dynamisch relevanten Daten anzuzeigen. Allein aufgrund der Display-Pixelgröße genügen Menschen wenige Daten[116] für eine parallele Anzeige. Dies gilt ebenso für Big-Data-Volumina. Das Netzwerk ist bei dieser Übertragungsanforderung kein Engpass. Es werden die relevanten SQL-/MDX- etc. Abfragen zum Server geschickt. Als Antwort erhält der Client die geforderten Datensätze zur Anzeige.
Die in der Visualisierung dargestellten Aggregationen werden somit nicht lokal berechnet, sondern bestenfalls In-Memory, alternativ hardwarenah in den Analytischen Anwendungen (vgl. Abschnitt 4.3).

Die Kompetenz der Software beruht dabei in interaktiver Ad-hoc-Analyse und hochwertiger Visualisierung. Technologien, die in ihrer Architektur auf diese Entkopplung achten, erhalten dabei automatisch die Flexibilität, auf verschiedene Datentöpfe zuzugreifen.

Shneidermann-Mantra: Interaktiv-iterative Visualisierung

Der Kernansatz der interaktiv-iterativen Visualisierung folgt dem Prinzip »Overview first, zoom and filter, then details-on-demand«[117] (vgl. auch Abbildung 30):

- In einer ersten Ansicht wird zunächst ein stark verallgemeinerter Überblick der Daten dargestellt (Overview).
- Der Anwender kann dann interaktiv für ihn relevante Informationen selektieren, irrelevante Kontextinformationen ausblenden und die Darstellung sukzessive anpassen.
- Für identifizierte Zielinformationen und Zusammenhänge lassen sich schließlich Detailinformationen abrufen, z. B. in Form einer Rückverknüpfung zu den Ursprungswerten in der Datenbank.

[116] maximal wenige Tausend Einzelobjekte dargestellt z. B. im Scatter Plot. Üblicherweise ist die ergonomische Anzahl um Zehnerpotenzen kleiner.
[117] Ben Shneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In Proceedings of the IEEE Symposium on Visual Languages, pages 336-343, Washington. IEEE Computer Society Press, 1996. http://citeseer.ist.psu.edu/409647.html

Die interaktiv-iterative Visualisierung unterstützt damit Hauptanforderungen der Fachbereiche für ihre Arbeit:

- größere Mengen von Informationen zu verarbeiten, d.h. durch Visualisierungstechniken mögliche Muster und Trends zu erkennen (greater volume)
- größere inhaltliche Breite von Informationen zu vergleichen, d.h. durch Gegenüberstellung ähnlicher Ansichten bei verschiedener Auswahl (broader dimensionality)
- die Fähigkeit den Betrachtungswinkel schnell zu wechseln, d.h. um Standpunkte aus verschiedenen Sichten zu prüfen (variable perspective).

Beispiel einer interaktiv-iterativen visuellen Datenanalyse

Anlass der Analyse anhand einer Fragenkette sei eine fiktive Immobiliensuche in Kiel: Der Mitarbeiter startet auf einer Tabelle zu jahresabhängigen Preisen und Eigenschaften einzelner Immobilien. Diese Tabelle (vgl. Tabelle 10) ist lokal in einer Excel-Liste oder auf einem zentralen Datenserver über einen Konnektor zugänglich.

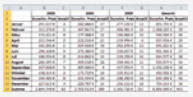
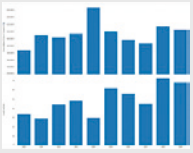
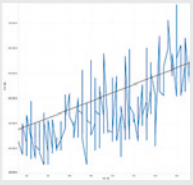
Optional: zur Orientierung wird die zugrunde liegende Tabelle kurz angezeigt.	
Für einen ersten Überblick bringt er den zeitlichen Verlauf als Balkendiagramme zur Anzeige.	
Die erste Detailansicht pro Stadtteil zeigt ihm, welche Lagen grundsätzlich teurer sind, und welche z. B. stärkere Schwankungen auftreten.	
Eine Querprüfung durch Perspektivenwechsel zeigt, in welchen Bereichen der Quadratmeterpreis mit der Fläche korreliert.	
Potenziale werden sichtbar, wenn die Eigenschaften (hier z. B. Anzahl Schlafzimmer) und der Preis betrachtet werden. 3 Schlafzimmer (grün) zeigen ab mittleren Größen einen deutlichen Preissprung!	
Damit rückt eine kleine Auswahl an Objekten in den Fokus, für die dieses Potenzial offen ist.	
Eine weitere Perspektive zeigt die Projektion in die Stadtfläche: die räumliche Einordnung und damit das Umfeld der Immobilie.	

Tabelle 10: Visuell unterstützte Ad-hoc-Analyse, beispielhaft mit Tableau Software

4.4.2 Fortgeschrittene Visualisierung und Visuelle Analytik

Die Herausforderungen im Umfeld von Big Data bezüglich der Dimensionen Volume, Velocity und Variety erfordern nicht nur auf technischer Ebene skalierbare Visualisierungslösungen. In der Praxis sind konkrete Problemstellungen oftmals nur unscharf oder informell formuliert, so dass der Datenbestand zunächst explorativ – d.h. zunächst mehr oder weniger ungerichtet – untersucht und erst im Zuge dieser Exploration die Analyse- bzw. Modellierungsaufgabe konkretisiert wird.

Der Cross-Industry Standard Process for Data Mining (CRISP-DM, vgl. Abbildung 32) beschreibt die unterschiedlichen Phasen der Datenanalyse in einem konzeptuellen Modell: Zu Beginn des Prozesses steht stets die Aufgabe, die involvierten Geschäftsfälle und –prozesse sowie die als Entscheidungsgrundlage dienenden Daten zu sammeln und im Sinne einer Anforderungsanalyse zu strukturieren (Phasen Business Understanding bzw. Data Understanding), wobei sich beide Teilaspekte gegenseitig beeinflussen. Vor der eigentlichen analytischen Verarbeitung der ausgewählten Daten (Phase Modeling) müssen diese im Allgemeinen bereinigt und ggf. ergänzt werden (Phase Data Preparation). Ergebnis der Analyse (Phase Modeling) ist ein Modell im weiteren Sinne, d.h. eine für die Entscheidungsfindung hinreichende Verdichtung der Datenbasis auf die wesentlichen Informationen. Eine nachfolgende Evaluation gegen bekannte Geschäftsfälle (Phase Evaluation) kann wiederum dazu führen, dass das bisherige Verständnis für die Geschäftssituation erweitert und somit wiederum eine verfeinerte Analyse auf einer nochmals verbesserten Datenauswahl vorgenommen wird. Sobald ein Modell als hinreichend ausdrucksstark für die Entscheidungsunterstützung im Geschäftsprozess angesehen wird, kann es z.B. im Rahmen des Reportings oder als Modul in Mashups oder Dashboards visualisiert werden.

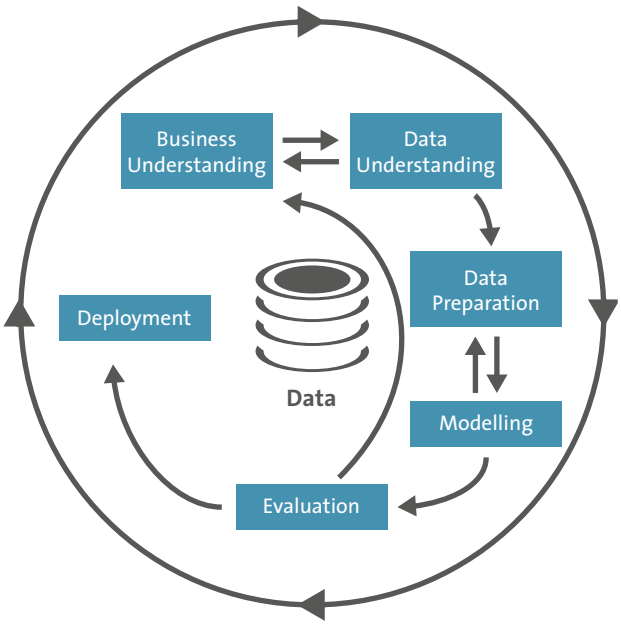


Abbildung 32: Cross-Industry Standard Process for Data Mining

Es ist dabei besonders wichtig zu beachten, dass diese Prozesssicht nicht nur akademischer Natur ist. Vielmehr beinhaltet auch in der Praxis eine Analyse fast immer Versuch und Irrtum: Das Geschäftsverständnis bzw. ein Datenmodell wird erst nach Betrachtung, Bewertung und Einordnung verschiedener (Teil-) Zwischenergebnisse erreicht.

Umso wichtiger sind deshalb Werkzeuge, welche ein solches iteratives Vorgehen in allen Phasen des CRISP-DM unterstützen. Die in Unterabschnitt 4.4.1 (vgl. S. 75) beschriebenen interaktiv-explorative Visualisierungen sind dabei insbesondere in den Phasen Data Understanding und Data Preparation von Bedeutung, während Dashboards in der Deployment-Phase (eines fertig entwickelten Modells) eingesetzt werden. Typische Data-Mining- bzw. Statistik-Lösungen, welche hauptsächlich in den Phasen Data Preparation und natürlich dem Modeling zum Einsatz kommen, realisieren dagegen oftmals einen Black-Box-Prozess – für einen gegebenen Datensatz und vorab festzulegende Parameter wird der komplette (und oftmals rechenintensive) Modellierungsprozess durchlaufen und lediglich das Endergebnis graphisch dargestellt. Ein Fine Tunig (oder gar eine Neumodellierung aufgrund zunächst falscher Annahmen) sind somit

vergleichsweise langen Zyklen unterworfen. Derartige Lösungen skalieren daher schlecht und sind insbesondere im Big-Data-Umfeld nicht für Anwendungen mit einem Fokus auf den Velocity-Aspekt geeignet.
Für die Realisierung komplexer Informationsvisualisierungen ist ein mehrstufiger Prozess notwendig. Dieser wird als Visualisierungspipeline[118] bezeichnet (vgl. Abbildung 33). Obwohl dieses Konzept deutlich vor dem Aufkommen des Themenkomplexes Big Data entwickelt wurde, besitzt es nach wie vor Gültigkeit. Lediglich die auf den jeweiligen Stufen involvierten Technologien haben sich teilweise geändert.

Visualisierungspipeline

Der erste Schritt bei der Datenvisualisierung ist dabei der Einsatz von Filtern, um beispielsweise Rohdaten in ein geeignetes Format zu konvertieren sowie für die Visualisierung irrelevante Datenpunkte oder Attribute zu entfernen. Auch die analytische Vorverarbeitung und die Ableitung statistischer Maße und Aggregate werden im Sinne der Visualisierungspipeline dem Filtern zugeordnet. Auf dieser Stufe findet also ein Großteil der Informationsverdichtung statt – so werden zum Beispiel aus vielen Millionen Einzelmeldungen aus mehreren Produktionsanlagen einige hundert oder tausend relevante Events (vgl. Unterabschnitt 4.2.2).

Als die Visualisierungspipeline 1990 eingeführt wurde, wurde außer in einigen Nischenanwendungen fast ausschließlich direkt in-memory oder mit RDBMS-Backends gearbeitet, weshalb die Filterstufe auch heute noch oft als integraler Bestandteil der jeweiligen Visualisierungslösung betrachtet bzw. implementiert wird. Dies schränkt jedoch die Skalierbarkeit bezüglich des Datenvolumens stark ein. Im Big-Data-Umfeld ist es daher vorzuziehen,

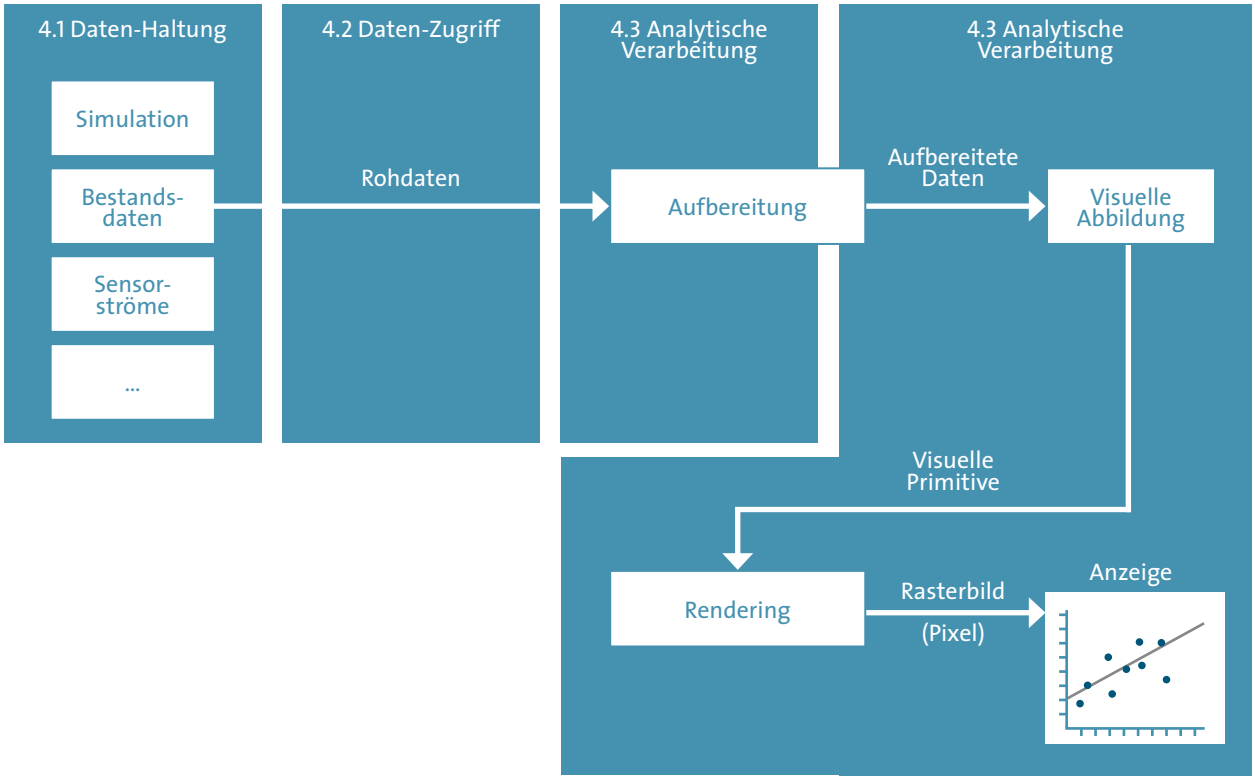


Abbildung 33: Visualisierungspipeline – komplexe Informationsvisualisierung als mehrstufiger Prozess

[118] Haber R. B., McNabb D. A.: Visualization idioms – A conceptual model for scientific visualization systems. In Visualization in Scientific Computing, IEEE Computer Society Press, 1990, S. 74–93.

entsprechende Technologien der Daten-Bereitstellung (vgl. Abschnitt 4.2) und Analytischen Verarbeitung (vgl. Abschnitt 4.3) über geeignete Schnittstellen zu integrieren.

In einem anschließenden Abbildungsmodul (Mapper) werden die Daten dann in eine darstellbare, d.h. geometrische Repräsentation überführt – Punkte, Linien, Flächen (in 2D) bzw. Volumen (in 3D) – in deren Eigenschaften (den sog. visuellen Variablen) wie Position, Größe, Form und Farbe einzelne Datenattribute kodiert werden. Auf dieser Stufe findet neben einer weiteren Informationsverdichtung[119] hauptsächlich eine Informationsgewichtung[120] statt.

Je nach Analyse- bzw. Kommunikationsziel ist es also notwendig, die visuelle Kodierung der Daten bzw. die Visualisierungstechnik adäquat auszuwählen.

Im letzten Schritt der Visualisierungspipeline wird die erzeugte geometrische Repräsentation der Daten von einem Darstellungsmodul (Renderer) in ein 2D-Pixelbild zur Ausgabe auf den verschiedenen Endgeräten[121] umgewandelt. Im Kontext von Big Data steht auf dieser Stufe vor allem die Frage nach der technischen oder Display-Skalierbarkeit – Bildschirmauflösung, Darstellungsverzögerung (und bei Mobilegeräten der Energieverbrauch) müssen berücksichtigt werden.

Benutzerinteraktion

Wie bereits auf S. 78 erläutert, stellt die Benutzerinteraktion einen wichtigen und wesentlichen Freiheitsgrad bei der Analyse und Exploration von Daten dar. Die Visulisierungspipeline erlaubt die Interaktion durch den Benutzer an jeder beliebigen Stelle. Vor allem bei unbekannten Daten ist eine Erkenntnis über die in den Daten vorhandene Information häufig erst durch die interaktive Manipulation der Visualisierungsparameter aller Visualisierungsstufen möglich. Die dafür erforderlichen, hohen Interaktionsraten setzen effiziente Algorithmen und Datenstrukturen sowie entsprechende Verarbeitungskapazitäten auf den darunter liegenden Schichten (Analytische Verarbeitung, Daten-Zugriff, Daten-Haltung) voraus.

Multiple koordinierte Ansichten

Ein wichtiger Aspekt der Flexibilität und (visuellen) Skalierbarkeit von Visualisierungswerkzeugen ist die Fähigkeit, unterschiedliche Aspekte der untersuchten Daten darzustellen.
Es ist oftmals nicht zielführend, die relevanten Zielinformationen in ein einziges Bild zu kodieren, da dies zu überladenen und schwer zu interpretierbaren Darstellungen führt. Gute Visualsierungslösungen bieten deshalb die Möglichkeit, unterschiedliche Teilaspekte in mehreren parallelen Fenstern darzustellen, wobei diese jedoch miteinander koordiniert sind – d.h., interaktive Selektion, Hervorhebungen und Markierungen (Brushing) in einer Ansicht führen zur unmittelbaren Anpassung aller weiteren verbundenen Ansichten (Linking). Auf diese Weise lassen sich

- in den Daten enthaltene Muster bezüglich eines Teilaspektes zuverlässiger und schneller aufspüren, und
- komplexe multi-dimensionale Filter (Zoom and Filter) lassen sich durch die Kombination mehrerer Einzelfilter aus verschiedenen Ansichten einfach und interaktiv definieren.

Die Abbildung 34 zeigt ein Beispiel dieses Ansatzes für die Analyse raumzeitlicher Daten. Die Aspekte Raumbezug (2D-Karte) und zeitliche Dynamik (Zeitgraph) werden jeweils in einer eigenen Ansicht dargestellt, beide Ansichten sind jedoch durch interaktive Hervorhebungen und Selektion (per Maus) miteinander verknüpft.

---

[119] typischerweise auf wenige Dutzend bis einige tausend Graphikprimitive
[120] z. B. werden leuchtende Farben vor einem sonst gedeckten Hintergrund eher wahrgenommen als Variationen der Form in einer großen Anzahl von Einzelsymbolen.
[120] PC-Monitor, Tablet, Smartphone...

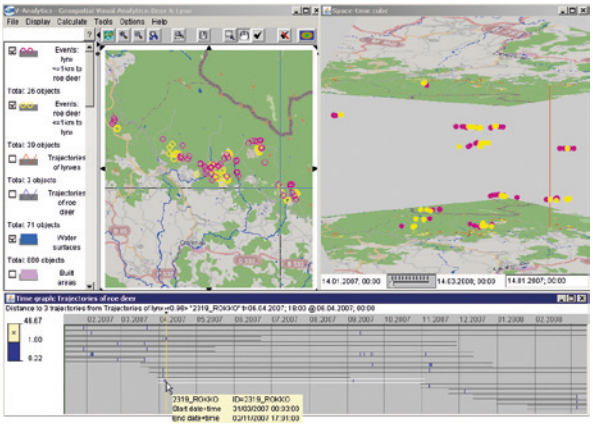


Abbildung 34: Beispiel für multiple koordinierte Ansichten

Die Abbildung zeigt Aspekte von Raum und Zeit von Begegnungen zwischen Rotwild (Beute) und Luchsen (Jäger) – eine 2D-Kartendarstellung für den Raumbezug (oben links), eine Zeitlinienansicht für den Zeitbezug (unten) sowie den kombinierten Raum-Zeit-Bezug im Space-Time-Cube (oben rechts).

Visuelle Analytik

Der Begriff Visuelle Analytik[122] (VA) bezeichnet eine allgemeine Methodik zur Analyse und Erkenntnisgewinnung aus unterschiedlichsten Datenquellen und verschiedenen Anwendungsgebieten.

Es handelt sich bei der Visuellen Analytik also weder um »die eine definitive« Analysemethode, noch um einen völlig neuen Ansatz. VA stellt vielmehr eine konsequente Weiterentwicklung und logische Konvergenz von Ansätzen aus den Disziplinen interaktive Visualisierung, Data Mining, Self-Service BI und maschinelles Lernen dar.
Das Hauptziel von VA-Ansätzen ist dabei die Multiplikation des analytischen Potentials von Mensch und Computer durch eine effektive Kombination interaktiver Visualisierungstechniken mit rechnergestützter Datenanalyse.

Zu diesem Zweck kombiniert VA Methoden und Techniken aus den Disziplinen Statistik, Data Mining, maschinelles Lernen und Modellierung einerseits sowie Visualisierung und Interaktionsdesign andererseits.

Diese Kombination erlaubt eine synergetische Kooperation zwischen dem Analysten und dem Computer, in der beide Seiten ihre jeweiligen Stärken einbringen:

- Rechenleistung des Computers für automatische Verfahren (Statistik, Clusterverfahren, ...) über sehr großen Datenmengen und/oder in Echtzeit, d.h. die enge Integration von Technologien aus den Ebenen Analytische Verarbeitung (vgl. Abschnitt 4.3) und Daten-Zugriff (vgl. Abschnitt 4.2), was seinerseits natürlich eine geeignete Dateninfrastruktur (vgl. Abschnitt 4.1) voraussetzt.

- Menschliche kognitive Fähigkeiten, insbesondere zur intuitiven Mustererkennung, Kreativität, Flexibilität und die Befähigung zum Schlussfolgern sowie zum Querdenken bzw. zu Ad-hoc-Analogieschlüssen, sowie implizites (domänenspezifisches) Hintergrundwissen. Insbesondere in letzterem sind heutige Expertensysteme auch nach jahrzehntelanger Forschung massiv unterlegen.[123]

Visuelle Repräsentationen sind das effektivste Kommunikationsmittel, um Information in das menschliche Bewusstsein zu tragen und die menschlichen Fähigkeiten zur intuitiven Mustererkennung sowie zum Schlussfolgern anzuregen. Interaktive Visualisierungen stellen somit einen hocheffizienten Zwei-Wege-Kommunikationskanal zwischen Mensch und Maschine dar[124].

VA-Ansätze sind deshalb im Allgemeinen gegenüber klassischen (rein statistisch-analytischen) Verfahren überlegen bei:

---

[122] engl. Visual Analysis

[122] Bachmann, R.; Malsch, T. & Ziegler, S. Jürgen Ohlbach, H. (Ed.) Success and failure of expert systems in different fields of industrial application. GWAI-92: Advances in Artificial Intelligence, Springer Berlin Heidelberg, 1993, 671, 77-86

[124] »Visualisation offers a method for seeing the unseen.« Vgl.: B. McCormick, T. DeFanti, and M. Brown. Definition of Visualization. ACM SIGGRAPH Computer Graphics, 21(6), November 1987, p.3
»An estimated 50 percent of the brain's neurons are associated with vision. Visualization [...] aims to put that neurological machinery to work.« [ibid.]

- unbekannten, verrauschten, unvollständigen und/oder widersprüchlichen Daten
- komplexen Problemstellungen mit unscharf oder nur informell definierten Anforderungen
- Untersuchung von Phänomenen in einem komplexen und/oder nur implizit gegebenen Kontext.

Schon das Sehen beinhaltet immer bereits eine Analyse![125]

Visual Analytics Loop

Grundlage eines VA-Ansatzes ist dabei das konzeptuelle Modell der Visual Analytics Loop[126] (vgl. Abbildung 35). Analog zur Visualisierungspipeline werden auch hier Daten in interaktive visuelle Repräsentationen überführt, und analog zum Data Mining werden mittels analytischer Verfahren Modelle der Daten erzeugt.

VA kombiniert diese beiden Aspekte, indem

- Modelle nicht in einem monolithischen Black Box Prozess erzeugt werden, sondern iterativ in mehreren Schritten, mit interaktiver Methoden- und Parameterauswahl bei jedem Schritt, und
- nicht nur die Eingangsdaten, sondern auch die aktuellen Modelle – d.h. insbesondere auch die Teilergebnisse aus Zwischenschritten des Modellierungsprozesses! – visualisiert werden; die Modellvisualisierung dient dabei selbst als interaktives Interface für die Parametrisierung des nächsten Modellierungsschritts.

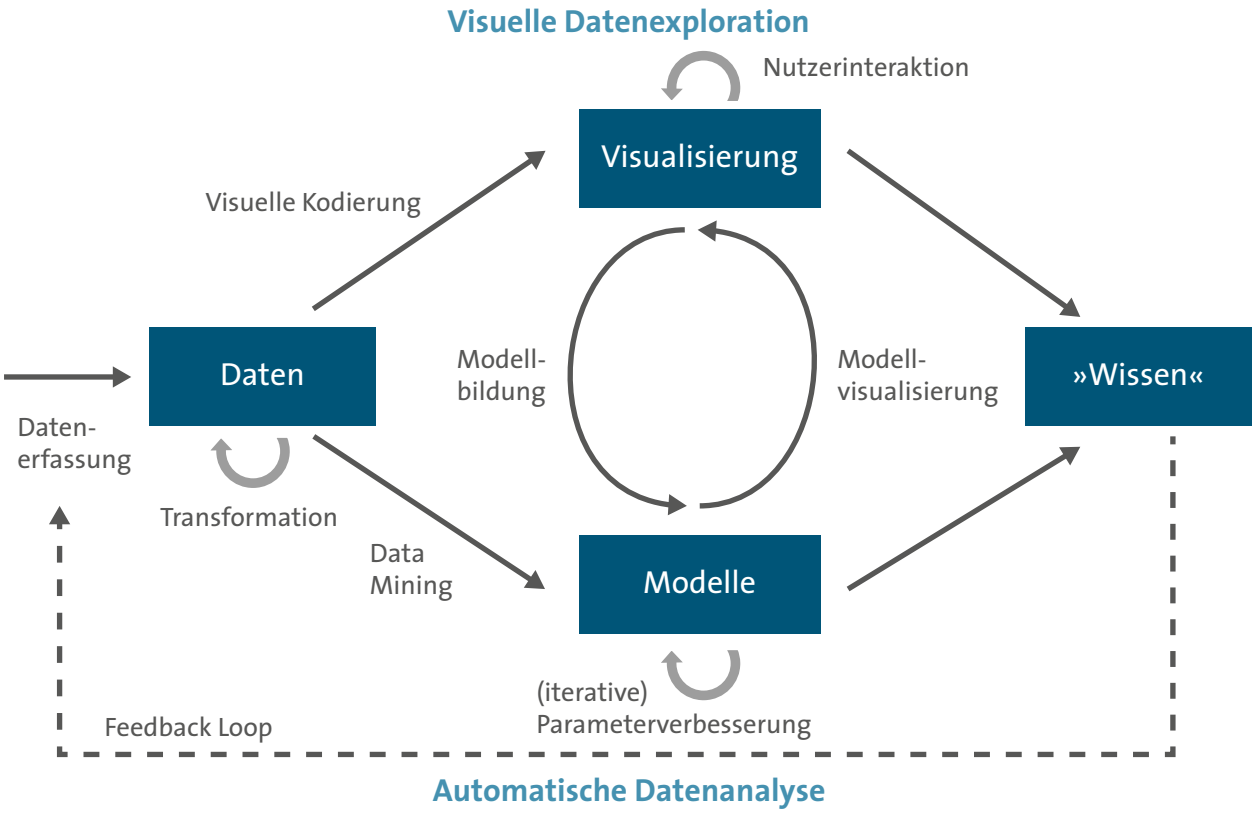


Abbildung 35: Konzeptuelles Modell des Visual Analytics Loop

[125] »An abstractive grasp of structural features is the very basis of perception and the beginning of all cognition.« – R. Arnheim. Visual Thinking. University of California Press, Berkeley 1969, renewed 1997, p. 161
»Detect the expected and discover the unexpected« – J. Thomas and K. Cook. Illuminating the Path. NVAC, 2006. http://nvac.pnl.gov

[126] Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. Visual analytics: Definition, process, and challenges. Springer Berlin, Heidelberg, 2008, 154–175.

Der Analyst kann somit wie erwähnt jedes Zwischenergebnis bewerten, einordnen und den weiteren Analyseprozess dementsprechend neu ausrichten (gerichtete Suche). Jeder Zwischenschritt erzeugt zudem neues oder vertieftes Verständnis über das untersuchte Phänomen im Bewusstsein des Analysten, mehr als es die (visuelle) Bewertung eines einzelnen Endergebnisses es je könnte.

Als zusätzlicher Nutzen der iterativen visuellen Analyse können Zwischenergebnisse, die zu relevanten Entscheidungen für den weiteren Analyseverlauf geführt haben, als »Schnappschüsse« (Checkpoints) gespeichert werden. Ein daraus erzeugtes »visuelles Analyse-Logbuch« kann bei der späteren Kommunikation der Ergebnisse helfen, einzelne Bewertungen belastbar und für Dritte nachvollziehbar zu machen, und so potentiell die Entscheidungsfindung zu verbessern und zu beschleunigen. Gute VA- und Self-Service BI Lösungen sehen oft entsprechende Funktionalität zum späteren Abspielen solcher Schnappschüsse, oft als Storyboard, d.h. in einer vom Analysten bestimmten Auswahl und Reihenfolge sowie individuell annotiert, vor.

Einordnung von Visualisierungswerkzeugen

Der Übergang zwischen einfachen Visualisierungslösungen und fortgeschrittenen Visual Analytics-Frameworks ist dabei fließend. Nicht-interaktive Info-Grafiken eignen sich im Allgemeinen nur für die Kommunikation eines eng begrenzten Sachverhalts, lassen sich aber selbst von Endanwendern sehr schnell mit den allermeisten Standardlösungen (z. B. Excel-Spreadsheet-Visualisierungen) erzeugen. Produkte, die Self-Service-BI-Lösungen bereitstellen, erlauben im Allgemeinen größere Datenmengen dank Anbindung an fast beliebige, skalierbare Back-Ends sowie mehr Flexibilität bei der Erstellung semi-interaktiver, aufgaben- bzw. anwendungsspezifischer Grafiken und Dashboards[127], richten sich typischerweise aber eher an erfahrene Anwender. Werkzeuge zur explorativen visuellen Datenanalyse (EDA) bzw. visueller Analyse sind überwiegend spezialisierte Frameworks, oft aus dem Forschungsumfeld, welche mehrere komplementäre Visualisierungs- und Analysetechniken hoher Komplexität für den professionellen Data Scientist bzw. den Visualisierungsexperten bereitstellen.[128]

Während unterschiedliche Visualisierungstechniken einzelne Aspekte des CRISP-DM adressieren (vgl. Abbildung 32), deckt die Visuelle Analytik als Methodik den gesamten Prozess ab:

- im Sinne der visuellen Exploration (EDA) während der Phasen Business Understanding und Data Understanding
- im Sinne des Visual Debuggings während der Data Preparation, des Modeling und der (visuellen) Evaluation, sowie
- im Sinne des (erweiterten) Visual Reportings zeitnaher, belastbarer und nachvollziehbarer Bewertungen für die Entscheidungsunterstützung.

Im Gegensatz zum klassischen Data Mining betont die Visuelle Analytik dabei ausdrücklich eine iterative Vorgehensweise in kleinen Teilschritten mit sofortiger Evaluation der erzielten Zwischenergebnisse, wie in Abbildung 36 durch die zusätzlichen Rückkoppelungspfeile angedeutet.

[127] z. B. Tableau, Qlikview
[128] Ein bekanntes Beispiel für diese Kategorie sind die Geo-Analysewerkzeuge der ArcGIS Spatial Analysis Workbench

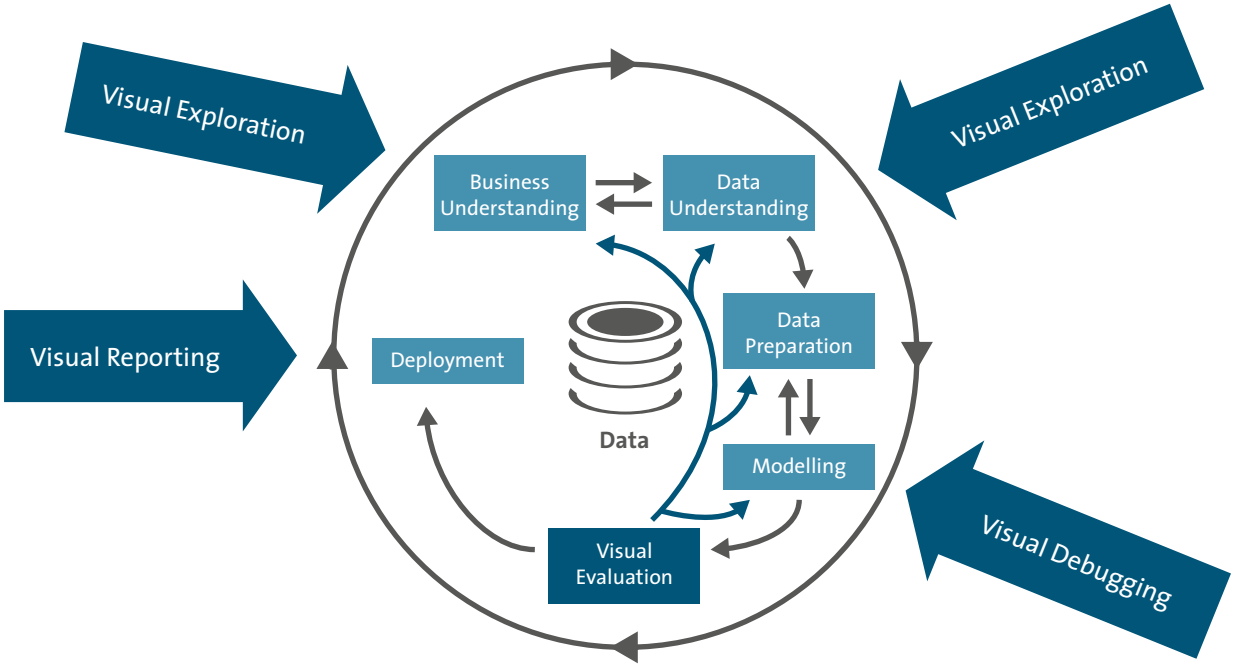


Abbildung 36: Bezug der VA-Methodik zum CRISP-DM

4.4.3 Real-time Intelligence

Das Real-time Intelligence ist insofern ein Spezialfall der Informationsvisualisierung bzw. visuellen Analyse, als hier der Fokus primär auf dem Big-Data-Aspekt Velocity liegt: Es liegen quasi zu keinem Zeitpunkt die aktuellsten Daten[129] vor. Stattdessen erfolgt die Visualisierung bzw. visuelle Analyse kontinuierlich auf den hereinkommenden Streaming-Daten. Typischerweise wird dazu ein gleitendes Zeitfenster aus dem Datenstrom extrahiert[130] und dieses grafisch dargestellt.

Durch das Streaming ändert sich der aktuelle Inhalt des beobachteten Ausschnitts fortwährend, so dass die Visualisierung effektiv eine Echtzeit-Animation der Daten vornimmt. Dies bedingt abhängig von der Aktualisierungsrate des Datenstroms natürlich eine ausreichend schnelle Anbindung an das Daten-Backend sowie eine entsprechend durchsatzstarke Visualisierungspipeline – es besteht hierbei also primär die Herausforderung der Volumen-Skalierbarkeit.

Prädiktive Analytik

Ein weiterer Aspekt des Real-time Monitorings bzw. der konstanten Analyse gesammelter und aggregierter Datenströme ist die Aktualität der im Rahmen der (visuellen) Analyse extrahierten Strukturen und Zusammenhänge sowie der auf dieser Basis erstellten Vorhersagemodellen.

Eine Analyse kann prinzipiell immer nur auf einem »Schnappschuss« der Daten ausgeführt werden. Gleichzeitig können insbesondere Veränderungen des Marktumfelds allgemein (Kontext), aber auch sukzessive Änderungen des modellierten Geschäftsprozesses selbst[131] dazu führen, dass Vorhersagemodelle mit der Zeit ungenauer werden oder neu hinzu gekommene Konstellationen nicht adäquat abdecken.

[129] im Sinne einer Datenbanktabelle
[130] z.B. die letzten X Messwerte aus einem Sensorkanal
[131] z.B. durch Modernisierung von Produktionsanlagen

Prädiktive Analytik ist ein Ansatz, bei dem der Modellierungsschritt der Analyse auch nach der Deployment-Phase (vgl. Abbildung 32, Abbildung 36), d.h. im Wirkbetrieb, in angemessenen zeitlichen Abständen und automatisiert durchgeführt wird. Voraussagemodelle lassen sich so weiter trainieren und passen sich fortlaufend neuen Situationen im Betriebsablauf an.

4.4.4 Zusammenfassung

Die Hauptanforderungen für den Einsatz von Visualisierungen sowie visueller Analyse zielen darauf ab, den Benutzern vier Freiheitsgrade einzuräumen:

- Er kann Informationen mit Kollegen und Dritten »teilen«, lokal, via Intranet, alternativ via Internet (Visualisierung als Kommunikationsmittel).
- Er kann Analyseschritte einzeln/in der Gruppe durchführen (interaktive Visualisierungen als Kollaborationswerkzeug).
- Er kann autonom Inhalte auswählen und zusammenstellen, um diese im Kontext zu analysieren.
- Er kann im Dialog mit den Daten Analyseschritte spontan zu einem Analyse-Workflow zusammenfassen und Ergebnisse zur Kommunikation visuell aufbereiten (Self-Service BI).

Hinzu kommen grundsätzliche qualitative Anforderungen beim Einsatz. Dazu zählen die Use Cases:

- Der Anwender kann neue Objekte hinzufügen (d.h. aus verfügbaren Datenbeständen auswählen und aggregieren).
- Der Anwender kann relevante Elemente auswählen (Informationszugang).
- Der Anwender kann Entwürfe bzw. Zwischenstände seiner Analyse in privaten Bereichen ablegen.
- Der Anwender kann Dritten, Kollegen etc. Zugang zu den Analysen einräumen.
- Dritte können bei Berechtigung die Analysen kommentieren, verändern und anreichern.
- Arbeitsergebnisse können im Fachbereich erzielt werden, ohne dass ein Programmierer nötig ist.

4.5 Daten-Integration

Big Data stellt das Arbeitsfeld Daten-Integration vor neue Herausforderungen – bietet allerdings mit den Hadoop- und Event-Processing-Technologien gleichzeitig eine Plattform, diese Herausforderungen komplementär zu etablierten Integrationslösungen zukünftig in den Griff zu bekommen.
Faktoren wie hohe Datenvolumina und –raten oder unvollständige Schemata führen insgesamt zu einer Neudefinition der Daten-Integration: Weg vom Vorgehen des »Extract-Transform-Load«, hin zu einem »Extract-Load-Transform«. ETL wird zum ELT.

4.5.1 Daten-Konnektivität

Die herkömmlichen Technologien zur Integration von Anwendungen und Daten sind für Big-Data-Lösungen weiterhin notwendig und von Bedeutung. Diese Technologien sind heute im unternehmensweiten produktiven Einsatz in verschiedenen Business-Intelligence-, SOA-, Anwendungs- und B2B-Integrationsszenarien. Eine Wiederverwendung in Big-Data-Architekturen ist notwendig, da die meisten werthaltigen Unternehmensdaten heute in strukturierter Form in existierenden Anwendungen (z. B. ERP, CRM) und Datenbanken vorliegen. Des Weiteren sind die etablierten Technologien schrittweise in Richtung von Big Data Systemen und Szenarien erweitert worden.

Etablierte Integrationstechnologien sind in der Lage, folgende Datenquellen zu integrieren und sie einer Big-Data-Umgebung zur Verfügung zu stellen:

- Datenbank- und Dateien – auf Basis von SQL (ODBC, JDBC) oder nativen Zugriffs- oder Replikationsmethoden (z. B. Change Data Capture)
  - Relationale Datenbanksysteme (RDBMS), z. B. DB2, Oracle, SQL Server
  - Data-Warehouse-Datenbanken
  - Mainframe- und Midrange-Datenbanken, z. B. DB2, IMS, Adabas, VSAM
  - CSV- oder XML-Dateien
- Anwendungen – auf Basis von API-Schnittstellen oder Adaptern
  - Eigenentwicklungen, z. B. Java, .NET, C++, Mainframe (z. B. COBOL)
  - Standardlösungen, z. B. SAP, Oracle, Microsoft
  - Cloud-Anwendungen – SaaS (Software as a Service), z. B. SalesForce, SAP Cloud
- Middleware – auf Basis von technologischen Standardschnittstellen
  - Standardschnittstellen, z. B. Web Services, REST API, Email, XML
  - Messaging-Systeme, z. B. JMS, Websphere MQ, webMethods
- Elektronische Nachrichten – auf Basis von B2B-Adaptern und Schnittstellen
  - EDI und industrie-spezifische Formate, z. B. FIX, SWIFT, ACORD, HL7, HIPAA
  - Email-Systeme.

Enterprise-Service-Bus-Technologien

Durch den Einsatz von Integration-Middleware können diese unterschiedlichen Datenquellen über standardisierte Schnittstellen zugänglich gemacht werden. Hier haben sich Enterprise-Service-Bus-(ESB-)Technologien für die Echtzeit-Integration etabliert. Ein ESB (z. B. webMethods, webSphere, Talend) ist eine robuste, standardkonforme Plattform, die die gängigen Standards für den Datentransfer und für Web Services unterstützt, wie beispielsweise XML, SOAP, REST und JMS. Adapter eines ESB ermöglichen es, Daten aus bestehenden Anwendungen und Systemen schnell, sicher und in Echtzeit auszutauschen, ohne dass die Anbindung an verschiedenen Datenquellen jeweils individuell implementiert werden muss.
Adapter können mit Transformationsregeln kombiniert, mit anderen Adaptern orchestriert und in Services eingebunden werden, um sie über Standard-Schnittstellen bereitzustellen. Durch die Erweiterungen eines ESB in

Richtung einer Event Driven Architecture (EDA) ist man in der Lage, neue Big-Data-Datenströme und -Ereignisse zu verarbeiten, dies wird z. B. durch die Integration von CEP, In-Memory und Low-Latency-Messaging Middleware ermöglicht.

ETL-Plattformen

Die andere wesentliche Technologie zur Anbindung von verschiedenen Datenquellen sind Datenintegrations- oder ETL-Plattformen. Im Unterschied zu einem ESB liegt bei ETL-Plattformen (z. B. Infosphere, Informatica, SAS, Talend) der Fokus auf dem Transfer und der Transformation von großen und komplexen Datenmengen, die häufig im Batch-Verfahren durchgeführt werden. Durch den Schwerpunkt auf Daten bilden bei ETL-Plattformen die Themen Datennormalisierung, Datenqualität und Metadaten eine stärkere Rolle als bei einem ESB. Auch wenn sich ETL- und ESB-Plattformen aufeinander zu bewegen und eine Abgrenzung zunehmend schwierig wird, empfiehlt sich bei Integrationsszenarien mit hohen Anforderungen an die Echtzeit-Verarbeitung und an die Applikationsanbindung (inkl. Datenveränderungen) der Einsatz eines ESB. Andere Technologien wie Daten-Qualität, Daten-Virtualisierung, Master Data Management (MDM), SOA-Governance, API-Management, Business Process Management (BPM) oder spezialisierte Integrationslösungen (z. B. Finanzmärkte) ergänzen diese Plattformen.

Neue Konnektivitäts-Anforderungen

Im Rahmen von Big Data kommen zu den existierenden Integrationslösungen neue Konnektivitäts-Anforderungen hinzu:

- Hadoop – Zugriff auf Daten, die in Hadoop gespeichert sind oder die Integration von Datenquellen mit Hadoop HDFS und MapReduce, z. B. der Import und Export von Daten aus relationalen Datenbanken mit Hilfe von Apache Sqoop.

- NoSQL-Datenbanken (z. B. MongoDB, Apache Cassandra, CouchDB, Neo4J) – NoSQL-Datenbanken stellen unterschiedliche API's zur Datensuche und -veränderung bereit. In den meisten Fällen werden die Operationen auf Basis von Internet-Protokollen (z. B. HTTP, REST), Dokument-orientierten Datenstrukturen (z. B. JSON) und API's für unterschiedliche Programmier- und Skriptsprachen (z. B. Java, C++, .NET, JavaScript, PHP) angeboten.

- Analytische Datenbanken (z. B. IBM Netezza, SAP HANA, Oracle Exalytics, Teradata) – Die Integration erfolgt meistens über Standard-SQL-Schnittstellen (JDBC, ODBC), die zum Teil produktspezifisch erweitert wurden (z. B. Teradata SQL-H).

- In-Memory-Datenhaltungssysteme (z. B. Terracotta BigMemory, Pivotal GemFire) – Integration von In-Memory-basierten Datenspeichern, z. B. auf Basis von API's (z. B. Java Standard JSR107) oder Query-Schnittstellen.

- Cloud-Datenhaltung (z. B. Microsoft Azure, Amazon RDS, Google BigQuery) – Daten, die in Cloud-Datenbanken gespeichert sind, können durch bereitgestellte API's (z. B. REST) oder anbieterspezifische Schnittstellen verarbeitet werden.

- Social Media (z. B. z. B. Facebook, Twitter) – Die Integration mit Social-Media Plattformen basiert auf den jeweiligen API's. Diese API's sind Plattform-spezifisch und unterscheiden sich im Umfang der bereitgestellten Funktionalität, der Datenstrukturen, der Limitierungen (z. B. Datendurchsatz) und Identifizierungsmechanismen.

- Unstrukturierte Daten (z. B. Texte, elektronische Dokumente, Log-Daten) – Diese Daten werden meist über herkömmliche Dateisysteme und Übermittlungsprotokolle (z. B. FTP) zur Verfügung gestellt. Elektronische Dokumente werden in Unternehmen in Dokumenten-Management-Systeme verwaltet und über zugehörige Schnittstellen bereitgestellt.

- Multimedia-Daten, z. B. Audio und Video

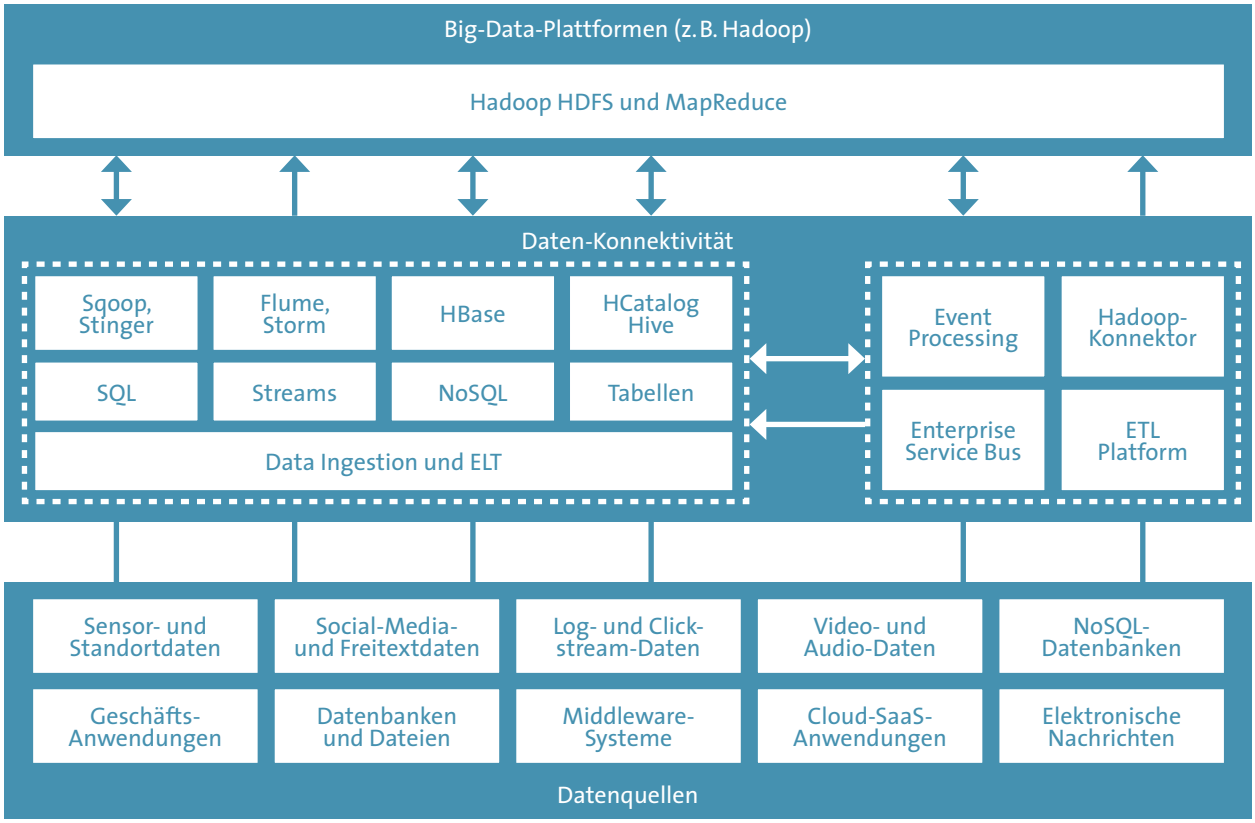


Abbildung 37: Etablierte und neue (grün) Datenintegrationskomponenten im Kontext von Big Data

- Lokationsdaten, z.B. GPS Informationen von mobilen Endgeräten
- Maschinen- und Industrie-spezifische Datenschnittstellen, z.B. Sensoren.

Durch die Verwendung der etablierten Integrationsplattformen und die neuen Big-Data-Anforderungen läßt sich die in Abbildung 37 dargestellte Big-Data-Integrationsarchitektur ableiten.

Hinter den genannten neuen Datenquellen verbergen sich unterschiedliche Datenstrukturen und Datenvolumina, die in unterschiedlichen Geschwindigkeiten einer Big-Data-Plattform bereitgestellt werden müssen. Es lassen sich hierbei folgende Kategorien ableiten, die sich zwischen Massen-Import und Echtzeit-Übermittlung von Daten bewegen:

- **Massen- bzw. Batch-Verarbeitung:** Übertragung von großen Datenmengen über Dateisysteme (FTP) oder Datenbank-Exporte (Entlade-Routinen).
- **Datenströme**: Das kontinuierliche Einlesen von Maschinendaten in Hadoop (z.B. mit Apache Flume) oder die Übermittlung von Nachrichten über Messaging-Middleware (z.B. JMS) zu ESB und CEP-Plattformen.

Integration von unstrukturierten Daten

Es spielt keine Rolle, ob nun Massendaten (z.B. Tagesabzug aller Twitter Feeds) oder Datenströme (z.B. neue Einträge einer Hotelbewertung in ein Reiseportal) verarbeitet werden – der Inhalt eines Datensatzes kann je nach Quellsystem unstrukturiert vorliegen. Um aus diesen semantisch noch nicht greifbaren Datensätzen Inhalte zu erkennen, müssen diese mit geeigneten Werkzeugen

vorab analysiert werden. Die Analyse der Rohdaten mit Einbeziehung evtl. vorhandener Metadaten ergibt dann je nach Verwendungszweck einen Mehrwert wie »negatives Sentiment« (»ungenießbares Frühstück«) für ein bestimmtes Hotel. Technologien aus den Bereichen von Suchmaschinen, maschinenlernender Systeme (vgl. Abschnitt 4.3.7) oder der künstlichen Intelligenz finden hier ihre Anwendung.

Hadoop-Integration mit Echtzeitdaten

Apache Flume[132] ist eine Hadoop-Schnittstelle für die effiziente Erfassung, Aggregation und Transport von großen Datenströmen in das HDFS. Es hat eine einfache und flexible Architektur auf Basis von Datenflüssen (data flows). Flume ermöglicht die Integration von Datenströmen aus mehreren Quellen (z. B. Web-Logs), gleicht die Anlieferungsgeschwindigkeit mit der Verarbeitungsrate ab, gewährleistet die Datenübertragung und skaliert horizontal, um große Datenvolumina verarbeiten zu können.

Hadoop-Integration mit relationalen Datenbanken

Apache Sqoop[133] ist ein Werkzeug für die effiziente Übertragung von Massendaten zwischen Hadoop und strukturierten Datenspeichern, wie relationalen Datenbanken, konzipiert. Sqoop importiert Daten aus Datenbanken in strukturierter Form nach Hadoop (z. B. HDFS, Hive, HBase). Sqoop kann auch verwendet werden, um Daten aus Hadoop zu extrahieren, d.h. Daten können strukturiert in relationale Datenbanken und Data-Warehouse-Systemen exportiert werden. Sqoop stellt Datenbank-Konnektoren bereit (z. B. Oracle, mySQL, SQL Server) oder kann beliebige Datenbanken über Standard-SQL-Schnittstellentechnologien (z. B. JDBC) integrieren.

Relationaler Datenzugriff auf Hadoop mit HCatalog

Apache HCatalog ist eine tabellen-orientierte Zugriffsschicht für Hadoop, die es ermöglicht, Daten zu lesen und zu schreiben. HCatalog ist eine Abstraktionsschicht für HDFS, um Hadoop-Daten über eine relationale Datenstruktur zugänglich zu machen. Die relationale Sicht wird mit Hilfe von Metadaten-Definitionen (Tabellen und Spalten) bereitgestellt und kann dann über Pig oder Hive verwendet werden. Somit können strukturierte Daten aus Dateien oder relationalen Datenbanken einfach in eine Hadoop-Umgebung übertragen werden. Spezialisierte Anbieter und Technologien, wie z. B. Hadapt, Facebook Presto oder Hortonworks Stinger ermöglichen darüber hinaus einen interaktiven SQL-basierten Zugriff auf Hadoop.

Grafische Entwicklungsumgebung für die Hadoop-Integration

Verschiedene Anbieter (z. B. Talend, Syncsort, Datameer) stellen grafische Entwicklungsumgebungen für die Hadoop Integration bereit (vgl. Abbildung 38). Diese Werkzeuge erlauben es, ohne tiefe Hadoop-Programmierkenntnisse Daten zu integrieren und zu analysieren. Mit der Unterstützung von Teamkollaboration und definierten Betriebsverfahren erleichtern sie hierbei typischerweise alle Phasen eines Integrationsprojektes (Design, Dokumentation, Deployment und Monitoring).

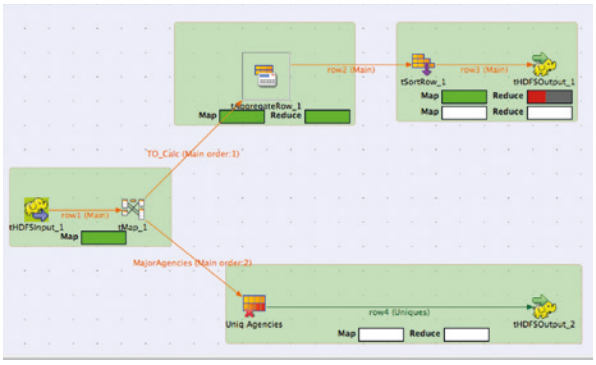


Abbildung 38: Grafische Entwicklung von Hadoop-Integrationsszenarien am Beispiel von Talend

[132] http://hortonworks.com/hadoop/flume/
[133] http://hortonworks.com/hadoop/sqoop/

4.5.2 Data Ingestion – von ETL zu ELT

Ausgangssituation

Schon im Umfeld des klassischen Data Warehousing der Prä-Big-Data-Ära spielte das Thema der Datenextraktion aus unterschiedlichsten Quellen, der Transformation der Daten in die gewünschte Zielformate und das Laden in ein Data Warehouse seit jeher eine gewichtige Rolle in Business-Intelligence-Projekten. Der Begriff Extract-Transform-Load (ETL) steht dabei einerseits für die entsprechenden Softwareprozesse und andererseits für eine breite Palette von Werkzeugen, die diese Prozesse überhaupt erst mit vertretbarem Aufwand ermöglichen.

Die Gestaltung der ETL-Workflows ist in vielen Fällen eine herausfordernde Aufgabenstellung: Daten aus unterschiedlichsten Quellen zu integrieren, dabei durch Datenbereinigungen und Korrekturen die Qualität der gewonnenen Informationen sicher zu stellen und gleichzeitig enge Zeitfenster einzuhalten – dies sind nichttriviale Anforderungen. Die zufriedenstellende Lösung dieser Aufgaben bedingt einige Voraussetzungen:

- gute Kenntnisse der Quellsysteme,
- sauber designte Datenstrukturen in den Ziel-Data-Warehouse-Systemen,
- Erfahrung in der Gestaltung der Ladeprozesse und
- nicht zuletzt Software-Werkzeuge, die gleichzeitig leistungsfähig und effizient in der Anwendung sind.

Daher ist es nicht weiter erstaunlich, dass die Implementierung der ETL-Stränge in DWH-Projekten erfahrungsgemäß leicht einen Anteil von bis zu 70% des Gesamtaufwands ausmachen kann.

Kimball-Methodologie

Diese Situation wird dadurch vereinfacht, dass sich im Laufe der Zeit für das ETL standardisierte Vorgehensweisen herausgebildet haben, die allgemein anerkannt sind und die sehr gut zu den – ebenfalls standardisierten – Datenmodellen passen, die in klassischen Data Warehouses eingesetzt werden. Einer der Vorreiter und maßgebliche Mitgestalter dieser Standards ist Ralph Kimball[134]. Kimball hat in den neunziger Jahren des letzten Jahrhunderts den Begriff der dimensionalen Modellierung geprägt, mit der sich Fragestellungen des Data Warehouse Designs nahezu unabhängig vom Anwendungsgebiet sehr strukturiert lösen lassen. Durch diese Kimball-Methodologie wurde unter anderem eine Vielzahl von Fachbegriffen geprägt, die heutzutage Allgemeingut sind, wie »Dimension«, »Fakt« oder »Langsam veränderliche Dimension« (ein Historisierungsverfahren). Kimball hat aber auch die Handhabung der ETL-Prozesse maßgeblich beeinflusst[135].

Der Einfluss dieser Grundlagenarbeiten geht so weit, dass heutzutage in diversen Datenbanksystemen und ETL-Tools Optimierungen für Kimball-Prozesse eingebaut sind. Ein Beispiel dafür ist die Unterstützung so genannter Star Join Queries in Datenbankystemen – das sind Abfragen, die auf Faktendaten in einem dimensionalen Datenbankschema ausgeführt werden. Ein weiteres Beispiel sind Funktionen für die Verarbeitung langsam veränderlicher Dimensionen in ETL-Tools.

In Data Warehouse-Umgebungen, die auf herkömmlichen RDBMS aufsetzen und in denen es um die Integration von Daten geht, die aus verhältnismäßig einfach strukturierten Geschäftsdatenquellen stammen, funktionieren diese herkömmlichen Ladeprozesse und die entsprechenden Werkzeuge weiterhin sehr gut. Übliche Datenquellen, wie ERP- oder PMS-Systeme und eine Vielzahl individueller betrieblicher Softwaresysteme, liefern Daten in

[134] Vgl. Kimball, Ralph: The Data Warehouse Toolkit, 3rd Edition: The Definitive Guide to Dimensional Modeling. John Wiley & Sons, 2013

[135] In ihrem Buch [Caserta, Joe; Kimball, Ralph: The Data WarehouseETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, 2004] beschreiben die Autoren praxisnahe Verfahren für die Implementierung von ETL-Prozessen, die als Blaupause in DWH-Projekten eingesetzt werden können – und dies weitgehend unabhängig von den verwendeten Werkzeugen.

strukturierten Formaten und überschaubaren Mengen. Die Herausforderungen bei der Integration in ein Data Warehouse sind daher in erster Linie logischer und nicht technischer Natur.

Big Data verändert auch im Anwendungsgebiet ETL so manche liebgewonnene Vorgehensweise und sorgt für neue Herausforderungen, wie auch neue Chancen.

ETL-Herausforderungen durch Big Data

Die Big-Data-Herausforderungen für ETL-Prozesse liegen auf der Hand: Die teils sehr großen Datenmengen und vor allem die Geschwindigkeit, mit der neue Daten generiert werden, erfordern eine hoch performante Plattform für ETL-Prozesse. Dies betrifft in erster Linie die Operationen des Bereinigens und des Umformens der Daten, um sie für Analysen zu erschließen[136]. Software, die für eine SMP-Ablaufumgebung, also die Ausführung auf einzelnen Standardservern, entwickelt wurde, kann hier Probleme bekommen, ausreichend zu skalieren.

Sollen beispielsweise Sensorik-Daten aus einer technischen Großanlage[137] verarbeitet werden, dann kommen leicht mehrere hunderttausend Datensätze pro Sekunde als Eingangsmenge zustande. Solch ein Datenvolumen kann von konventionellen ETL-Tools bei weitem nicht mehr online verarbeitet werden. Die Daten werden daher einfach im ursprünglichen Format abgespeichert – ein typisches Big-Data-Vorgehen.

Natürlich verschiebt sich durch dieses Verfahren das Integrationsproblem einfach nur zeitlich nach hinten: Irgendwann müssen die gespeicherten Daten erschlossen werden und auch die Verarbeitung der ruhenden Datenbestände muss in einem sinnvollen Zeitraum erledigt sein.

Gleichzeitig stellen die neuen nicht-, semi- und multistrukturierten Datenquellen herkömmliche Werkzeuge vor neue Herausforderungen. Klassische ETL-Werkzeuge, sind dafür entwickelt worden, mit strukturierten Daten umzugehen. Für diesen Anwendungsfall sind sie hoch optimiert. Die Processing-Pipelines setzen voraus, dass die Input-Datenstrukturen präzise definiert sind (Fixed Schemas), was in einer Big-Data-Aufgabenstellung häufig nicht möglich ist. Selbst, wenn Datenschemata prinzipiell vorhanden sind, so sind diese nicht selten unvollständig definiert oder inkonsistent. Big-Data-ETL-Werkzeuge müssen daher Daten mit lose definierten und sich ändernden Schemata verarbeiten können. Die Definition einer Datenstruktur kann häufig erst am Ende einer Verarbeitungskette passieren (Late Schemas, vgl. Abschnitt 5.9).

Beispiel Sensorik-Daten

Diese Schwierigkeiten lassen sich wiederum gut am Beispiel von Sensorik-Daten nachvollziehen: Besteht die Aufgabenstellung beispielsweise darin, Diagnosedaten von Kraftfahrzeugflotten zu analysieren, so hat man es mit unterschiedlichsten Formaten aus den verschiedenen Steuergeräten und Wertespeichern der Aggregate zu tun. Diese Formate sind häufig schlecht dokumentiert, je nach Zulieferer gibt es Varianten und je nach Firmware-Stand eines Steuergeräts können sich Änderungen des Schemas von einem zum anderen Auslesezeitpunkt ergeben.

Eine weitere typische Klasse von Big-Data-Daten – Social-Media-Daten – sind sehr strukturiert, wenn es um die Metainformationen geht (Datum, User, Standort etc.). Es werden gut dokumentierte, einfache XML- oder JSON-Formate verwendet. Die Inhalte dagegen (Tweets, Posts, Blogeinträge) sind unstrukturierte Texte, die durch Textanalyseverfahren erschlossen werden müssen. Bestenfalls erleichtern (Hash-)Tags diese Arbeit.

Big-Data-Denkweise – neuer Umgang mit Daten

Doch nicht nur die Datenmengen und Strukturen sind es, die ein Umdenken notwendig machen. Auch der Umgang mit Daten ändert sich im Zuge der Big-Data-Denkweise. In der neuen Welt der Big Data Analytics mit

[136] Das ist das T in ETL: Transform.
[137] z. B. einem Kraftwerk oder einem Windpark

ihren Methoden des Machine Learnings, des Data Minings und der Predictive Analytics ist es wichtig, eine möglichst große Menge des Rohstoffs Daten vorzuhalten. Nur Daten, die über einen längeren zeitlichen Verlauf gesammelt wurden, ermöglichen den Einsatz dieser Verfahren. Gleichzeitig wird bisweilen erst im Verlauf einer Analyse klar, welche der gesammelten Daten wichtig sind und zu einem Ergebnis beitragen. Das notwendige exploratives Vorgehen bedingt, dass die Daten ungefiltert gesammelt werden. Schlussendlich ist es banaler Weise viel einfacher, die Rohdaten zunächst unverändert auf einem preiswerten Speichermedium abzulegen und die Strukturierung und Weiterverarbeitung dann durchzuführen, wenn die endgültige Verwendung geklärt ist.
In einem modernen Data Warehouse werden neben den strukturierten dimensionalen Daten, die klassisch behandelt werden können, gezielt lose strukturierte Daten in einem Langzeitspeicher – typischer Weise einem Hadoop-Cluster – vorgehalten, um diese bei Bedarf zu analysieren. In einem DWH für ein Webshop-System beispielsweise könnten neben den strukturierten Stamm- und Transaktionsdaten[138], die Logdaten der Webserver im Rohformat gespeichert werden. Auf der Grundlage der Protokolldateien lassen sich dann regelmäßig durch Clickstream-Analysen fortlaufende Optimierungsvorschläge für das Webshop-Frontend ableiten. Dabei geht es um große Datenvolumina, die ausreichend schnell verarbeitet werden müssen, und um semistrukturierte Daten[139].

Neudefinition der Daten-Integration

Die beschriebenen Faktoren – hohe Datenvolumina und -Raten, unvollständige Schemata, fehlende Notwendigkeit, in eine dimensionale Zielstruktur zu laden – führen insgesamt zu einer Neudefinition der Daten-Integration: Weg vom Vorgehen des »Extract-Transform-Load«, hin zu einem »Extract-Load-Transform« (ELT)[140]. In einem Big-Data-Umfeld ist der Aufwand für Extraktion (Extract) der Daten häufig vernachlässigbar. Sensorik-Daten werden bereits in großen Mengen geliefert, Social Media-Dienste habe einfach abzufragende APIs, Webserver protokollieren die Benutzerzugriffe sehr umfangreich: Die Daten liegen in Form von Textdateien bereits vor.
Folgt man dem Paradigma des »Speichere jetzt – Verarbeite später«, dann reduziert sich das Laden der Daten (Load) auf einen einfachen Transportvorgang. Eventuell sind dabei Netzwerk-Bandbreiten zu berücksichtigen. Durch Komprimierungsverfahren, Caching und den Einsatz von Cloud-Diensten lassen sich aber auch große Datenmengen und weite Entfernungen gut in den Griff bekommen.

Die Transformationen schließlich werden aufwendiger. Die Verarbeitung wenig strukturierter Daten lässt sich schlechter optimieren und verbraucht deutlich mehr Ressourcen (Prozessor, Speicher, IO). Selbst, wenn die Daten nicht in eine dimensionale Form gebracht müssen, weil es den Anwendungsfall BI nicht gibt, so sind doch auch für statistische und andere Datenanalysen oft umfangreiche Vorverarbeitungen[141] erforderlich. In einer MPP-Umgebung, die viel Rechenleistung zur Verfügung stellt – wie ein ausreichend dimensionierter Hadoop-Cluster – können diese Transformationen auch bei großen Datenmengen noch hinreichend schnell durchgeführt werden.

Anforderungen an das »neue ETL«

Bevor es nachfolgend darum geht, wie Hadoop die Daten-Integration unterstützen kann, folgt an dieser noch eine kurze Definition der drei wichtigsten Anforderungen an das »neue ETL«. Um Big-Data-Anwendungsfälle optimal zu unterstützen, sollten mindestens diese Kriterien erfüllt werden:

- **Performance:**
  Große Datenvolumina und -Raten müssen hinreichend schnell verarbeitet werden können
- **Flexibilität:**
  Nicht-, semi- und polystrukturierte Daten müssen einfach verarbeitet werden können.

[138] Artikel, Kunden, Bestellungen etc.
[139] Webserver sind etwas »unordentlich«, was ihre Protokollierung angeht.
[140] Obwohl der Begriff ELT der passende für diese neue Art des Vorgehens ist, hat es sich doch eingebürgert, beim althergebrachten ETL zu bleiben.
[141] Extraktionen, Bereinigungen, Filterungen, Ersetzung fehlender Werte etc.

- **Effektiv der Handhabung:**
  Die Datenaufbereitungsprozesse sollten möglichst durch komfortable integrierte Entwicklungsumgebungen graphisch gestaltet werden können. Die Verwendung von Skriptsprachen ist nur dann sinnvoll, wenn (langfristig) Entwickler-Know-how vorhanden ist.

Hadoop-Sicht auf ETL: ELT

Die Aufbereitung großer, wenig strukturierter Datenmengen ist ein sinnvoller Anwendungsfall für die Big-Data-Basistechnologie Hadoop. Hadoops HDFS ist ein auf der Hand liegender Speicherort für große Datenmengen – sowohl für Staging-Zwecke, wie auch als Langzeitspeicher. Damit befinden sich die Daten bereits in einer Umgebung, in der potentiell hohe MPP-Verarbeitungskapazitäten zur Verfügung stehen.

Eine hohe Rechenleistung zahlt sich gerade da aus, wo es nicht allein um große Datenmengen geht, sondern gerade auch um komplexe Aufbereitungsmethoden. Die Extraktion von Informationen aus unstrukturierten Formaten, wie Texten oder Blobs (Bilder, Audiodateien) ist rechentechnisch aufwendig und nimmt Prozessoren stark in Anspruch. Hier kann Hadoop seine Stärke der linearen Skalierbarkeit besonders gut ausspielen. Gibt es Aufgabenstellungen, die temporär eine hohe Rechenleistung verlangen, dann lässt sich ein Hadoop-Cluster einfach durch neue Rechenknoten erweitern; besonders leicht geht das, wenn Hadoop als Software as a Service genutzt wird.

Das Transformieren der Daten in einem Hadoop-Cluster nach der Extraktion und dem Laden (ELT) liegt also nahe. Zusätzlich existieren in Hadoop native Technologien, die sich für diese Aufgabenstellung anbieten. Für den Anwendungsfall ETL sind die beiden Technologien Pig und Hive (vgl. Unterabschnitt 4.2.1) besonders interessant. Beide Technologien wurden entwickelt, um Hadoop einfacher in der Anwendung zu machen und die Notwendigkeit zu reduzieren, MapReduce-Jobs direkt in Java zu entwickeln.

4.6 Daten-Governance und -Sicherheit

Daten-Governance und -Sicherheit gewährleisten, dass die verschiedenen Schritte von den Rohdaten bis zur Gewinnung von Erkenntnissen für die Entscheidungsvorbereitung in existierende Technologien, Prozesse und Compliance-Vorgaben großer Unternehmen einbetten.

4.6.1 Daten-Sicherheit

Bei der Umsetzung von Big-Data-Projekten sind die Themen Datenschutz und IT-Sicherheit stark in den Vordergrund gerückt. Die Big-Data-Technologien bieten technologische Optionen an, um die Anforderungen im Bereich der Sicherheit erfüllen zu können.

Verschlüsselung

Verschlüsselungstechnologien werden zum einen im Bereich der Datenspeicherung[142] in der Big-Data-Plattform eingesetzt. Im Umfeld der Hadoop- und DWH-Plattformen werden heutzutage Verschlüsselungsmöglichkeiten angeboten, die durch die Hersteller direkt oder in Kombination mit der Nutzung von Betriebssystem-Funktionalitäten realisiert werden können. Zum anderen ist aber auch der Einsatz von verschlüsselten Kommunikationskanälen zum sicheren Austausch von schützenswerten Daten für die im Fluss befindlichen Daten[143] zu implementieren, die meistens auf SSL/TLS-Funktionen oder unter VPN-Einsatz im Weitverkehrsbereich abgebildet werden.

Multi-Mandantenfähigkeit

Big-Data-Technologie ist multi-mandantenfähig (auch mandantentauglich), wenn sie erlaubt, auf demselben Server oder Cluster oder demselben Software-System – z. B. Hadoop, mehrere Mandanten[144] zu bedienen, ohne dass diese gegenseitigen Einblick in ihre Daten, Jobs, analytischen Modelle, Benutzerverwaltung und Ähnliches haben. Eine Big-Data-Plattform, die dieser Eigenschaft

[142] Data at Rest
[143] Data in Motion
[144] Kunden oder verschiedene Unternehmenseinheiten

genügt, bietet die Möglichkeit der disjunkten, mandantenorientierten Daten-Haltung, Jobausführung, Visualisierung und Konfiguration sowie Entwicklung von Analytischen Applikationen und ihres Customizings.

Data Masking

Beim Data Masking handelt es sich um eine Technologie für die Anonymisierung bzw. Verfremdung von Daten, die mittlerweile auch für Big-Data-Systeme wie Hadoop verfügbar sind. Die eingesetzten Methoden sind somit auch Maßnahmen des Datenschutzes. Data Masking unterscheidet sich von der Verschlüsselung von Daten dadurch, dass es keine 1:1-Abbildung zwischen Originaldaten und verfremdeten Daten geben muss. Zudem bleiben die Daten meist lesbar. Data Masking bezieht sich nicht allein auf personenbezogene Daten und ist daher weiter gefasst als die reine Anonymisierung und Pseudonymisierung von Personen- und Adressdaten. Ziel des Verfremdens der Originaldaten ist die sogenannte Data Leakage Prevention (Verhinderung von Datenlecks). Die Data-Masking-Technologie wird oft zur Verringerung des Risikos von Verstößen gegen die Daten-Sicherheit in nicht produktiven Umgebungen oder zur Erstellung von Testdaten höherer Qualität und Rationalisierung von Entwicklungsprojekten eingesetzt.

Custodian Gateways

Im Bereich der Verwertung und Vermarktung von persönlichen Daten ist es absolut notwendig, die Datenverwertungsmodelle und Konzepte einer Governance zu unterlegen, um im Sinne des Verbraucherschutzes, aber auch der Wirtschaft Möglichkeiten zur Verwertung digitaler Informationen abzubilden. In diesem Umfeld etablieren sich erste Treuhandmodelle und -konzepte (vgl. Abschnitt 8.2).

Identitäts- und Zugangs-Management

Um die Sicherheit von Big-Data-Plattformen und ihrer Softwarekomponenten zu gewährleisten, werden heute übliche Identitäts- und Zugangs-Management-Lösungen mit den Big-Data-Softwaretechnologien integriert.

Diese ermöglichen die Speicherung und Verwaltung der Benutzer, Gruppen sowie die Zugriffsprivilegien auf Daten, Applikationen, Geräte und Systeme. Hierzu werden meist Unternehmens-LDAP Directories wie OpenLDAP und ADS genutzt sowie Identitäts-Management-Systeme zur zentralen Verwaltung und Lifecycle-Management von Benutzern, Gruppen, ihrer Rechte und Zugriffsprivilegien eingesetzt.

4.6.2 Daten-Governance

Unter Daten-Governance versteht man eine Kombination von Prozessen, Technologien und Wissen, mit der sich nachhaltig wertvolle und qualitativ hochwertige Informationen gewinnen lassen. Zur Daten-Governance tragen mehrere Disziplinen bei, die mit ihrem Zusammenwirken den Daten-Lebenszyklus vollständigen abbilden. Fragen wie:

- Woher kommen die Daten?
- Was bedeuten diese Daten?
- Wer trägt die Verantwortung für diese Daten?
- Handelt es sich um datenschutzrechtlich relevante Daten?

werden aus Sicht der IT und der Fachabteilungen beantwortet.

Durch die neue Datenvielfalt und die zunehmende Anzahl von Datenquellen in Big-Data-Projekten ist es notwendig, die Daten eindeutig zu beschreiben. So können nutzenbringende Analysen durchgeführt und Entscheidungen getroffen werden.
Metadaten sind Informationen über Merkmale anderer Daten. Metadaten beschreiben die Daten auf technologischer und fachlicher Ebene. Technische Metadaten sind z. B. der zugrundeliegende Datentyp (numerisch, alphanumerisch) oder ein Ziffernformat (z. B. Kreditkartennummer). Fachliche Metadaten sind z. B. eindeutige betriebswirtschaftliche Feldbeschreibungen. Metadaten können aber nicht nur Daten beschreiben, sondern auch Daten-Integrations- und Datentransformations-Prozesse, um transparent zu machen, wie Daten entstanden sind bzw. verändert wurden.

Mit Data Lineage (Abbildung 39) oder der Datenabstammung beschreiben Metadaten den Prozess, aus welchen ursprünglichen Daten ein aggregierter oder transformierter Wert entstanden ist.

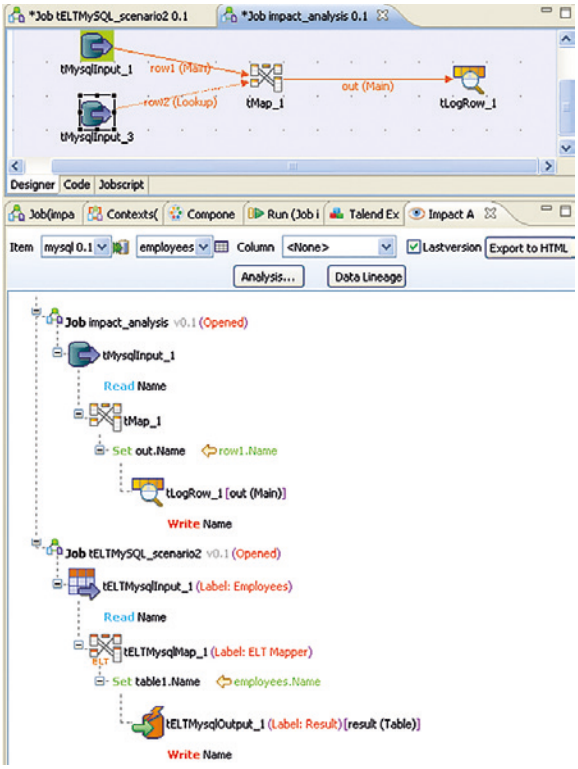


Abbildung 39: Data Lineage – Wo kommen die Daten her

Metadaten-Management-Systeme

Metadaten können in Metadaten-Management-Systemen verwaltet werden, d.h. Metadaten können gespeichert, ausgetauscht, ausgewertet und visualisiert werden. Metadaten-Management ist in klassischen Business-Intelligence-Plattformen eine etablierte Methodik mit bewährten Technologien. Im Kontext von Big-Data-Lösungen sind diese Technologien und Konzepte anzuwenden, aber auch zu erweitern, da z. B. durch den explorativen Big-Data-Analyseansatz meist erst zu einem späteren Zeitpunkt Erkenntnisse über Daten und Datenzusammenhänge gewonnen werden können.

Metadaten bilden eine wesentliche Grundlage für die Datenqualität. Datenqualitätsprozesse helfen Unternehmen primär dabei, nur auf sauberen Daten zu arbeiten, können aber zusätzlich auch Plausibilitäten und Compliance-Regeln abprüfen und somit Risiken reduzieren. Die Datenqualität kann in mehreren Schritten überprüft werden (vgl. Tabelle 11), um die Glaubwürdigkeit, die Nützlichkeit und die Interpretierbarkeit zu verbessern.

Schritt	Erläuterung
1	Mit Daten-Profiling werden Werkzeuge und Mechanismen zur Verfügung gestellt, die es erlauben, schlechte Datenqualität (Inkonsistenzen, Duplikate, etc.) möglichst automatisiert zu erkennen. Die Prüfungen werden anhand von vordefinierten oder individuell definierten Regeln und Metadaten durchgeführt.
2	Die Daten-Standardisierung und Daten-Bereinigung kann z. B. Rohdaten mit Hilfe von kommerziell verfügbaren Referenzdaten angereichert oder ergänzt werden, z. B. mit Informationen aus externen Quellen[145]. Hierbei werden die Rohdaten exakt oder mit entsprechenden Näherungsalgorithmen[146] gegen Referenzlisten[147] verglichen. Je nach geschäftlicher Bedeutung können oder müssen dann Bereinigungsmaßahmen ergriffen werden. Nicht immer sind automatisierte Bereinigungsverfahren[148] möglich, hier müssen dann interdisziplinäre Teams und Fachanwender eingebunden werden.
3	Die Überwachung und Kontrolle der Datenqualität besteht aus einer laufenden Messung, die Aussage darüber liefert, ob sich die Qualität der Daten verbessert bzw. verschlechtert hat.

Tabelle 11: Schritte zur Überprüfung der Datenqualität

145 z. B. Unternehmens-Datenbanken von Dun & Bradstreet
146 z. B. Soundex, Fuzzy
147 z. B. Verbots- bzw. Sanktionslisten
148 z. B. eine Konsolidierung im Kundenstamm

Master Data Management

Master Data Management ist ein Konzept, um die Datenqualität der Geschäftsobjekte[149] eines Unternehmens kontinuierlich auf einem hoen Stand zu halten. Der Einsatz von Master Data Management ist auch im Rahmen von Big-Data-Lösungen vorteilhaft. So reicht es z. B. nicht mehr aus, Kundendaten nur aus eigenen Unternehmensanwendungen zu konsolidieren. Vielmehr sind auch verfügbare Daten von Geschäftspartnern oder aus sozialen Netzwerken einzubeziehen. Ein Multikanal-Vertrieb ist hierfür ein illustratives Anwendungsbeispiel: Er muss Millionen von Kunden-Masterobjekten mit anderen Informationen verknüpfen bzw. anreichern; dazu gehören z. B. die Kundenhistorien aus CRM-Anwendungen, die Nutzungsprofile aus Web-Logs oder die Kundenprofile aus LinkedIn.

4.6.3 Veränderungen in der Data Governance bei Big Data

Im Rahmen von Big Data ändern sich einige der klassischen Regeln und Prinzipen im Bereich der Data Governance (vgl. Tabelle 12). Aufgrund der großen Datenmengen ist es oftmals nicht mehr betriebswirtschaftlich sinnvoll, bestmögliche Datenqualität zu gewährleisten und exakte Ergebnise aus den Datenanalysen zu erhalten. Oft wird eine gewisse Unschärfe in Kauf genommen und ist für die jeweilige Fragestellung auch absolut vertretbar, geht es doch nicht um Finanzbuchhaltung, sondern z. B. Stimmungsanalysen von Kundenmeinungen, Lokalisierung von Ereignissen oder Wahrscheinlichkeiten von Vorhersagen. Ob nun 26-28% der Kunden die Farbe eines neuen Produkts mögen oder ob es genau 27,87% sind, macht keinen wirklichen Unterschied. Während im klassischen BI 2+2 immer 4 ergibt (und das Ergebnis ansonsten falsch), ist ein Ergebnis von ~3.8 in vielen Fällen von Big Data vertretbar.

Traditionelle Data Governance	Big Data Governance
Maximale Datenqualität	→ Vertretbare Datenqualität
Konkrete Antworten Fest Definierte Fragen	→ Wahrscheinlichkeiten → Explorative Analyse
Proprietäre Daten	→ Öffentliche/Web Daten
Daten-Silos	→ Daten-See
Silo-Zugangs-Kontrolle	→ Granulare Zugangs-Kontrolle
Strukturierte Daten	→ Unstrukturierte Daten
Persistente Daten ETL-Prozesse Relationale Datenmodelle	→ Datenströme → ELT-Prozesse → Schemafreie Datenmodelle

Tabelle 12: Neue Aspekte von Data Governance in Big-Data-Szenarien

[149] z. B. Kunde, Produkte, Lieferanten

5 Big-Data-Lösungs-Architekturen und -szenarien

Herkömmliche Lösungen sind angesichts der mit Big Data assoziierten Herausforderungen (»3 V«) sowohl aus technischer als auch aus betriebswirtschaftlicher Sicht eng limitiert.
Hadoop bietet eine Antwort auf diese Herausforderungen und hat sich daher als Kern einer modernen Datenarchitektur und Ergänzung herkömmlicher Lösungen fest etabliert.
Aus dem Zusammenspiel von Hadoop und herkömmlichen Lösungen ergeben sich drei typische Rollen für Hadoop in einer Big-Data-Zielarchitektur – Hadoop als:

- preiswerter Langzeit-Parkplatz für Daten,
- Basis für die Erforschung von Daten,
- unternehmensweite Plattform.

Das Kapitel 5 zeigt, dass es für jedes Einsatzszenario die passende Architektur gibt. Meist bestimmen der Datentyp sowie die Anforderungen an die Verarbeitung die Auswahl der Bausteine in der Zielarchitektur. Daher orientiert sich die Diskussion der Zielarchitekturen an den Datentypen der verschiedenen Einsatz-Szenarien: Clickstream-Daten, Social-Media-Stimmungsdaten, Server-Logdaten, Sensordaten, Maschine-zu-Maschine-Kommunikation, Standortdaten und Freitext-Daten.
Zum Abschluss des Kapitels wird das Zusammenspiel von Big Data und Business Intelligence thematisiert. Ein Beispiel hierfür ist die Entlastung eines traditionellen Data-Warehouses durch Hadoop.

5.1 Warum eine neu entstehende Datenarchitektur für Big Data?

Herkömmliche Datenarchitekturen

Gegenwärtig nutzen die meisten Unternehmen mindestens eine analytische Anwendung zur Unterstützung von Entscheidungen im täglichen Geschäft. Stark vereinfacht sieht die Architektur dieser Lösungen so aus:

- Daten liegen in strukturierter Form vor und stammen überwiegend aus transaktionalen Unternehmensanwendungen wie ERP, CRM oder SCM.
- Die Daten werden in relationalen Datenbanken oder Data Warehouses gehalten. Hierfür werden Rohdaten meist mit dem bekannten Prozess Extract – Transform – Load umgewandelt. Nach dieser Umwandlung werden die Rohdaten meist nach kurzer Zeit gelöscht. Somit finden nur strukturierte und verdichtete Daten Eingang in das Data Warehouse.
- Für die Aufgaben der Analytik und Visualisierung wird Standard-Software wie Business Objects, Hyperion, Cognos eingesetzt, die für das Zusammenwirken mit den transaktionalen Anwendungen (in diesem Fall von SAP, Oracle und IBM) optimiert ist.

Hadoop als Kern einer modernen Datenarchitektur und Ergänzung herkömmlicher Lösungen

Im Unterabschnitt 4.1.1 wurde bereits ausgeführt, wie limitiert herkömmliche Lösungen sind – mit dem Ergebnis, dass die Datenmengen die verfügbaren Budgets übersteigen (vgl. S. 37). Ebenfalls in diesem Unterabschnitt

wurde Hadoop als Werkzeug beschrieben, die Grenzen zu überwinden:

- Hadoop macht Skalierbarkeit bezahlbar. Datenhaltung auf Hadoop ist circa 20x günstiger pro Terabyte als Alternativen wie zum Beispiel traditionelle Speicherlösungen oder Enterprise Data Warehouses.

- Mit Hadoop ist es möglich, Daten erst zu speichern und spontan oder später aufschlussreiche Fragen zu stellen. Erst zum Zeitpunkt der Analyse werden die Daten strukturiert. Techniker nennen dies »Schema on Read«.

Hadoop ist jedoch heute noch kein Ersatz für die traditionellen Datenspeicher im Unternehmenseinsatz, sondern dient als deren Ergänzung. Aus den Hadoop-basierten Big-Data- Projekten, an denen im Jahre 2014 in zahlreichen deutschen Groß-Unternehmen gearbeitet wurde, kristallisieren sich drei Modelle für Zusammenarbeit zwischen Hadoop und herkömmlichen Lösungen heraus:

1. **Hadoop als billiger Langzeit-Parkplatz für Daten.**
   Daten aus unterschiedlichsten Quellen und mit unterschiedlichsten Formaten landen im Hadoop-Cluster, wo sie analog zum ankommenden Rohöl in einer Erdöl-Raffinerie zu Zwischenprodukten verarbeitet werden. Der Prozess Extract – Transform – Load bereitet die dann verdichteten Daten auf ihr Leben in einem Enterprise Data Warehouse vor. Die Rohdaten verbleiben für lange Zeit im kostengünstigen Hadoop-Cluster. Auch alte, schon verdichtete Daten aus dem EDW werden am Ende ihrer Lebenszeit aus Kostengründen wieder Richtung Hadoop ausgelagert. Somit wird Hadoop ein aktives Archiv, die Daten bleiben – anders als bei einer Löschung oder Auslagerung auf Band – weiter im Zugriff. Das Gros der Abfragen und Analysen erfolgt aber weiterhin in herkömmlichen Lösungen.

2. **Erforschung von Daten auf Hadoop.**
   Diese Phase wird sowohl in der Erdölindustrie, als auch in der Business-Intelligence-Gemeinschaft Exploration genannt. Vielfältige Analyse-Werkzeuge laufen direkt auf Hadoop und durchdringen das Datenmeer auf der Suche nach Mustern und Zusammenhängen, bis sie schließlich in der Fahndung nach dem schwarzen Gold fündig werden.

3. **Hadoop als unternehmensweite Plattform.**
   Die zweite Generation von Hadoop ermöglicht es großen Unternehmen, einen firmenweiten Big-Data-Shared-Service anzubieten.

Die Komplementarität von Hadoop mit herkömmlichen Lösungen verdeutlicht die Abbildung 40.

Hadoop etabliert sich als unternehmensweite Plattform

Schon in der ersten Generation hat sich Hadoop rasch zu einer vielversprechenden Plattform entwickelt, um große Datenmengen preiswert zu speichern und skalierbar zu verarbeiten.

Mit der zweiten Generation ermöglicht Hadoop es nun großen Unternehmen, einen firmenweiten Big-Data-Shared-Service anzubieten – also einen gemeinsamen Infrastruktur-Pool, auf dem verschiedene interne und externe Kunden mit ihren oder mit gemeinsamen Daten arbeiten können.

Viele Unternehmen gehen dazu über, Hadoop als einen unternehmensweiten Shared-Service bereitzustellen – oft als »Daten-See« bezeichnet. Der Wert eines solchen Hadoop-Daten-Sees wächst exponentiell, je mehr Daten in diesem See landen und je mehr Anwendungen auf diesen Daten-See zugreifen. Mehr und mehr Daten werden für Jahrzehnte beibehalten (vgl. Abbildung 41).

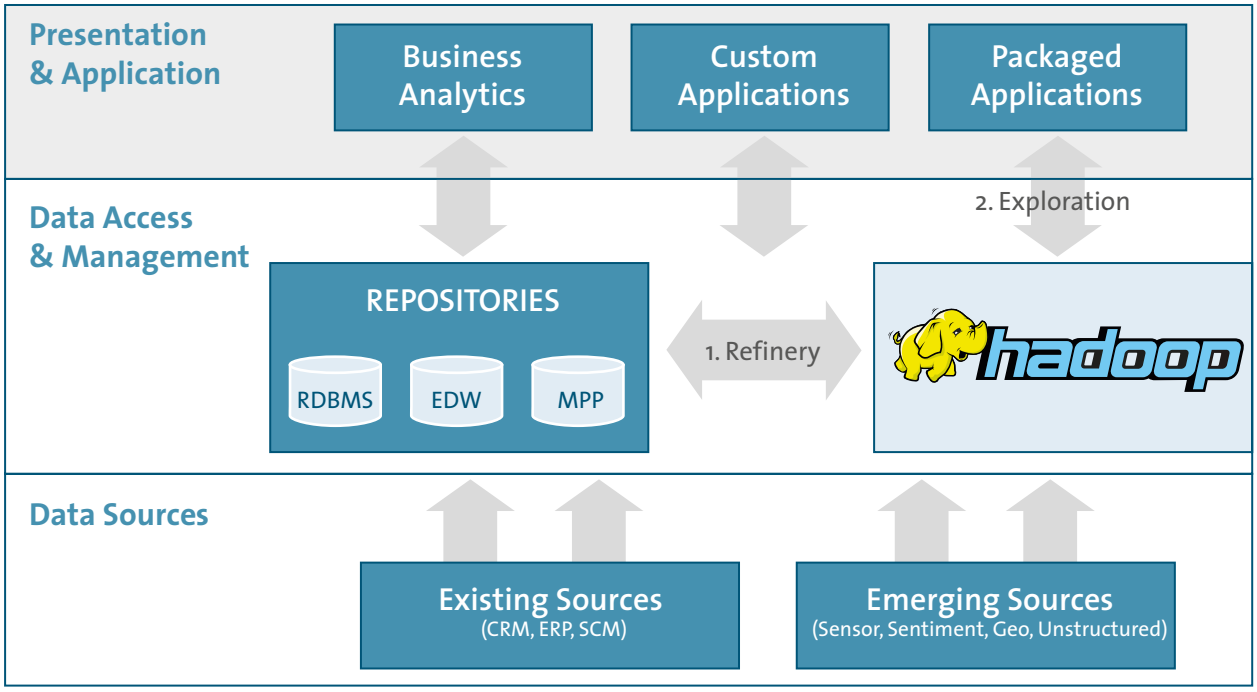


Abbildung 40: Zusammenspiel von Hadoop mit herkömmlichen Lösungen (vereinfacht)

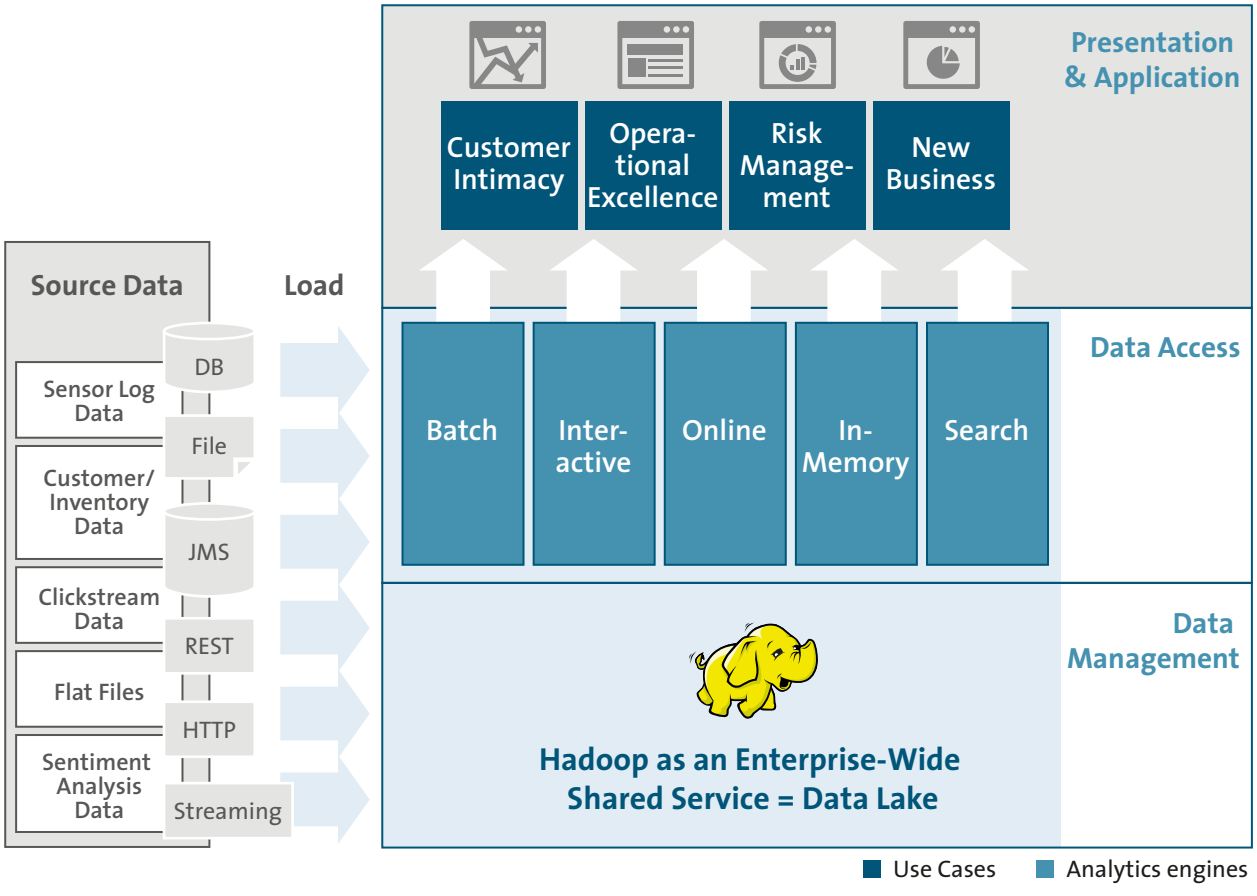


Abbildung 41: Hadoop als unternehmensweite Plattform

Die Reise in Richtung des »Daten-Sees« beginnt oft mit einer Anwendung bzw. einem Einsatzszenario, das Hadoop in der Organisation verankert. Im Zeitverlauf kommen dann weitere Anwendungsfälle hinzu, die zum Beispiel auf größere Kundennähe, Effizienzsteigerungen im Betrieb, besseres Risiko-Management oder neue Geschäftsmodelle zielen. Oft werden folgende Muster sichtbar:

- Wo Fachabteilungen die Reise in Richtung Hadoop angetreten haben, steht oft der Wunsch nach einer 360-Grad-Sicht auf die Kunden im Vordergrund. Durch die Kombination von fragmentierten Datensätzen bietet dabei Hadoop einen Mehrwert. Tiefe und zeitnahe Einblicke in das Kundenverhalten werden möglich.

- Treibt im Unternehmen die IT-Abteilung de Hadoop-Einsatz, so stehen oft die Kosten der Datenhaltung im Fokus. Hadoop senkt die EDW-Ausgaben und Speicherlösungen.

Der Einsatz von Hadoop als unternehmensweite Plattform bringt eine Reihe von Vorteilen. Am wichtigsten ist, dass größere Fragen gestellt werden können. Zu tieferen Einsichten gelangt man, weil jeder autorisierte Benutzer mit dem Pool von Daten in vielfältiger Weise interagieren kann. Mehr Daten führen typischerweise zu besseren Antworten.

Ähnlich einer privaten Cloud-Infrastruktur bewirkt der Daten-See als Shared Service in der Organisation zahlreiche Effekte:

- höhere Geschwindigkeit in der Daten-Bereitstellung,
- schnellere Lernkurve und verminderte Komplexität im Betrieb.
- konsequente Durchsetzung von Datenschutz und -sicherheit sowie Governance,
- verbesserte Kapitaleffizienz im Vergleich zu dedizierten Clustern für jedes Projekt.

Für jedes Einsatzszenario die passende Architektur

Für eine moderne Daten-Architektur bildet Hadoop somit eine kardinal wichtige Grundlage. Auf Hadoop sammeln sich riesige Datenmengen zum Beispiel aus sozialen Medien, Einkäufen im Internet, Weblogs, Videos, Maschinen oder Sensordaten von Geräten. Diese »neuen« Datenquellen weisen alle Merkmale von Big Data auf. Oftmals wurden diese Daten als minderwertige Assets oder sogar als »erschöpfte Daten« betrachtet, deren Speicherung und Auswertung zu teuer war. Aber es sind genau diese Arten von Daten, die von der Datenanalyse zur Big-Data-Analyse führen – mit vielen Einblicken für geschäftliche Vorteile.

Sicher sind diese Arten von Daten streng genommen nicht wirklich »neu«[150], jedoch waren aus heutiger Sicht nie sehr viele vorhanden. Mit Hilfe von Hadoop lernen Unternehmen, diese Arten von Daten als wirtschaftliche, zugängliche und tägliche Quellen für Einsichten und Wettbewerbsvorteile zu betrachten – aus Daten, die sie früher löschten, vernichteten oder auf Band speicherten.

In den Abschnitten 5.2 bis 5.8 dieses Kapitels wird gezeigt, dass es für jedes Einsatzszenario passende Big-Data-Architekturen gibt. Meist bestimmen der Datentyp sowie die Anforderungen an die Verarbeitung die Auswahl der Bausteine in der Zielarchitektur.

[150] Es gibt sie schon seit geraumer Zeit – Textdaten zum Beispiel seit dem alten Ägypten.

5.2 Lösungsszenarien mit Clickstream-Daten

Clickstream-Daten bieten Informationen von unschätzbarem Wert für Internet-Vermarkter. Analysten überprüfen den Clickstream (Datenstrom) auf der Suche danach, welche Webseiten die Besucher aufrufen und in welcher Reihenfolge. Diese Informationen entstehen aus einer Reihe von Mausklicks (dem Clickstream), die jeder Besucher ausführt. Die Clickstream-Analyse kann aufzeigen, wie Nutzer Produkte recherchieren und wie sie ihre Online-Käufe tätigen.

Clickstream-Daten werden oftmals eingesetzt, um die Produktrecherche und Kaufüberlegungen der Besucher einer Website zu verstehen. Mit Hilfe von Clickstream-Analysen können Online-Vermarkter Produktseiten und verkaufsfördernde Inhalte optimieren und somit die Wahrscheinlichkeit erhöhen, dass sich ein Besucher über die Produkte informiert und anschließend auf »Kaufen« klickt. Dank umfangreicher Aufzeichnungen von realen Verhaltensmustern können Online-Vermarkter die Wirksamkeit verschiedener Werbemittel und Calls-to-Action beurteilen – mit der Gewissheit, dass ihre Ergebnisse statistisch signifikant und reproduzierbar sind. So kann es bei bestimmten Produkten sein, dass Videos Besucher häufiger zum Kauf anregen als Whitepaper. Bei einem anderen Produkt ist es dafür möglich, dass ein Whitepaper ein besseres Ergebnis erzielt als ein Datenblatt.

Die Clickstream-Analyse gibt Aufschluss über das Kundenverhalten während des eigentlichen Kaufvorgangs. Mit Hilfe von Verhaltensmustern aus Millionen Verkaufsprozessen können Vermarkter verstehen lernen, weshalb ganze Kundengruppen einen Kaufvorgang an der gleichen Stelle abbrechen. Sie können außerdem sehen, welche Produkte Kunden zusammen kaufen und dann Preis- und Werbestrategien entwickeln, um die Produktpakete zu verkaufen, die ihre Kunden durch ihr Online-Verhalten definieren.

Clickstream-Daten sind jedoch nicht nur für Online-Händler von Konsumgütern geeignet. Die Clickstream-Analyse kann von allen Unternehmen genutzt werden, um zu erfahren, wie gut ihre Website die Bedürfnisse ihrer Kunden, Mitarbeiter oder verbundenen Unternehmen erfüllt.

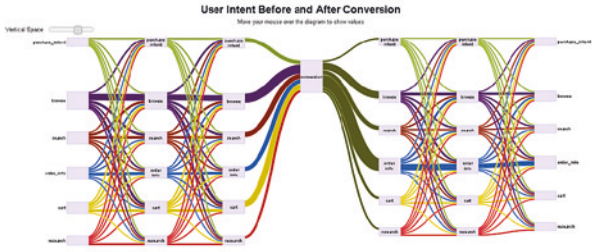


Abbildung 42: Sankey-Diagramm – Darstellung von Besucheraktivitäten[151] auf einer Webseite vor und nach einem Event

Mit Hilfe von Tools wie Omniture und Google Analytics sind Web-Teams bereits jetzt in der Lage, ihre Clickstreams zu analysieren, Big Data fügt jedoch wichtige Vorteile hinzu.

- Bei Big Data werden die Clickstream-Daten mit anderen Datenquellen wie z. B. CRM-Daten zu Kundenstrukturen, Umsatzdaten von stationären Geschäften oder Informationen zu Werbekampagnen verknüpft. Die Kombination mit zusätzlichen Daten bieten häufig wesentlich umfassendere Informationen als eine isolierte, KPI-basierte Analyse des Clickstreams.

- Bei den Mengen an Rohdaten ist eine einfache Skalierung des Big-Data-Systems kardinal wichtig. Hadoop erlaubt die Speicherung der über Jahre gesammelten Daten ohne exponentielle Kostensteigerungen. Das bietet Anwendern die Möglichkeit, eine temporale Analyse oder einen Jahresvergleich des Clickstreams zu erstellen. Anwender können auf der Basis einer Analyse von Daten mehrerer Jahre tiefergehende Muster aufdecken, die der Wettbewerb möglicherweise übersieht.

Zur datenschutzkonformen Erfassung der relevanten Interaktionen von Besuchern einer Webseite bieten sich Pakete wie Celebrus oder Webtrends an. Alternativ können auch die Logfiles der Webserver direkt ausgewertet werden.

[151] Die Dicke der Linien ist proportional zu der Anzahl von Besuchern, die diesen Weg genommen haben.

5.3 Lösungsszenarien mit Social Media Stimmungsdaten

Die Analyse von Meinungen, Stimmungen oder Einstellungen zu Themen oder Produkten war lange Zeit traditionellen Marktforschungsunternehmen vorbehalten. Die Popularität von Social Media hat dazu geführt, dass online große Mengen von unstrukturierten Stimmungsdaten in Social-Media-Einträgen, Blogs, Online-Rezensionen und Nutzerforen von den Nutzern selbst publiziert werden und für die Analyse genutzt werden können.

Mit den traditionellen Marktforschungsinstrumenten wie Befragungen und Fokusgruppen wird gezielt eine Stichprobe der Zielgruppe zu einem Thema befragt und von diesen Ergebnissen verallgemeinert. Die Analyse von Stimmungsdaten setzt dagegen bei einer möglichst großen Datenmenge an und extrahiert hieraus die Inhalte, die der Fragestellung zuzuordnen sind. Die Analyse von Stimmungsdaten kann dabei helfen, die öffentliche Meinung zu einer bestimmten Fragestellung oder die Einstellung der Kunden gegenüber einem Produkt kennenzulernen und die zeitliche Entwicklung zu verfolgen.

Stimmungsanalysen auf der Basis von Online-Daten haben vor allem den Vorteil, dass die Daten nicht aufwändig erhoben werden müssen und immer auf dem neuesten Stand sind. Allerdings lassen sich bestimmte Faktoren nicht mehr kontrollieren. Zum Beispiel ist es schwierig, Aussagen zur Repräsentativität der Beiträge zu machen und Effekte wie Meinungsbeeinflussung und Shit Storms richtig einzuschätzen.

Zunächst müssen die qualitativen Daten extrahiert, strukturiert und angereichert werden, um ein quantifiziertes Ergebnis zu bekommen. Dazu werden Inhalte daraufhin durchsucht, ob sie positive, neutrale oder negative Bewertungen enthalten. Anbieter von Webanalysesystemen unterscheiden sich z. B. dadurch, welche Skalen sie einsetzen oder welche Arten von Stimmungen ausgewertet werden und ob die Analyse automatisch oder in Kombination mit manuellen Auswertungstechniken durchgeführt wird. Eine erste Übersicht gibt der Social Media Monitoring Tool Report 2013[152]. Ein spezielles Sentiment Analysis Tool auf der Basis von SAP HANA hat SAP auf den Markt gebracht.

Eine weitere technische Basis zur automatischen Stimmungsanalyse bietet Hadoop. Hadoop speichert und verarbeitet riesige Mengen an komplexen, unstrukturierten Inhalten. Social-Media-Einträge können mit Hilfe von Apache Flume in das HDFS zum Echtzeit-Streaming geladen werden. Apache Pig und Apache Mahout ordnen die unstrukturierten Daten und bewerten die Stimmungsdaten mit Hilfe von fortgeschrittenen Methoden für Maschinelles Lernen (vgl. Abschnitt 4.3.7).

Nach der Bewertung der Meinungen können die Daten aus den Social Media mit anderen Datenquellen kombiniert werden. Mit Hilfe des HDFS-Datenpools lassen sich CRM-, ERP- und Clickstream-Daten zusammenführen, um Meinungen z. B. einem bestimmten Kundensegment zuzuordnen. Die Ergebnisse können anschließend mit Business-Intelligence-Tools wie Microsoft Excel, Platfora, Splunk oder Tableau veranschaulicht werden (vgl. Abbildung 43).

Eine erweiterte Analyse ist die Ermittlung von Emotionen[153] wie Freude, Ärger, Gefallen, Sorge. Die Verfahren und Komponenten der Analyse werden hier am Beispiel der PKW-Diskussion skizziert.

Die Analyse extrahiert aus einem großen Nutzerforum[154] die in den Beiträgen geäußerten Emotionen, um ein Stimmungsbild für ein Fahrzeug und dessen Teilaspekte[155] zu erschließen. Auf dieser Basis können aus dem Benutzerforum quantitative Maße z. B. für die Produktzufriedenheit ermittelt werden.

[152] http://www.goldbachinteractive.com/aktuell/fachartikel/social-media-monitoring-tool-report-2013
[153] www.emotionsradar.com
[154] www.Motortalk.de
[155] Verbrauch, Robustheit,...

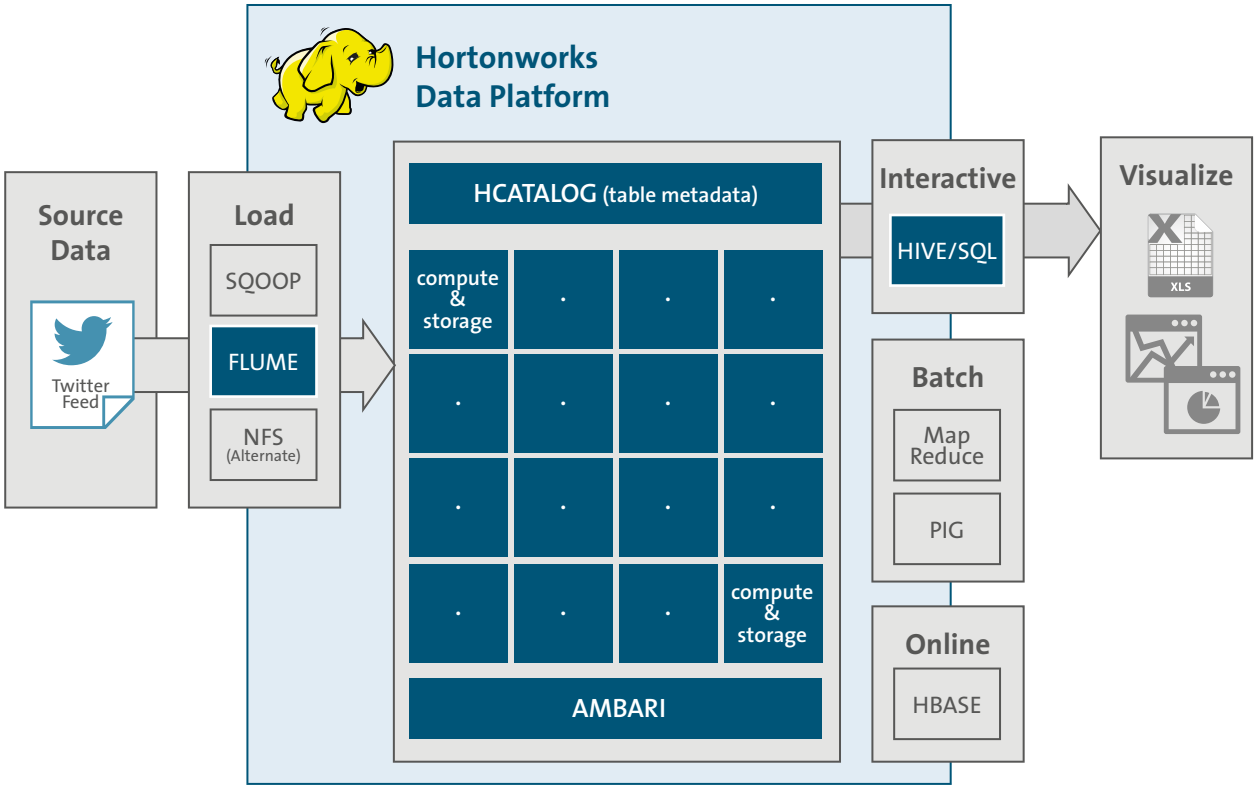


Abbildung 43: Anwendung der Hortonworks Data Platform für die Analyse von Twitter-Daten

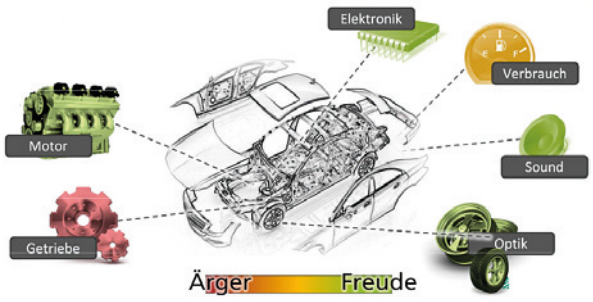


Abbildung 44: Beispiel-Szenario – Emotions-Analyse bei PKW

Im PKW-Beispiel (vgl. Abbildung 44) wird der Ausgangstext zusammen mit allen Annotationen in einem UIMA-Dokument gespeichert. Mittels Apache UIMA kann eine Sequenz von Buchstaben oder Worten durch eine Annotation gekennzeichnet werden, z. B. wird eine Wortfolge als Nennung eines Autotyps annotiert. Darüber hinaus können auch Relationen zwischen Annotationen im Text markiert werden, z. B. Ärger/<Automarke>/ <Fahrzeugtyp>/Verbrauch. UIMA erlaubt einen schnellen Zugriff auf diese Annotationen und kann sie als XML-Dokument auslagern. Die Schritte bei der Sprachverarbeitung und Analyse sind in Tabelle 13 aufgelistet.

Verarbeitungsschritt	Beispiel Emotions-Analyse
Download, Bereinigung und Normalisierung	Beiträge zu ausgewählten Automarken downloaden, HTML-Code entfernen.
Anreicherung mit Metadaten	Automarke, Datum, Diskussionsthread.
Sprachenerkennung	nicht notwendig, da rein deutsch-sprachige Beiträge
Satzsegmentierung	im Einsatz
Wortstammreduktion	-
Lemmatisierung	-
POS-Tagging	(Wortartenerkennung) im Einsatz
Parsing	-
Koreferenzauflösung	-
Eigennamen-Erkennung	Orte und Personen (welche oft Aliasnamen haben)
Domänenspezifische Eigennamen	Automarken und –typen, Einzelteile eines PKWs, Werkzeuge
Topic-Modell	im Einsatz
Phrasenextraktion	Ermittlung von Phrasen die Freude, Ärger, Sorge ... ausdrücken
Relationsextraktion	Zusammenhang zwischen Auto/Autoteil und Emotionsausdruck

Tabelle 13: Schritte der Sprachverarbeitung am Beispiel Motortalk

Das Ergebnis der Analyse ist ein Überblick über die aufgetretenen Emotionen im Zusammenhang mit einer bestimmten Automarke und/oder einem zugehörigen Bauteil. Außerdem lässt sich auch erschließen, wie häufig die entsprechenden Ausdrücke genannt wurden, was ein Indiz für die Relevanz eines bestimmten Themas ist.

5.4 Lösungsszenarien mit Server-Logdaten

Server-Logdaten-Management beschreibt den Umgang mit großen Volumen an computergenerierten Logausgaben. Es umfaßt das Sammeln, die zentrale Aggregation, die Langzeitspeicherung, die Analyse der Logdaten (in Echtzeit und im Batch-Betrieb) sowie die Suche in den Logdaten und daraus generierte Berichte.
Die Quellen für Logdaten sind vielfältig und reichen von Betriebssystemen und Applikationen über Netzwerkgeräte bis zu Sicherheits- und Überwachungseinrichtungen. Mit der Virtualisierung von Servern kommen die Host-Systeme als neue Bestandteile der IT-Landschaft hinzu, die ebenfalls große Mengen an Logdaten generieren. In der Summe erhöhen sich damit nicht nur die Logvolumina erheblich, sondern auch die Zahl der Logformate nimmt in gleicher Weise zu. Die Möglichkeit einer sinnvollen manuellen Auswertung und Analyse wird dagegen immer geringer. Dies gilt besonders dann, wenn Logdaten miteinander korreliert werden müssen, um Beurteilungen

einer Situation oder Auswertungen von Fehlerzuständen vorzunehmen.

Die Herausforderungen einer effizienten Verarbeitung ergeben sich zum einen aus den Logvolumina, die in größeren Organisationen mehrere 100 GB pro Tag betragen können, und zum anderen aus der Vielfalt der zu bearbeitenden Logformate.

Durch die Zentralisierung werden Daten verfügbar und im Zusammenhang auswertbar gemacht und bildet daher die Grundlage für die Überwachung und Kapazitätsplanung des IT-Betriebs genauso wie für die Durchführung von Sicherheitsmaßnahmen oder die Protokollierung der Einhaltung von Richtlinien und Vorschriften.

Verdächtige Veränderungen oder Vorgänge möglichst rasch – am besten in Echtzeit[156] – zu erkennen, ist der Anspruch der IT-Sicherheit. Die Auswertung von Logdaten bietet hier einen sehr guten Ansatzpunkt. Unverzichtbar wiederum ist sie für den Nachweis oder die Spurensuche nach Sicherheitsvorfällen. Weltweit ist eine Zunahme an Richtlinien und Gesetzen zum Umgang mit Daten zu verzeichnen. Dabei handelt es sich um Vorschriften, die den Schutz der Daten vor Manipulation oder die Nachweisbarkeit von Transaktionen beinhalten, z. B. ISO/IEC 27001 oder PCI DSS Requirement 10[157]. Für die Nachweisfähigkeit bei Audits spielen Logdaten eine zentrale Rolle.

Durch die Korrelation von Logdaten aus verschiedenen Quellen werden Zusammenhänge sichtbar gemacht, die ansonsten verborgen sind. Ein System zur Auswertung von Server-Logdaten bietet – im Gegensatz zu reinen Monitoring-Werkzeugen wie Nagios oder Incinga – Funktionen, um Auffälligkeiten im zeitlichen Verlauf von Kenngrößen zu untersuchen, einzugrenzen und im Detail zu analysieren (Drill-Down). Treten in Applikationen z. B. überlange Reaktions- oder Antwortzeiten auf, kann in einer Querschnittsbetrachtung festgestellt werden, wo zur gleichen Zeit Lastspitzen auftreten, die Hinweise auf bestehende Schwachstellen liefern.

Die Auswertung von Kennlinienverläufen auf der Basis von älteren und aktuellen Logdaten ermöglicht Projektionen dieser Kenngrößen und auf dieser Basis auch eine fundierte Kapazitätsplanung. Die grafische Auswertung macht die Geschwindigkeit des Ressourcenverbrauchs ebenso sichtbar wie den Zeitpunkt, bei dem die vorhandene Kapazität erschöpft sein wird. Schwachstellen sind identifizierbar und Optimierungsmaßnahmen können eingeleitet und letztlich in ihrer Wirksamkeit überprüft werden.

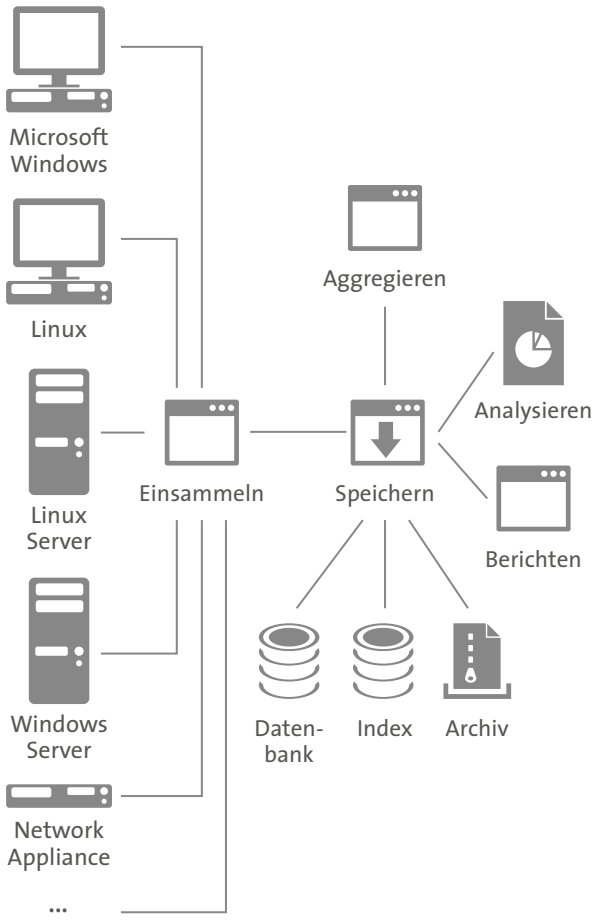


Abbildung 45: Allgemeine Architektur eines Systems für Server-Logdaten-Management

[156] 2011 erfolgte zum Beispiel ein Angriff auf das Playstation-Netzwerk von Sony, der drei Tage andauerte. Am vierten Tag sperrte Sony das Netzwerk, das fast einen ganzen Monat außer Betrieb blieb. Namen, Geburtstage und Kreditkartennummern von nahezu 25 Millionen Kontoinhabern wurden gestohlen. Sony gab die Sicherheitsverletzung sechs Tage nach Sperrung des Netzwerkes bekannt. Angaben von Sony zufolge wurde die Öffentlichkeit nicht früher verständigt, da Sony Zeit brauchte »um den Umfang der Sicherheitsverletzung, der nach mehrtägiger forensischer Analyse durch externe Experten deutlich wurde, zu verstehen.«

[157] In den USA sind zu beachten: Payment Card Industry Data Security Standards (PCI DSS), Sarbanes-Oxley (SOX), Health Insurance Portability and Accountability Act (HIPAA), Federal Information Security Management Act (FISMA), Gramm-Leach-Bliley Act (GLBA).

Einsammeln der Logdaten

Die erste Hürde beim zentralisierten Server-Logdaten Management stellt das Einsammeln (vgl. Abbildung 45) der Logdaten dar. Je nach Quellsystem stehen unterschiedliche Programme bereit, die Logdaten über eine Netzwerkschnittstelle in standardisierter Form bereitstellen können wie z. B. die unterschiedlichen syslog-Varianten. Allerdings verfügt nur ein Bruchteil der in der Anwendung befindlichen Applikationen über die Möglichkeit den syslog-Dienst für das Logging zu benutzen. Viel häufiger werden separate Logdateien benutzt, die sich in Aufbau und Inhalt stark unterscheiden. Unter Windows gilt dies gleichermassen für die Nutzung der Windows Management Instrumentation (WMI).

Speicherung der Logdaten

Bei der zentralen Speicherung sind zwei wesentliche Arten zu unterscheiden, die je nach Anwendungsfall heranzuziehen sind. Für die Archivierung von Logdaten zu Beweiszwecken werden die Daten in ihrer Rohform in Dateien – zum Beispiel im HDFS – gespeichert. Eine Zusammenfassung verschiedener Quellen in gemeinsame Dateien ist dabei möglich, solange die Ursprungsquelle erkennbar bleibt. Andere Logdaten werden dagegen in Speichermedien abgelegt, die eine schnelle Durchsuchung gewährleisten. Dafür kommen dann Indizes oder NoSQL-Datenbanken infrage.

Aggregation der Logdaten, Analyse, Reports

Je nach Anwendungsfall ist es von Vorteil, Logdaten auf die eine oder andere Weise zu aggregieren – z. B. um Auswertungen von Transaktionen zu vereinfachen. Für diesen Zweck haben sich MapReduce-Werkzeuge wie Hadoop etabliert, die große Datenmengen effizient verarbeiten können. Die Implementierung der Verarbeitungsschritte kann der Anwender mittlerweile auch Zusatzprogrammen wie Hive oder Pig überlassen. Diese erzeugen aus einer Anfrage, die in einer Hochsprache formuliert ist, die passende Verarbeitungssequenz. Die Ergebnisdaten können wiederum zur weiteren Auswertung abgespeichert werden.

Bei der Analyse von Logdaten wird die Stärke eines zentralen Managements der Server-Logdaten besonders deutlich: die Visualisierung von Zusammenhängen. Daher ist die geeignete Darstellung der Analysen und mögliche Weitergabe durch Berichte ein wichtiger Bestandteil einer Architektur für das Server-Logdaten-Management.

5.5 Lösungsszenarien mit Sensordaten

Beispiele für Sensordaten

Von Kühlschranken und Kaffeemaschinen bis hin zu den Strom-, Wärme- und Wasserzählern in den Häusern – Sensordaten sind allgegenwärtig. Eine wahre Flut von Sensordaten erzeugen:

- Maschinen, die Fließbänder antreiben,
- Mobilfunkmasten, die Telefonanrufe senden,
- mobile Geräte wie Smartphones und Tablets und
- Gadgets wie Google Glass oder Samsung's Gear Watch.

Vielfach ist es unmöglich, dass Menschen diese Daten sammeln. Man stelle sich Messungen aus dem Inneren einer Erdöl-Pipeline in der gefrorenen Tundra vor. Kein Mensch wäre dafür zu gewinnen, aber Sensoren können die Messungen ohne Pause, zuverlässig und kostengünstig vornehmen.

Sensoren erfassen außerdem Daten aus Natur, Wissenschaft oder Medizin. Beispiele sind:

- Daten über meteorologische Prozesse,
- Daten über Bohrmechanismen für Ölquellen,
- Bodendaten für landwirtschaftliche Zwecke,
- Vitalparameter von Patienten.

Intelligentes Transport-Management

Ein bedeutendes Einsatzgebiet für Big Data mit Sensordaten stellt das Intelligente Transport-Management dar. Hier geht es um den Einsatz von Streaming-Technologie mit dem Ziel, Echzeit-Informationen über Verkehrszustände

zu gewinnen und das Verkehrsmanagement in verkehrsreichen Städten zu optimieren. Als Datenquellen stehen Videos von Kameras an neuralgischen Verkehrspunkten, Sensoren in Fahrzeugen sowie im Stadtbereich installierte Sensoren zur Verfügung. Die Auswertung dieser unterschiedlichen Quellen erzeugt ein Echtzeit-Lagebild der Verkehrssituation. Mit Verkehrsanalyse-Modellen lässt sich die Verkehrsentwicklung prognostizieren. Insgesamt wird so eine dynamische Verkehrsleitung ermöglicht.

Gesundheitswesen

Im Gesundheitswesen wird die Streaming-Technologie eingesetzt, um die Datenströme der medizinischen Geräte in Echtzeit auszuwerten und die Patientenüberwachung in Krankenhäusern zu verbessern.

Von besonderer Bedeutung sind solche Einsatzszenarien bei der Frühgeborenen-Überwachung. Das Children's Hospital Los Angeles sammelt seit 2007 in seinen pädiatrischen Intensivstationen Sensordaten, die alle 30 Sekunden von jedem Patienten erhoben werden. Dieser Daten-Pool enthält mehr als eine Milliarde einzelne Messungen. Die Ärzte haben vor, diese Daten zur genaueren Diagnose und Prognose von medizinischen Vorfällen einzusetzen. Die Herausforderung besteht dabei darin, medizinisch relevante Muster zu finden.

Exploration und Förderung von Erdöl und Erdgas

Bei der Förderung von Öl und Gas wird Streaming eingesetzt, um den Bohr- und Förderungprozess in Echtzeit zu überwachen und auf Anomalien schnell reagieren zu können. Bei Abweichungen im Bereich der Vibrationen und des Drehkräfteverhaltens müssen zeitnah präventive Maßnahmen ergriffen werden. In diesen Use Cases werden sowohl strukturierte als auch unstrukturierte Daten analysiert. Dazu gehören Echtzeit-Wetterdaten, Satellitenbilder, Gravitäts- und Magnetfelddaten, Audio-Daten aus den Geräuschmessungen an den Bohrplattformen, seismische Daten sowie Meeresinformationen wie Wellenhöhe oder Windgeschwindigkeit.

Proaktive Wartung

Die Fähigkeit, einen Geräteausfall vorherzusagen und proaktiv zu reagieren, ist ungemein wertvoll. Schließlich ist es wesentlich günstiger, proaktive Wartungsmaßnahmen durchzuführen als für Notreparaturen zu bezahlen oder Geräte auszutauschen. Wenn der Kühlraum eines Restaurants ausfällt, verliert das Restaurant Tausende von Euro durch verdorbene Lebensmittel und die Einnahmen eines Tages. Ähnliche Problemstellungen gibt es in vielen Industriezweigen – im Betrieb von Produktionsanlagen, Mobilfunkmasten oder medizinischen Großgeräten. Die Herausforderung besteht darin, in den übertragenen diagnostischen Sensordaten mit Hilfe von Algorithmen subtile Muster zu identifizieren. Die ausgewerteten Daten zeigen an, ob eine günstige Reparatur aller Wahrscheinlichkeit nach einen teuren Austausch verhindern kann.

Industrie 4.0

Machine-to-Machine (M2M) steht für den automatisierten Informationsaustausch zwischen Sensoren/Aktuatoren in Maschinen, Automaten, Fahrzeugen oder Containern untereinander oder mit einer zentralen Leitstelle. M2M-Anwendungen bilden auch das Rückgrat für das Internet der Dinge[158].

Im Internet der Dinge werden Objekte intelligent und können über das Internet untereinander Informationen austauschen.

Ziel des Internets der Dinge ist es, die virtuelle mit der realen Welt zu vereinen. Die Event- und Datenvolumina, die in der M2M- und IoT-Umgebung entstehen, sind immens und steigen durch das Wachstum und die weitere Verbreitung von Sensoren dramatisch an.

In die Lösungswelt von M2M integrieren sich auch mehr und mehr Video- und Audio-Datenquellen, die im Bereich Sicherheit oder in diagnostischen Verfahren Einzug halten. Die zeitnahe Auswertung dieser hochvolumigen und zum Teil unstrukturierten Daten wird mehr und mehr

[158] IoT – Internet of Things

in die Wertschöpfungskette der Unternehmen integriert. Deshalb sind Analyse-Verfahren für Data in Motion zwingend erforderlich. Hier setzen sich Streaming- und Real-time Analytics-Verfahren durch.

Einsatz von Streaming-Technologie

Beim Streaming werden lang laufende, kontinuierliche Abfragen bzw. Analysen auf Datenströmen aus der Produktion anstelle von einmaligen Abfragen und Analysen auf gespeicherten Datensätzen durchgeführt. Viele Sensordaten stehen im Kontext von aktuellen Situationen, haben eine Unschärfe aufgrund Ihres zeitlichen und technischen Entstehens und müssen zeitnah in Korrelation mit anderen Informationen gebracht werden, um die Wertschöpfung aus Unternehmenssicht abzubilden.

Die durch die Sensorik in M2M-Applikationen erzeugten Events müssen nicht alle persistiert werden. Vielmehr ist im Bereich der Big-Data-Plattformen die Filterung von relevanten Events und korrelierten Informationen zur Weiterverarbeitung und Speicherung ein wichtiges Element der Beherrschbarkeit großer Datenmengen geworden.

Der Einsatz von Streaming-Technologien kann den TCO einer Big-Data-Lösung signifikant reduzieren, da nur relevante Daten in den Data-Stores weitergeroutet, gespeichert und prozessiert werden müssen.

Des Weiteren geht es darum, aus den Datenströmen in Real-Zeit (low latency) durch Analytische Funktionen und CEP-Prinzipien direkte Auslöser für unternehmensrelevante Events und Prozessverarbeitungen zu entdecken und zu verarbeiten.

Streaming Technologien wie Storm, InfoSphere Streams oder z. B. GemFire erlauben es, aus den hochvolumigen Event-Datenströmen direkt unscharfe Events (Veracity) auszufiltern, relevante Datenkontexte analytisch zu ermitteln und daraus Business-relevante Rückschlüsse z. B. in Echtzeit-Visualisierungs-Dashboards den Business-Bereichen zur Verfügung zu stellen.

Sensoren liefern Big Data – Hadoop ermittelt ihren Wert

Zwei Probleme schränken derzeit die Nutzung von Sensordaten ein:

- ihr Umfang und
- ihre Struktur.

Hadoop ist in der Lage, diese Probleme zu lösen. Sensoren messen und übertragen kleine Datenmengen effizient, allerdings sind sie immer in Betrieb. Mit zunehmender Anzahl an Sensoren und im Verlauf der Zeit können sich die Bytes oder Kilobytes von jedem Sensor schnell zu Petabytes anhäufen. Mit traditionellen Datenspeicher-Plattformen stellt ein solcher Datenstrom ein Problem dar. Die Kosten zur Speicherung der Daten können ein Unternehmen veranlassen, entweder die Sammlung von Daten einzuschränken[159] oder deren Speicherung zu begrenzen[160].

Hadoop bietet eine effizientere und kostengünstigere Speicherung dieser Daten. Dank Hadoop verwandeln sich große Mengen an Sensordaten von einem Problem zu einem Vermögenswert.

Sensordaten sind zudem zum Zeitpunkt ihrer Erhebung in der Regel unstrukturiert und werden durch einen mechanischen, sich wiederholenden Prozess generiert.

Apache Hive kann die Sensordaten in Übereinstimmung mit ihren Metadaten[161] umwandeln. Die Daten werden dann in HCatalog in einem geläufigeren Tabellenformat präsentiert, auch wenn die zugrunde liegenden Daten noch in ihrem ursprünglichen Format in HDFS vorhanden sind.

[159] durch Einschränkung der Anzahl der Sensoren
[160] durch Löschen von Daten über einer bestimmten Menge oder nach einem bestimmten Zeitraum
[161] z. B. Zeit, Datum, Temperatur, Druck oder Neigung

5.6 Lösungsszenarien mit Standortdaten

Standortdaten sind eine Untergruppe der Sensordaten, da das Gerät seinen Standort erkennt und Daten zu seinem Längen- und Breitengrad in vorgegebenen Intervallen übermittelt. Kommerziell interessant wird diese spezielle Form von Sensordaten mit der massenhaften Verbreitung von GPS-fähigen Geräten z. B. in Mobiltelefonen oder Kameras, oder aber auch in professionellem Equipment in Fahrzeugen[162]. Logistikunternehmen bieten an, GPS-Tracker einer Sendung hinzuzufügen, um so als Endkunde den Versand einer Ware verfolgen zu können.

Verbraucherorientierte Unternehmen wollen Standortdaten nutzen, um zu erfahren, wo sich potentielle Kunden zu bestimmten Tageszeiten aufhalten – sofern der Kunde der Verwendung der ortsbezogenen Informationen zugestimmt hat. Personalisierte Werbung mit Ortsbezug wird häufig als Anwendungsfall genannt. Einem Nutzer einer App auf dem Smartphone kann dann ein Rabatt-Gutschein übermittelt werden, sobald dieser sich vor einer Filiale befindet.

Startups wie Waze nutzen Ortsinformationen, die über die App auf Smartphones erfasst werden, zur Verkehrsflussermittlung und zur Aktualisierung des Kartenmaterials, welches von dem Routenplaner genutzt wird.

Eine weiteres viel versprechendes Einsatzgebiet von Standortdaten bildet das Flotten-Management. Logistikunternehmen, die Sendungen an Privathaushalte zustellen, können feingranulare Standortdaten, die in regelmäßigen Zeitabständen gesammelt werden, nutzen, um Fahrerrouten zu optimieren. So lassen sich Lieferzeiten verkürzen und Kraftstoffkosten sowie Unfallrisiken vermindern. Logistikunternehmen möchten zu jeder Tageszeit wissen, wo sich ihre mobilen Vermögenswerte befinden. Allgemein lassen sich über diese Datenquelle Themen adressieren wie

- Verringerung des Leerlaufverbrauchs
- Einhaltung von Vorschriften, die die Mindestruhezeit und die maximale Lenkzeit vorgeben,
- Vorbeugung von Unfällen durch die Erkennung von unsicherem Fahrverhalten.

Das sogenannte Geofencing wird hierbei genutzt, um Bereiche zu definieren, an denen sich GPS-Sensoren, welche z. B. an einer Wechselbrücke angebracht sind, zu einer bestimmten Zeit befinden dürfen. Verlässt ein solcher Sensor den definierten Bereich, wird ein Alarm ausgelöst – dies kann auch in Kombination mit anderen Sensoren hinsichtlich Erschütterungen, Temperatur etc. erfolgen.

Versicherungen bieten heute – häufig bei Flottenfahrzeugen – an, das Fahrverhalten des Fahrzeuges zu überwachen. Somit können Ereignisse wie heftiges Bremsen, extremes Beschleunigen etc. mit einem Ortsbezug gesehen werden und entsprechenden Häufungspunkte und –zeiten ermittelt werden. Dies kann dann wiederum zur maßgeschneiderten Erstellung von Tarifen führen. Mit Car2Car-Kommunikation können lokationsbezogene Informationen dazu dienen, Warnungen an nachfolgende Fahrzeuge zu senden.

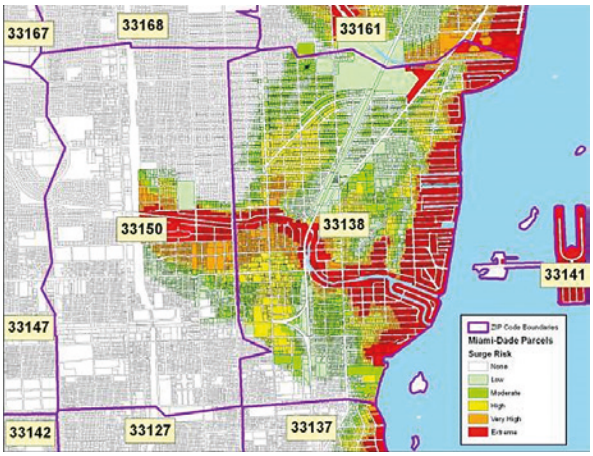


Abbildung 46: Simulationen von Überflutungsflächen mit Geodaten

[162] OBUs – On Board Units

Die Verknüpfung von bereits existierenden Daten mit einem Ort[163] ermöglicht generell die Formulierung von neuen Fragestellungen, die in sich eine größere Unschärfe haben können. Dies können Fragestellungen hinsichtlich des Umkreises um einen Ort[164] oder aber auch Prognosen basierend auf Wetterinformationen sein, wo sich eine Rauchwolke in einigen Stunden höchstwahrscheinlich hin ausbreiten wird. Spezialisierte Plattformen bieten hier leistungsstarke Algorithmen an, um z. B. eine Ausbreitungsprognose mit Katasterdaten zu kombinieren. So können dann rechtzeitig Warnungen an Bewohner der betroffenen Gebiete herausgegeben werden (vgl. Abbildung 46[165]).

5.7 Lösungsszenarien mit Freitext-Daten

Je nach Einsatzgebiet und betrachteter Datenquellen gibt es vielfältige Einsatzbereiche für Big-Data-Technologien auf natürlichsprachigen Texten, von denen im Abschnitt 5.7 einige erläutert werden.

Marketing und Marken-Pflege

Mittels Werkzeugen des Social Media Monitoring beobachten Marketing-Abteilungen Web-Seiten, Blogs, Foren und andere Social Network-Kanäle, um Informationen darüber zu erhalten, welche Meinungen und Ideen Kunden bzw. Anwender zu den Produkten und Dienstleistungen des Unternehmens haben. Werden an dieser Stelle gezielt linguistische und semantische Technologien genutzt, so ist eine qualitativ deutlich bessere Erschließung und Durchdringung der Daten möglich. Im Sinne des »BI auf Text« werden hier zunächst Entitäten und Strukturen in den Texten erkannt, welche dann wiederum in vielfältiger Art und Weise gruppiert und analysiert werden können, wie Abbildung 47 beispielhaft zeigt.

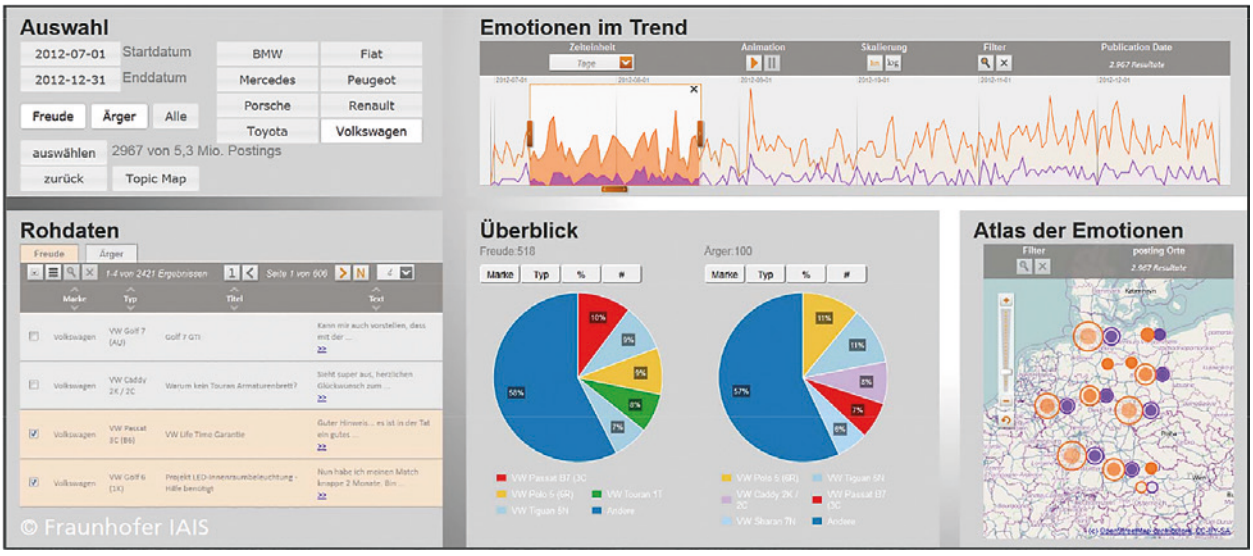


Abbildung 47: Visuelle Datenexploration im Demonstrator »Living Lab Big Data« auf der CeBIT 2013[166]

[163] z. B. über die Auflösung von Straßennamen zu einer Geolokation
[164] z. B. einen Supermarkt
[165] Simulationen von Überflutungsflächen sind mit Geodaten deutlich genauer durchzuführen als z. B. auf Ebene der Postleitzahlen.
[166] Quelle: www.iais.fraunhofer.de/bigdata-studie.de

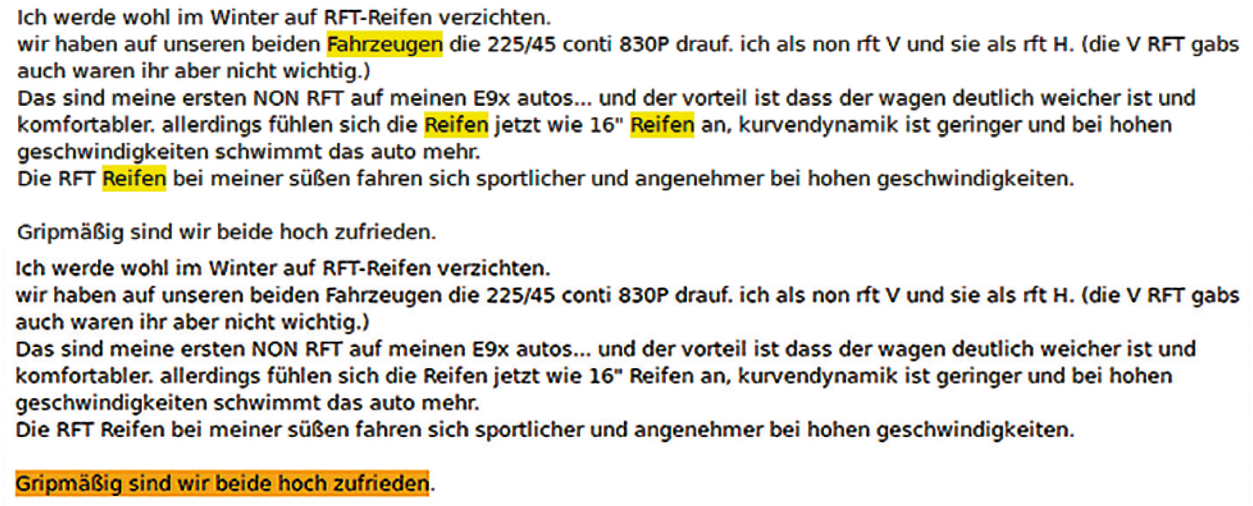


Abbildung 48: Technische Terme und Stimmungsdaten in einem Forum-Beitrag aus dem motor-talk Portal

Im Umfeld der Marken-Pflege kann ein Zusammenhang zwischen Marken, Typen, Ausstattungmerkmalen und Stimmungsdaten hergestellt werden. Texte werden angereichert um Annotationen, welche die gezielte Auswahl und Analyse von Texten unterstützen, und des weiteren Lese- und Interpretationshilfe sind. In folgendem Beispiel sind technische Terme und Stimmungsdaten als erkannte Entitäten farblich hervorgehoben:

Anreicherung von Nachrichten

Unter anderem für die Pressearbeit ist es enorm wichtig, dass schnell große Mengen an Informationen erschlossen werden können. Hierfür ist es notwendig, dass die Rohtexte automatisiert mit Hintergrundinformationen angereichert werden. So kann eine Vielzahl an Quellen beobachtet werden, etwa die Webseiten von Print-Medien, Rundfunk- und Fernseh-Anbietern oder auch reine Nachrichtenportale. Die Meldungen werden mittels der in Unterabschnitt 4.3.3 beschriebenen Verfahren analysiert und somit Informationen zu Personen, Orten, Organisationen, Ereignissen etc. extrahiert. Ebenso werden zu den erkannten Entitäten weiterführende Informationen gesucht, etwa zugehörige Artikel bei Wikipedia. Mit all diesen Fakten werden die Nachrichten angereichert, so dass dem Betrachter anhand der visuellen Darstellung eine schnelle Erschließung der Inhalte ermöglicht wird.

Analyse von Service-Reports

Im Umfeld des Kundendienstes fallen meist viele Berichte an, die zwar in Teilen wohlstrukturiert sind (man denke an Geräte-Nummern oder Messwerte), jedoch neben diesen weitere wertvolle Informationen in Freitext-Form enthalten, wenn etwa der Monteur im Rahmen der Dokumentation der Service- und Reparatur-Arbeiten einen mehr oder weniger präzisen Report verfasst, in dem er die Symptome, den Fehler und die Schritte zur Fehlerbehebung beschreibt. Zieht man außerdem Fehlerbeschreibungen der Kunden selbst hinzu, etwa wenn diese über ein Web-Portal ihr Problem beschreiben, so fallen einerseits zwar vielfältige Daten und Informationen an, die aber andererseits ohne Big-Data-Techniken und semantische Analysen nicht wirklich erschlossen werden können.

Linguistik und Semantik helfen an dieser Stelle, die Texte aufzubereiten, Produktbezeichnungen zu erkennen, Problembeschreibungen zu vereinheitlichen und all diese Information sinnvoll zu gruppieren und zu ordnen, so dass zum Beispiel Erkenntnisse und häufige Fehler und Schwachstellen abgeleitet werden können.

Wettbewerbs- und Innovations-Management

In Zeiten immer kürzerer Innovations-Zyklen ist ein professionelles Innovations-Management für Unternehmen überlebenswichtig. Hierzu zählen interne Prozesse und eine Kultur, um neue Ideen zu fördern, aber ebenso die aktive Beobachtung des jeweiligen Marktes in Bezug auf Fragen wie: Welche neuen Trends gibt es? Was machen meine Wettbewerber? Gibt es neue Player im Markt? Was wünschen sich meine Kunden?

Hierzu sind intensive Recherchen vor allem auf öffentlichen Quellen notwendig, wobei im besten Falle Inhalte aus dem sogenannten Deep Web, also etwa Patent-Datenbanken, einbezogen werden sollten. Wichtig hierbei ist einerseits ein enger Fokus auf das jeweilige Anwendungsgebiet; andererseits müssen Signale möglichst früh erkannt werden, bevor sie allgemein bekannt werden.

An dieser Stelle sind Big-Data-Technologien notwendig, um insbesondere Quellen im Internet zu überwachen und die immens großen Datenmengen zu bewältigen. Darüber hinaus kommen semantische Verfahren zum Einsatz, welche auf einem Wissensmodell über die relevanten Produkte und Technologien, bekannte Wettbewerber, Partner und Kunden und ähnlichen Informationen bestehen. Basierend darauf erschließt Semantik Zusammenhänge, kann Trends aufzeigen oder auf neue Marktteilnehmer hinweisen, indem zum Beispiel erkannt wird, dass ein bisher unbekanntes Unternehmen in relevanten Märkten operiert bzw. Meldungen über interessante Technologien veröffentlicht.

Patent-Recherche

Sobald Unternehmen im Rahmen des Innovationsprozesses ihre eigenen Ideen patentrechtlich schützen wollen, kommen sie mit Prüfern im Patentamt in Kontakt, welche eine Reihe von Fragen zu klären haben: Werden die erforderlichen formellen Richtlinien eingehalten? Gibt es bereits identische oder zumindest hinreichend ähnliche Erfindungen, die einer Patentierung entgegenstehen? Passt der Inhalt des Patents zum angestrebten Schutz?

Um all dies schnell, aber dennoch rechtssicher und nachvollziehbar beantworten zu können, benötigt der Patentprüfer Zugriff auf vielfältige Informationen, vor allem Patentdatenbanken und Fachliteratur. Nicht alle Informationen sind lokal verfügbar, es gibt eine Vielzahl an Datenquellen und Formaten etc.

Big-Data-Techniken und semantische Verfahren helfen an dieser Stelle dabei, Inhalte zu erschließen sowie Bezüge zu relevanten Funktionen, Materialien, Technologien etc. deutlich zu machen, so dass zum Beispiel darauf hingewiesen werden kann, dass in anderen relevanten Dokumenten sehr ähnliche Beschreibungen entdeckt wurden, wobei Ähnlichkeiten hier anhand eines fachlichen Wissensmodells ermittelt werden. Hierdurch erhält der Prüfer gezielt Hinweise auf potenziell relevante Dokumente und kann diese gezielt sichten, um somit eine fundierte Entscheidung treffen zu können.

Risiko-Management und Compliance

Versicherungsgesellschaften besitzen riesige Mengen an unstrukturierten, textbasierten Schadensfalldaten. Sie haben außerdem Zugriff auf andere, strukturierte und unstrukturierte Datensätze (öffentlich und privat), die sie mit Schadensfalldaten kombinieren können, um ihre Risikobeurteilung zu verbessern oder Missbrauchsfälle aufzudecken.

5.8 Lösungsszenarien mit Video- und Sprachdaten

Auswertung von Medien-Archiven

In Medienarchiven werden mittlerweile große Mengen an Inhalten gespeichert, deren Erschließung ohne Methoden der Sprachverarbeitung kaum mehr möglich ist. Dies können zum Beispiel Archive der Medienanstalten, Inhalte aus dem Bereich eLearning oder auch andere Audio- und Video-Inhalte sein. Als ein Beispiel-Szenario wurde eine News-Discovery-Applikation entwickelt, welche umfangreiche Video-Archive, zum Beispiel aus Debatten und Ausschusssitzungen des Deutschen Bundestages, analysiert und auswertet. Nachdem mittels Transkription die Audio- und Video-Daten in Text konvertiert wurden, werden wiederum die im Unterabschnitt 4.3.3 beschriebenen Schritte der Sprachverarbeitung angewandt, wobei hier zusätzliches Wissen in Form einer Ontologie über Mitglieder des Bundestages, Parteienzugehörigkeit etc. genutzt wird. Im Ergebnis ist es nicht nur möglich, gezielt einzelne Video-Beiträge zu suchen, sondern auch Auswertungen zu Schwerpunkten und Positionen der Parteien, der Aktivität der einzelnen Abgeordneten, Meinungsführern etc. vorzunehmen, wie in der Abbildung 49 anhand der Diskussionen um »Stuttgart 21« dargestellt:

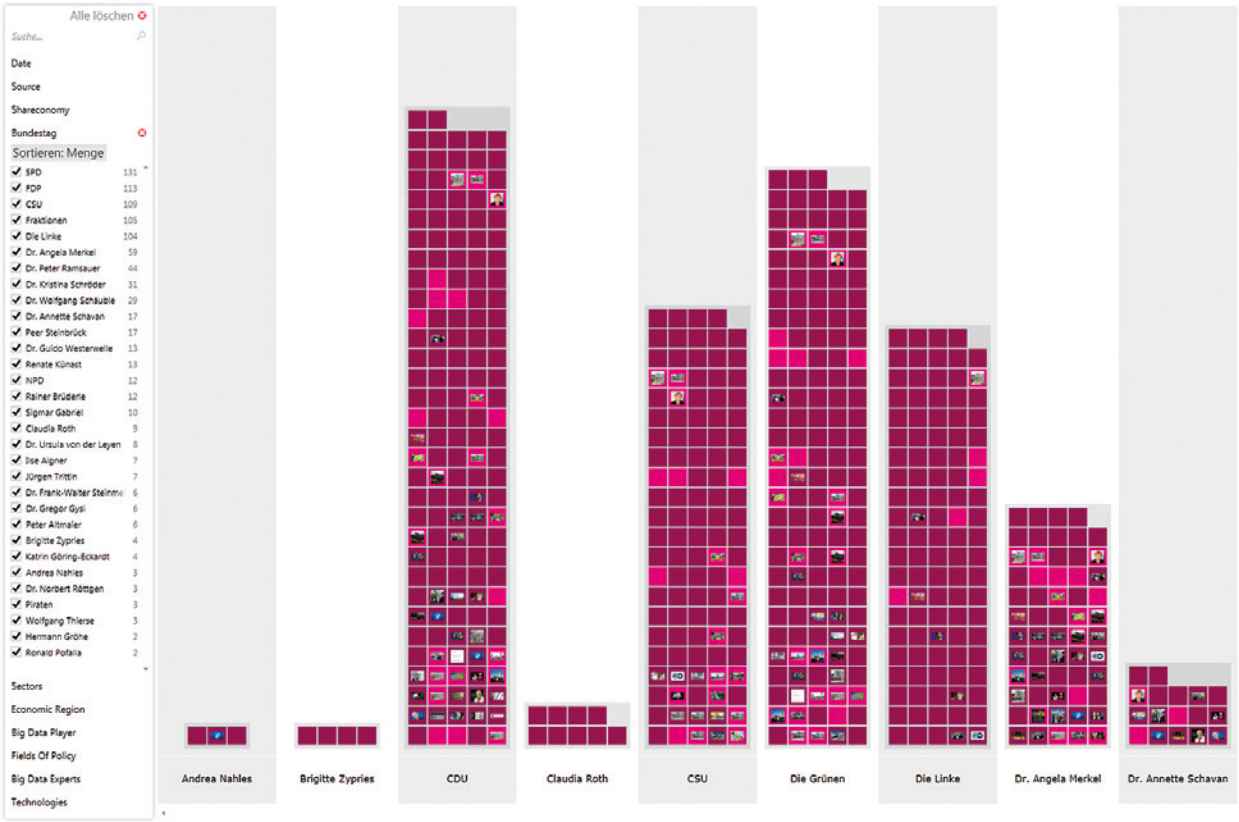


Abbildung 49: Inhaltliche Erschließung von Video-Archiven

5.9 Big Data und Business Intelligence

Evolution von Business Intelligence zu Big Data

Seit vielen Jahren ist Business Intelligence (BI) in Unternehmen etabliert; BI-Werkzeuge und Architekturen befinden sich in den meisten Unternehmen im Einsatz. Häufig sind umfangreiche Investitionen in BI-Systeme geflossen, die durch das Aufkommen von Big Data nach Möglichkeit nicht entwertet werden sollten. Im Abschnitt 5.9 wird eine Architektur vorgestellt, die die Evolution von BI in Richtung Big Data vorsieht.

Anforderungen an eine kombinierte Business-Intelligence-/Big-Data-Architektur

In vielen Unternehmen sind BI-Systeme im produktiven Einsatz, die auf einem Reporting-Werkzeug und relationalen Datenbanken aufsetzen. In diesen Systemen sind zahlreiche Geschäftsregeln definiert und umfangreiche Datenbestände zusammengeführt, die als Data Warehouse bezeichnet werden. Ein Data Warehouse kann auch einen wichtigen Bestandteil einer Big-Data-Architektur bilden, denn Big-Data-Analysen benötigen oft den Zugriff auf eine integrierte Datenbasis des Unternehmens.[167]

Seit einigen Jahren rücken nun neue Datenquellen in den Blick von Analysten[168], die sich mit den bestehenden BI-Architekturen nicht oder zumindest nicht effizient verarbeiten lassen. Der Gedanke liget nahe, BI-Architekturen nach Möglichkeit so weiter zu entwickeln, dass auch neue Datenquellen erschlossen werden können.

Abbildung 50: Anforderungen an eine kombinierte Business-Intelligence-/Big-Data-Architektur

Die an eine kombinierte Business-Intelligence-/Big-Data-Architektur gestellten Anforderungen sind in der Abbildung 50 zusammengefasst. Am Markt ist gegenwärtig keine singuläre Technologie sichtbar, die diesen Nutzeranforderungen entsprechen könnte. Diese Situation lässt sich am besten an einem Beispiel aus der Praxis verdeutlichen, bei dem die Aktivitäten von Kunden auf der Webpräsenz eines Unternehmens mit Hilfe von Weblog-Daten analysiert werden sollen[169]. Diese Informationen werden entweder direkt von dem Webserver im laufenden Betrieb protokolliert oder es wird spezieller Code in die Webseite eingefügt, der Interaktionen an einen Erfassungsserver sendet. Bei Webseiten mit einer hohen Besucherzahl[170] entstehen dabei Daten im Bereich von GB bis zu TB pro Tag. Häufig sind diese Daten in einem technischen Format abgespeichert, das sich für die Analyse in einer auf eine relationale Daten-Haltung setzenden BI-Architektur nicht eignet. Es ist auch nicht sinnvoll, den Rohdatenstrom in ein relationales Modell zu zwängen, wenn noch keine Klarheit über die zu stellenden Fragen herrscht. Von Interesse könnten z. B. die Pfade sein, über die Nutzer häufig auf der Webseite navigieren – sind sie bekannt, so kann in die Navigation mit personalisierten Angeboten direkt eingegriffen werden.

Als ein weiteres Hindernis für relationale BI-Architekturen kommt hinzu, dass sich solche Anfragen nicht auf einfache Weise in der quasi-Standardabfragesprache SQL formulieren lassen.

[167] Das ist z. B. bei der Überprüfung der Aussagekraft von Vorhersagemodellen wichtig.
[168] Vgl.: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Leitfaden des BITKOM, Berlin 2012
[169] Vgl. auch Abschnitt 5.4
[170] z. B. Webshops

Die alleinige Speicherung von Rohdaten über einen Zeitraum von mehreren Jahren in einem Data Warehouse ist ökonomisch nicht sinnvoll. Das gilt insbesondere dann, wenn für BI-Anwendungen aufgrund der Performance-Anforderungen eine In-Memory-Datenbank[171] zum Einsatz kommt. Sinkenden Preisen für Hauptspeicher stehen wachsende (Roh-)Datenmengen gegenüber, was kommerziell die Archivierung von Rohdaten in einer In-Memory-Datenbank ausschließt.

Komponenten einer hybriden, integrierten Architekturen

Die sinnvolle Antwort auf die in Abbildung 50 dargestellten Anforderungen sind hybride Architekturen, die eine kosteneffiziente Integration von BI und Big Data in einer integrierten Architektur verbinden. Bei der TCO-Betrachtung spielen neben den Kosten für Anschaffung und Betrieb einer solchen Architektur auch der Schulungsaufwand, die Einsatzbreite sowie die Zeit für Datenexperimente eine wichtige Rolle.
Am Beispiel der Weblog-Daten[172] werden im Folgenden die Einsatzbereiche möglicher Komponenten einer solchen hybriden Architektur vorgestellt.

Komponente 1: Erfassung, Speicherung und Vorverarbeitung

Es ist oft erforderlich, Weblog- oder andere Rohdaten unter neuen Gesichtspunkten zu analysieren. Für solche Situationen ist es von Vorteil, Rohdaten möglichst lange speichern zu können. Es wird dann möglich, unstrukturierte Rohdaten erst zur Laufzeit einer Analyse mit einer sinnvollen semantischen Struktur zu belegen. Dieser als Late Binding bezeichnete Ansatz unterscheidet Big Data wesentlich von klassischer Business Intelligence, bei der die Rohdaten zunächst in ein definiertes Schema transformiert werden.
Für die Erfassung, Speicherung und Vorverarbeitung von Rohdaten bieten sich Hadoop oder ähnliche verteilte Dateisysteme an. Durch die eingebaute Replikation von Daten können so auch größte Datenmengen auf Standard-Servern langfristig und ausfallsicher in der Rohform gespeichert werden. Das Hadoop-Ökosystem sieht auch die Möglichkeit vor, Vorverarbeitungsschritte Batch-orientiert zu absolvieren.

Komponente 2: Explorative Analyse und Mustererkennung

Stehen die Rohdaten in einem Hadoop-Cluster zur Verfügung, so können Analysten oder Data Scientists mit diesen Daten experimentieren und sie mit anderen Datenquellen zu kombinieren, um auf diese Weise neue Erkenntnisse aus den Daten zu gewinnen.
Dafür stellt Hadoop zahlreiche Werkzeuge bereit. In der Praxis zeigt sich jedoch, dass nur wenige Nutzer in Unternehmen die dafür benötigten Kenntnisse beseitzen oder erwerben wollen.
Aus dieser Situation gibt es einen Ausweg. Der erste Schritt im Late-Binding-Ansatz zur Verarbeitung von unstrukturierten Rohdaten sieht vor, den Daten eine Struktur aufzuprägen. Danach liegen strukturierte Daten vor, die sich effizient mit relationalen Datenbanken analysieren lassen. Eine Vorverarbeitung kann also in Hadoop angestoßen und in einer relationalen Datenbank weitergeführt werden. Alternativ arbeitet die relationale Datenbank ebenfalls massiv-parallel und bietet in SQL eine MapReduce[173]-Implementierung mit vorgefertigten Analysealgorithmen an. Beiden Ansätzen ist gemeinsam, dass die Komplexität der neuen Hadoop-Technologie vor dem Anwender verborgen wird: Er muss lediglich einen erweiterten Satz von SQL-Funktionen erlernen.
Die geschilderten Vorgehensweisen erweitern den Kreis von Nutzern deutlich, die in Hadoop gespeicherte, große Datenmengen analysieren können. Außerdem können die in einem Unternehmen genutzten Werkzeuge zur Visualisierung weiter verwendet werden, was den Einstieg enorm erleichtert.
Die von Hadoop bereitgestellten Werkzeuge wie Hive[174] werden zügig weiter entwickelt, um ihren Einsatz komfortabler zu gestalten. Viele Anwender werden es

[171] vgl. dazu Abschnitt 6.2
[172] vgl. auch Abschnitt 5.4
[173] vgl. Unterabschnitt 4.1.1
[174] vgl. Unterabschnitt 4.1.1

ungeachtet dieser Fortschritte vorziehen, in der bekannten Welt der massiv-parallelen, relationalen Datenbanksysteme zu verbleiben. Hier sind die Laufzeiten von Anfragen deutlich geringer. Experimente mit Daten sind schneller beendet, und so ist in kürzerer Zeit klar, ob ein gewählter Ansatz zielführend ist. Die Time to Insight ist ein maßgeblicher Faktor in TCO-Betrachtungen.
Neue Entwicklungen senken die Einsatzbarrieren für Hadoop weiter: So bietet die Hadoop-Komponente HCatalog – seit kurzem ein Teil des Hive-Projektes – die Möglichkeit, auf Daten in Hadoop aus anderen Werkzeugen heraus zuzugreifen und dabei auch Filter zu setzen[175].

Der direkte Durchgriff auf in Hadoop gespeicherte Daten ist von großer Bedeutung, da im Falle von Big Data das einfache Kopieren von Daten ohne Berücksichtigung von Filterkriterien sehr viel Zeit beanspruchen kann. Zusätzlich muss das jeweilige Zielsystem ebenfalls über entsprechende Speicherkapazitäten verfügen.
Für die explorative Analyse von Daten ist der Zugriff auf integrierte und qualitätsgesicherte – z. B. in einem Data Warehouse vorgehaltene – Unternehmensdaten wesentlich. In Weblog- und anderen Rohdaten sind häufig Identifikatoren wie Produktnummern enthalten. Für eine prädiktive Analyse ist dies vollkommen ausreichend. Für die Interpretation von Ergebnissen muss man z. B. wissen:

- welches Produkt sich hinter einer Identifikationsnummer verbirgt,
- ob das Produkt Gegenstand einer Werbekampagne war,
- welchem Teil des Webshops das Produkt an welchem Tag zugeordnet war[176],
- ob Lieferprobleme verzeichnet wurden,
- in welchen Zuständigkeitsbereich[177] das Produkt fällt,
- zu welcher Kategorie das Produkt gehört.

Das bedeutet: Der Durchgriff in ein Data Warehouse ist in dieser Phase der Analyse entscheidend, um eine Big-Data-Analyse im Gesamtkontext aller relevanten Unternehmensdaten bewerten zu können.

Komponente 3: Unterstützung von Geschäftsprozessen und BI

Bisher wurden die Komponenten 1 und 2 einer hybriden BI-Big-Data-Architektur betrachtet. Sie dienen primär der Erfassung und Analyse von großen, nicht-relationalen Daten. Die dritte Komponente einer solchen Architektur setzt die Aufgabe um, die Ergebnisse von Big-Data-Analysen in die Geschäftsprozesse einfließen zu lassen. Hierfür ist ein Data Warehouse prädestiniert, führt es doch bereits viele Daten aus Geschäftsprozessen in einer Plattform zusammen. Die aus Big Data gewonnenen Erkenntnisse können durch die Einbettung in das logische Datenmodell in den Gesamtkontext des Unternehmens gestellt werden.
Relationale Datenbanken, die architektonisch für einen analytischen Data Warehouse Workload ausgelegt sind, erlauben in einer gesicherten Art und Weise Tausenden Nutzern den Zugriff auf die Big-Data-Erkenntnisse, denn analytische Modelle etc. sind für ein Unternehmen schützenswerte Daten.

Außerdem lässt sich das Ergebnis einer Big-Data-Analyse kosteneffizient in einer für Data Warehousing optimierten Datenbank ablegen und vielfältig nutzen. Diese hybriden BI- und Big-Data-Systeme verfügen über Methoden zur Feinjustierung der Allokation von Systemressourcen zu Anfragen und bewegen Daten in Abhängigkeit von der Zugriffsfrequenz vollautomatisch zwischen den unterschiedlichen Speichermedien[178] innerhalb des Systems. So wird eine maximale Performance beim Zugriff auf die Daten gewährleistet. Und es entfällt die Notwendigkeit, manuell Duplikate zu erstellen, die dann z. B. in eine dedizierte In-Memory-Datenbank kopiert werden müssten.

Auf diese Weise können BI und Big Data in einer hybriden Architektur unter Ausnutzung von neuen Technologien wie In-Memory-Computing zusammengeführt werden. So wird ein angemessener TCO-Wert erreicht, wie er beim Einsatz von nur einer der beiden Technologien nicht möglich wäre.

[175] Beispiel: Nutzung der Weblog-Daten der letzten drei Monate
[176] z. B. Kleidung oder Kleidung und »Sale«
[177] z. B. Produktmanagement
[178] wie Hauptspeicher, SSDs und HDDs

Empfehlungen zur Umsetzung und Datenkultur

Im Kontext einer hybriden BI-/Big-Data-Architektur (vgl. Abbildung 51) wird häufig die Frage gestellt, welche Plattform sich für eine gegebene Aufgabe eignet.

Die Antwort ist von vielen Faktoren abhängig. Grundsätzlich sollte eine hybride Architektur einen hohen Grad von Vereinheitlichung aufweisen, so dass die Frage, welche Technologie grundsätzlich für einen Anwendungsfall zu verwenden ist, bei der Anbindung einer neuen Datenquelle zunächst keine Rolle spielt.

Darüber hinaus bestimmt auch der aktuelle Stand des Analyseprozesses, welche Technologie zunächst Verwendung findet. Zu Beginn der Datenexploration hängt noch viel mehr von den Erfahrungen der Akteure ab, welche Technologie am schnellsten die möglichen Antworten liefert. Hat sich eine Idee als erfolgsversprechend erwiesen, so sollte dann eine »Operationalisierung« erfolgen, d.h. die Umsetzung nach den Standards des Unternehmens. Auf diese Weise werden zwei Aufgaben gelöst:

- schnelle Erprobung von neuen Datenquellen und neuen Ideen auf ihren Mehrwert hin, ohne dass zu diesem Zeitpunkt übermäßig viel auf Standards zu achten wäre;
- gleichzeitig ist dieser Prototyp eine genaue Spezifikation, die als Nukleus die Umsetzung nach den gültigen Standards sowie die Wartung der Lösung im Produktivbetrieb erleichtert.

Ziel sollte es daher sein, flexibel zwischen den Technologien der hybriden BI-/Big-Data-Architektur wählen zu können. Dafür muss die Konnektivität der Systeme untereinander gegeben sein, um auch große Datenmengen aus

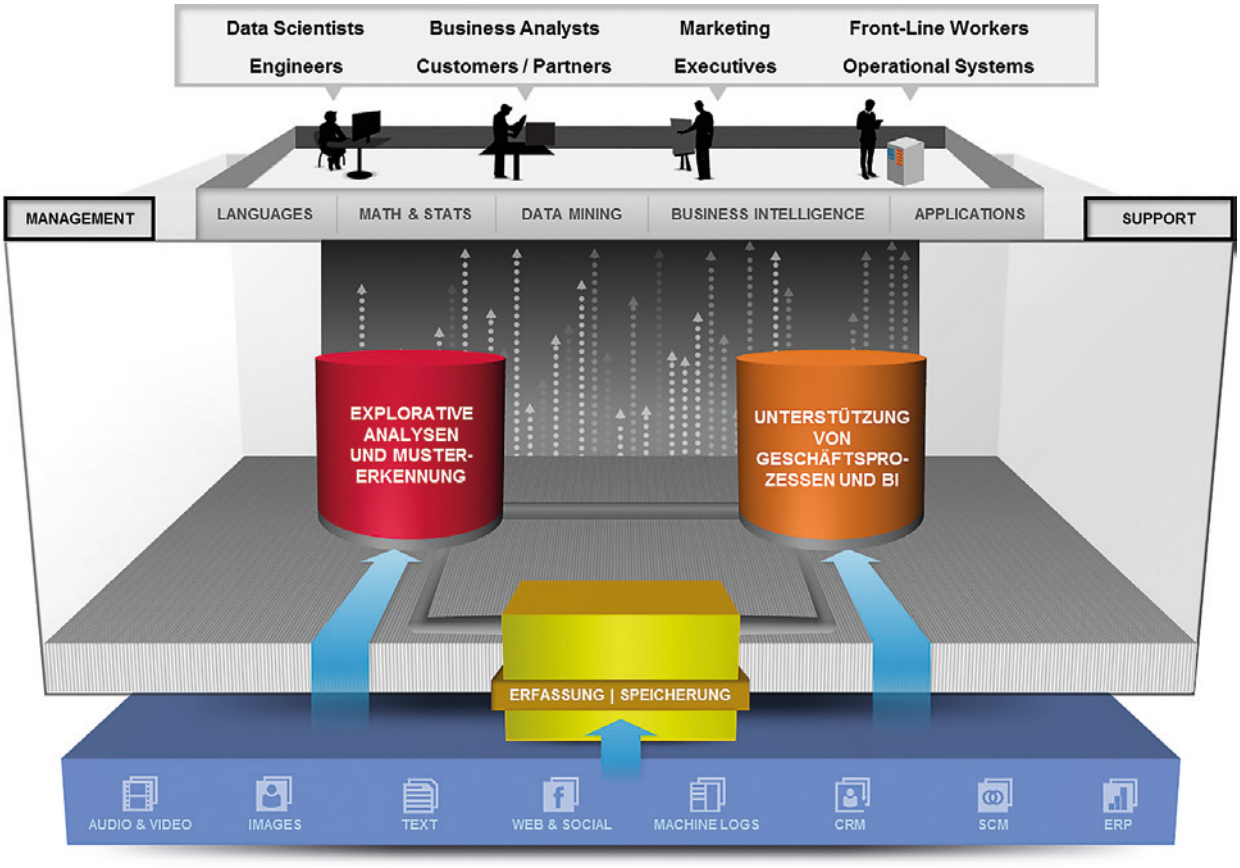


Abbildung 51: Komponenten einer hybriden BI-/Big-Data-Architektur

Quellen unterschiedlichen Typs miteinander verknüpfen und effizient verarbeiten zu können.

Datenkultur

Eine hybride BI-/Big-Data-Architektur stellt eine nicht zu unterschätzende technologische Herausforderung dar. Ihre Bewältigung reicht aber nicht aus. Ein langfristiger Erfolg setzt die Etablierung einer fördernden Datenkultur im Unternehmen voraus, die sich in drei Facetten widerspiegelt:

- Alle Daten sind als wichtig anzusehen. Nicht selten wird die Bedeutung von Daten erst in einem zukünftigen Kontext deutlich.
- Innovation setzt ganzheitliches Denken voraus, das Impulse aus einer Zusammenführung von Daten in einer hybriden Architektur erhalten kann.
- Vertrauensvolle Zusammenarbeit im Team ist ein Eckpfeiler für die Etablierung einer Datenkultur.

Big Data bedeutet also mehr als die drei V[179], mehr als Technologie und auch mehr als ein bestimmter Use Case. Es sind die Mitarbeiter, die eine Datenkultur ausbilden und die Technologien und Werkzeuge einsetzen, um aus Daten wirtschaftlichen Nutzen zu generieren.

Anwendungsbeispiel

Ein Beispiel aus der Wirtschaftspraxis soll die im Abschnitt 5.9 formulierten Zusammenhänge untermalen.
Bereits seit vielen Jahren setzt Ebay auf eine hybride Architektur (vgl. Abbildung 52), um neue datengetriebene Produkte zu entwickeln. Ein bedeutender Erfolgsfaktor hat sich dafür die Verfügbarkeit von Daten für viele Mitarbeiter erwiesen.
Besonders hervorzuheben ist: Die Nutzerkreise sind nicht auf eine Plattform beschränkt. Die Grenzen zwischen klassischer BI und Big Data sind fließend. Eine strikte Trennung zwischen beiden Welten führt letztlich zu doppelter Daten-Haltung und erschwert konsistente Antworten.

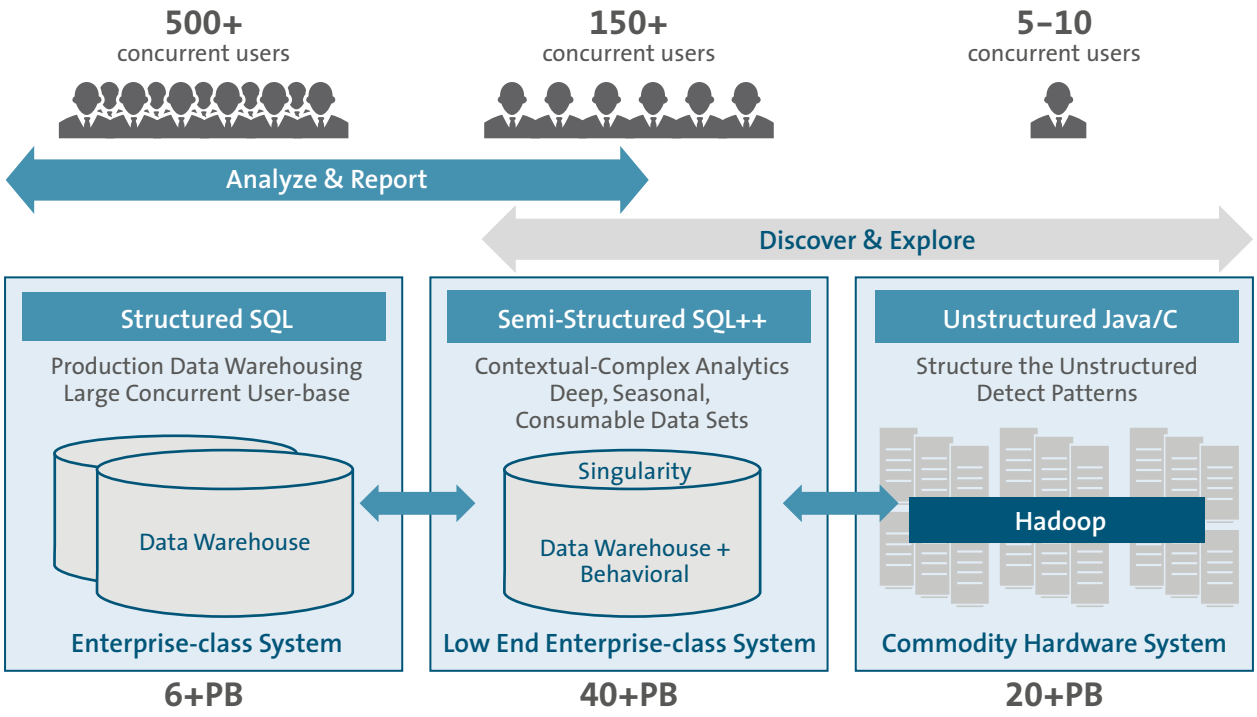


Abbildung 52: Big-Data-Architektur bei Ebay, Stand 2011[180]

[179] Volume, Velocity, Variety

[180] Quelle: http://www-conf.slac.stanford.edu/xldb2011/talks/xldb2011_tue_1055_TomFastner.pdf
Mittlerweile ist die eBay-Hadoop-Installation auf 365 PB und 50.000 Server gewachsen.

5.10 Data-Warehouse-Entlastung – Aktives Archiv in Hadoop

Ein Hadoop-Cluster kann sowohl am Beginn des Daten-Lebenszyklus, quasi als Einflugschneise für Daten, als auch an seinem Ende eingesetzt werden – zur Archivierung der Daten für eine spätere Analyse. Damit schafft Hadoop in existierenden EDW freie Kapazitäten, in die Unternehmen hineinwachsen können, ohne zunächst in die Erweiterung des EDW investieren zu müssen.

Entlastung des Enterprise Data Warehouse durch Hadoop

Hadoop erlaubt die wirtschaftliche Speicherung von Daten beliebiger Struktur auf unbegrenzte Zeit, insoweit das rechtlich zulässig ist. Hadoop kann man sich vereinfacht als ein neuartiges Data Warehouse vorstellen, das größere Datenmengen und mehr Arten von Daten speichern kann und außerdem flexiblere Analysen zulässt als etablierte EDW. Als Open-Source-Software für Standard-Hardware steht Hadoop aus Sicht der Wirtschaftlichkeit etwa um den Faktor zwanzig besser da als konventionelle Data-Warehouse-Lösungen. Im Unterschied zu konventionellen EDW Architekturen mit proprietärer, hoch optimierter Hardware wurde Hadoop so konzipiert, dass ein Betrieb auf handelsüblichen Servern mit preiswertem Speicher die Norm ist.

Und so verwundert es nicht, dass großen Unternehmen bereits erhebliche Kostenreduktionen gelungen sind, indem sie Hadoop zur Entlastung des EDW eingesetzt haben (vgl. Abbildung 53). Von Vorteil ist dabei, dass

- die Unternehmen ihr EDW nicht ersetzen müssen, denn Hadoop ergänzt ihre vorhandene Lösung;
- eine Reihe von EDW Herstellern eine Hadoop-Distribution in ihre EDW-Appliance eingebettet haben.

Hadoop am Anfang sowie am Ende des Daten-Lebenszyklus

Hadoop wird in absehbarer Zeit für das Data Warehouse vieler Unternehmen an Bedeutung zunehmen. Das EDW behält jedoch zunächst seine zentrale Rolle. Und so wird bereits für 2015 prognostiziert, dass mehr als die Hälfte der neu entstehenden Daten von Unternehmen

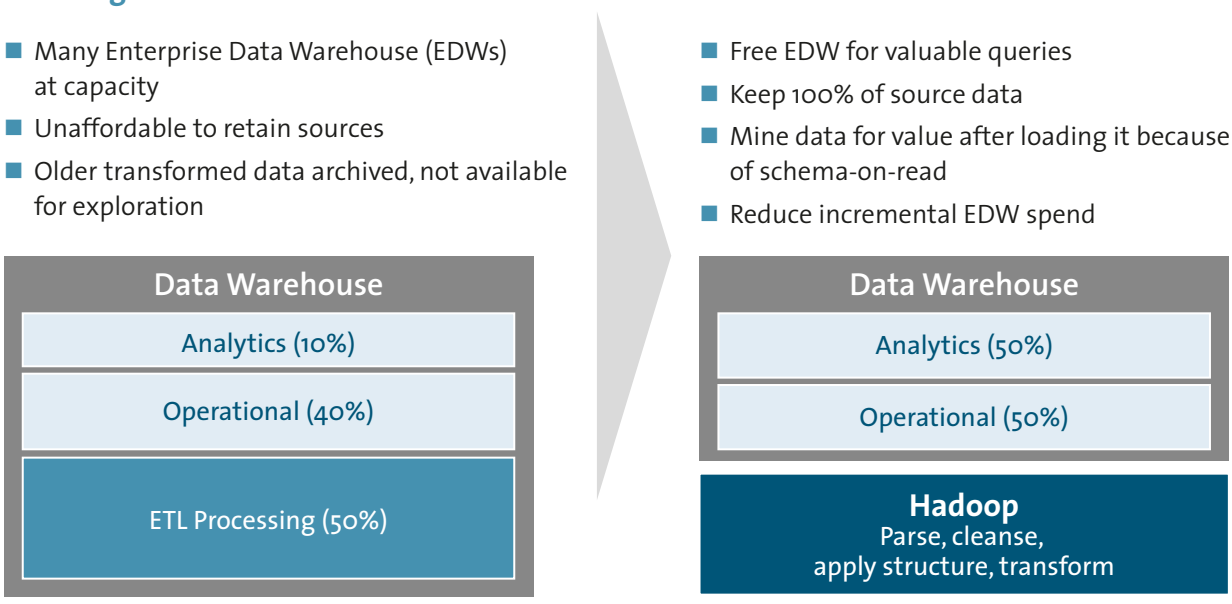


Abbildung 53: EDW-Entlastung – Einsatz-Szenario für Hadoop

erst einmal in einem Hadoop-Cluster Aufnahme finden. Hadoop-Cluster werden so de facto zum unternehmensweiten Eintrittstor für umfangreiche Datenmengen. Ein Hadoop-Cluster kann sowohl am Beginn des Daten-Lebenszyklus, quasi als Einflugschneise für Daten, als auch an seinem Ende eingesetzt werden – zur Archivierung für eine spätere Analyse.

Vielfältige Datentypen auf Hadoop

Es lohnt, sich die Sprengkraft von Hadoop genau vor Augen zu halten: Hadoop kommt mit jedem Datentyp klar – auch mit unstrukturierten Daten, dem am schnellstens wachsenden und vielleicht wichtigsten Datentyp. Die relationalen Datenbanken können unstrukturierte Daten nicht verarbeiten (vgl. Unterabschnitt 4.1.2). Wo eine relationale Datenbank eine vorher definierte, fixierte Struktur für die Speicherung von Daten voraussetzt, benötigt Hadoop lediglich eine Struktur zum Zeitpunkt der Analyse der Daten. Das bedeutet, dass ein Unternehmen strukturierte und unstrukturierte Daten erst einmal speichert und erst später beliebige Fragen zur Analyse stellen kann. Hadoop bietet eine unvergleichliche Flexibilität bei der Sammlung, Speicherung und Analyse großer Datenmengen – egal, ob es sich um Daten von Produktionsprozessen, Sensoren, Kundentransaktionen, von mobilen Endgeräten oder Social-Media-Plattformen handelt. Gleichermaßen wichtig ist die Fähigkeit von Hadoop zur Parallelverarbeitung und Skalierung. Mit dem Hadoop Framework werden Nutzer, die sich intensiv mit Hadoop befassen, bei der Speicherung und Transformation von Daten kaum an Grenzen stoßen, und so unterstützt Hadoop die Wettbewerbsfähigkeit von Unternehmen im Big-Data-Zeitalter.

Ein neuer Prozess: von ETL zu ELT

Hadoop verfolgt einen neuen Ansatz beim Umgang mit Daten. Traditionell werden die Quelldaten aus den Vorsystemen in einer Staging Area zunächst geladen, um dann in einem weiteren Schritt validiert und bereinigt zu werden, bis die verdichteten Daten schließlich über eine ODS-Schicht in das eigentliche EDW gelangen. Im Big-Data-Kontext stellt sich die Verarbeitung jedoch anders da. Hier werden die Ausgangsdaten aus den Vorsystemen 1:1 im Original Format im Hadoop Distributed File System (HDFS) dreifach redundant gespeichert. Die Bearbeitung der Rohdaten ist natürlich auch in dem Big-Data-Kontext notwendig. Um aus den polystrukturierten Daten brauchbare Erkenntnisse gewinnen zu können, ist auch hier eine Transformation notwendig, allerdings erst zum Zeitpunkt der Analyse. Daher nennt man den Prozess in Hadoop Extract, Load, Transform, im Gegensatz zum klassischen Extract, Transform, Load. Bei Hadoop werden die Transformationsprogramme über das MapReduce Verfahren auf alle verfügbaren Compute-Knoten des Hadoop-Clusters verteilt, ausgeführt und massiv-parallel in ein Zielverzeichnis abgelegt.

Dieser Prozess läuft als Stapelverarbeitung ab und skaliert automatisch mit steigender Anzahl der Compute-Nodes. Im klassischen EDW werden die Daten auf einem vorgelagerten Extraktions/Transformations/Lade Server verarbeitet. Damit dieser ETL-Server seine Aufgaben erfüllen kann, müssen die Daten aus den Quellen ausgelesen werden und anschließend wieder in der Relationalen Datenbank (RDBMS) abgelegt werden. Dieser Roundtrip ist bei steigenden Datenmengen der kritische Pfad für die gesamte Laufzeit. Allerdings ist die Software rund um diesen ETL-Server sehr ausgereift und benutzerfreundlich.
Im Kontext von Hadoop stehen unterschiedliche Script Sprachen wie Pig und Hive bzw. Scoop zur Verfügung, deren Einsatzfelder nicht immer klar abgegrenzt sind. Auch erste graphische Tools sind verfügbar, aber diese bieten noch nicht den Funktionsumfang kommerzieller ETL-Software. So sei Talend erwähnt als einer der ersten Open-Source-ETL-Implementierungen. Derzeit besteht mit Hadoop die Gefahr, dass die entwickelten ETL-Prozesse im Laufe der Zeit unübersichtlich und undurchsichtig werden, wie dies bereits bei früheren, nicht ETL-Server unterstützten EDW Implementierungen der Fall gewesen ist.

Beispiel aus der Praxis

Hervorzuheben ist, dass Hadoop ein EDW nicht ersetzt. Vielmehr schafft Hadoop Raum, damit der Nutzer sein EDW mehr für Analysen – und damit für stärker wertschöpfende Aufgaben – nutzt. Die Nutzer können die installierte EDW-Kapazität besser einsetzen und so zusätzliche EDW-Investitionen verschieben. Allein die Kosteneinsparungen sind bereits sehr überzeugend. So kostet die Speicherung von Daten im Volumen von einem Terabyte in einem traditionellen EDW circa $20,000 bis $80,000. Es gelang in dem Unternehmen[181], diese Kosten mit einem Hadoop-Cluster auf weniger als ein Zwanzigstel zu reduzieren. Von dieser Summe entfällt ein Viertel auf Hardware, ein weiteres Viertel auf Software und die verbleibende Hälfte auf Services. Die Ergebnisse sprechen für sich. Die über einen Zeitraum von fünf Jahren fälligen Zahlungen an den Hersteller des vorhandenen EDW sanken von $65m auf $35m, wobei ein Großteil der verbliebenen $35m auf Wartungskosten für das EDW entfielen. Zusätzlich stieg die EDW-Performance auf das Vierfache, da das EDW nun stärker auf besonders wertschöpfende interaktive Aufgaben fokussiert werden konnte und die restlichen Aufgaben einem Hadoop-Cluster übertragen wurden.

Schnittstellen zu klassischen Datenbanken

Transparente Integration von klassischen DBMS-Systemen und Hadoop erlaubt die Nutzung der besten Verfahren aus beiden Welten, der klassischen SQL und der NoSQL-Verfahren. In Anbetracht der massiv-parallelen Architektur des HDFS und der voraussichtlich umfangreichen Datenbewegungen zwischen der klassischen RDBMS und dem HDFS ist eine massiv-parallele Schnittstelle wesentliche Voraussetzung für die Skalierung. In der Offenheit der Schnittstellen und den erzielbaren Durchsatzraten unterscheiden sich die verschiedenen Hersteller deutlich. Auch der Umfang und die detaillierten Zugriffsverfahren im Sinne der Integrationstiefe stellen deutliche Unterschiede heraus:

- Microsoft bietet zum Beispiel mit der Polybase genannten Schnittstelle zu Hadoop in dem SQL Server Parallel Data Warehouse eine massiv-parallele Schnittstelle zwischen Hadoop und der relationalen Datenbank. Polybase erlaubt einen transparenten Zugriff auf Dateien aus dem Hadoop System sowohl lesend als auch schreibend. Ferner ist Polybase vergleichbar mit dem HCATALOG aus dem Hadoop-Ökosystem, mit dem Unterschied, dass hier direkt, mittels T-SQL die Daten aus Hadoop abgefragt werden können, ohne dass diese erste in die Datenbank importiert werden müssen. Es können als über die SQL-Syntax in einer Anweisung Daten aus Hadoop und dem Relationalen SQL-Server performant abgefragt werden. Die Ergebnisse dieser Abfrage wiederum können entweder an den Client, zum Beispiel Excel geliefert werden, oder für die Weiterverarbeitung wiederum als Files im Hadoop abgelegt werden. Polybase erlaubt eine Verschmelzung der beiden Welten vom klassischen RDMBS und Hadoop in einer Plattform.

- Im Abschnitt 5.5 wurde bereits auf die Hadoop-Komponente »HCatalog« eingegangen. Anhand der in HCatalog hinterlegten Metadaten über Daten in einem Hadoop-Cluster ist es mit der SQL-H™ von Teradata möglich, auf die Daten in Hadoop »on-the-fly« aus einem SQL-Statement heraus zuzugreifen, diese mit anderen Daten im Data Warehouse zu verknüpfen und einer analytischen Engine in Teradata zuzuführen, um z. B. Klickpfade in Weblogs zu bewerten, Machine Leaning zu betreiben etc.
  SQL-H™ nutzt dabei die Metadaten aus dem HCatalog, um bei partitioniert in Hadoop abgelegten Daten nur die wirklich benötigten Informationen heranzuziehen. Das gilt auch für den Fall, dass in der Anfrage nur ein Teil der Attribute benötigt werden. Der Datentransfer zwischen Hadoop und einem Teradata-System erfolgt durch die direkte Kommunikation von den einzelnen Rechnerknoten, um maximalen Datendurchsatz zu erreichen.

[181] Das Unternehmen wollte anonym bleiben.

6 Big Data im Kontext relevanter Entwicklungen

Als Basis für die Darstellung von Big-Data-Technologien in diesem Leitfaden leistete der im Abschnitt 3.2 vorgestellte Technologie-Baukasten gute Dienste. Es gibt jedoch wichtige Entwicklungen, die eine ganze Reihe von Komponenten aus dem Baukasten betreffen. Dazu gehören Cloud Computing (Abschnitt 6.1), In-Memory Computing (Abschnitt 6.2) und Open Source (Abschnitt 6.5).

Außerdem wird auf neue Entwicklungen bei Datenanalysesystemen (Stratosphere, Abschnitt 6.4) und Frameworks (Akka, Abschnitt 6.3) hingewiesen.

6.1 Neue Chancen für Big Data durch Cloud-Dienste

Dieser Abschnitt widmet sich der Nutzung der Cloud für Big Data, ob als Softwaredienst (SaaS) für BI, als Entwicklerplattform (PaaS) oder als Speicherinfrastruktur (IaaS) für Hadoop.

Es lässt sich feststellen: Die Cloud bietet eine Vielzahl von Werkzeugen, um die Nutzung von Big Data zu vereinfachen, zu beschleunigen sowie die Kosten und Risiken zu verringern.

Stand von Cloud-BI

In den vergangenen Jahren war BI das oft übersehene Stiefkind im Bereich Cloud-Lösungen und SaaS. Wenn es auch einige erfolgreiche SaaS-Anbieter für BI gab, so waren sie dennoch Nischenanbieter – verglichen mit den vier Hauptanwendungen für SaaS: Kundenbeziehungsmanagement (CRM), Collaboration, Personalmanagement (HCM) und Beschaffung.

Mit der zunehmenden Reife von Cloud-Technologien und dem breiteren Verständnis ihrer Vorteile[182] wird der Einsatz von Public- und Virtual-Private-Cloud-Lösungen für BI in den kommenden Jahren jedoch sicher zu nehmen. 37 % der Unternehmen, die Forrester Research in einer Umfrage zu dem Thema Ende 2013 befragt hatte, planen eine Ergänzung ihrer bestehenden BI-Lösung durch eine externe Cloud-Lösung, und 28 % wollten sie sogar vollständig durch eine Cloud-Lösung ersetzen.[183] Big Data ist in diesem Zusammenhang ein wichtiger Katalysator für das wachsende Interesse an Datenmanagement und analyse in der Cloud.

Big Data an ihrem Ursprung lassen

Eine ganze Reihe von Big-Data-Szenarien beruhen auf internen Daten, wie zum Beispiel Sensordaten von Fertigungsstraßen oder Videodaten von Sicherheitssystemen. In diesem Falle sprechen drei Gründe dafür, die Daten und ihre Verarbeitung im eigenen Rechenzentrum zu belassen:

- Die Daten sind zu sensibel, um sie einem Verlustrisiko auszusetzen, zum Beispiel Testserien oder Wartungsdaten von Maschinen oder auch Lage- oder Patientendaten.
- Die Daten sind zu umfangreich und zu volatil, um sie schnell genug in eine Cloud-Umgebung hoch zu laden, zum Beispiel bei Sensordaten.
- Zu hohe organisatorische Hürden innerhalb des Unternehmens verzögern eine rechtzeitige Verlagerung der Daten in die Cloud.

[182] vgl. zum Beispiel den Report »The Changing Cloud Agenda«, http://www.forrester.com/The+Changing+Cloud+Agenda/fulltext/-/E-RES72363?intcmp=blog:forrlink

[183] Daten aus Forrester's Q4, 2013 Forrsights Software Survey

In vielen anderen Anwendungsszenarien für Big Data kommt allerdings ein Großteil der Daten von außerhalb des Unternehmens oder wird mit diesen Daten angereichert, zum Beispiel sozialen Medien, demographischen Daten, Webdaten, Ereignissen, Feeds etc.

Big Data in der Cloud

In sozialen Medien ist nur ein Bruchteil der Daten relevant zum Beispiel für die Sentimentanalyse. 20 % aller Tweets beinhalten einen Link, den man öffnen muss um den Kontext zu verstehen. Riesige Volumina externer Daten müssen gefiltert, formatiert und für die weitere Analyse vorbereitet werden. Nach der Analyse muss häufig nur das aggregierte Ergebnis gespeichert werden (zum Beispiel der Klout Score in 4.3.7). Selten benötigt man die Datenquelle für Audit oder weitere Analyse. Alle Tweets der letzten zwei Jahre benötigen 0,5 PetaByte Speicher. Es ist wenig sinnvoll für ein Unternehmen, diese Rohdaten in seinem Rechenzentrum zu speichern.

Abhängig vom Anwendungsfall gibt es somit auch gute Gründe, warum Big Data in entsprechenden Fällen in der Cloud verbleiben sollte:

- Big Data erfordert ein ganzes Spektrum neuer Technologien, Fertigkeiten und Investitionen. Jedes Unternehmen muss sich fragen, ob es das wirklich alles in seinem Rechenzentrum braucht und entsprechend investieren will?

- Big Data beinhaltet oft riesige Mengen externer Daten. Ist es wirklich sinnvoll, diese Daten innerhalb der Unternehmes-Firewall zu speichern und zu verwalten?

- Je nach Anwendungsszenario wird möglicherweise technisches Know-how (z. B. im Bereich Data Science) benötigt, das im eigenen Unternehmen schwierig aufzubauen ist.

- Big Data erfordert eine Vielzahl von Dienstleistungen. Anwender werden sich eher auf die differenzierte Analyse großer Daten konzentrieren wollen – und weniger auf deren Verwaltung.

Folgerichtig bieten eine Reihe von Anbietern Lösungen für Big Data in der Cloud.

- Big-Data-Plattformen bieten diverse Dienste von der Speicherung großer Datenmengen bis zu dynamisch buchbarer Rechenkapazität für schnelle Analysen und Simulationen. Das Ergebnis wird in der Cloud gespeichert, nachdem der Rechencluster heruntergefahren wird.[184]

- Big Data Services konzentrieren sich auf die Vorbereitung, Anreicherung oder Verknüpfung verschiedener Datenquellen, ihre Analyse und einfacher Visualisierung.[185]

[184] Kombinationen bieten zum Beispiel Amazon mit S3 und EMR oder Microsoft mit Azure und HD Insight
[185] zum Beispiel Microsoft Power BI

6.2 In-Memory Computing

In-Memory Computing umfasst eine Anzahl von Technologien, die sich in unterschiedlichen Big-Data-Komponenten wiederfinden. Durch die Verlagerung der Datenverarbeitung von der Festplatte in den Rechner-Hauptspeicher (In-Memory) können Big-Data-Visualisierungen, -Analysen oder-Transaktionen massiv beschleunigt werden. Somit kann der geschäftliche Mehrwert schneller erbracht werden.

Historische Einordnung

Der Begriff In-Memory beschreibt ein Konzept, bei dem die Daten nicht – wie bei Systemen zur Verarbeitung von großen Datenmengen üblich – auf der Festplatte gespeichert und verarbeitet werden, sondern im Hauptspeicher. Dies hat den Vorteil, dass Zugriffe auf die Daten wesentlich schneller sind als bei althergebrachten Herangehensweisen.
Die Verarbeitung der Daten im Hauptspeicher ist, historisch gesehen, die einzige Art, da die klassische Von-Neumann-Architektur keine Festplatten kannte. In dieser Urform des Computers gab es nur den einen Speicher, den (Haupt-) Speicher. Musste ein Programm Daten verarbeiten, so griff es direkt auf die Stelle im Speicher zu, an der die Daten abgespeichert waren. Der Speicher erlaubt also Zugriffe auf jede Speicherstelle in beliebiger Reihenfolge, daher auch der Begriff Random Access Memory (RAM).

Bei sehr großen Datenmengen kommt eine reine hauptspeicherbasierte Herangehensweise jedoch schnell an ihre Grenzen, da der Speicher üblicherweise für den ganzen Datensatz nicht ausreicht. Ein gangbarer Weg ist, nur Teile des Datensatzes in den Hauptspeicher zu laden und diese unabhängig voneinander zu verarbeiten. Dies können, wie zum Beispiel bei Datenbanken üblich, blockbasierte Verfahren sein oder im Fall von analytischen Berechnungen spezielle Big-Data-Algorithmen. Der Vorteil dabei ist, dass die Größe der zu verarbeitenden Daten nicht mehr durch den Hauptspeicher begrenzt ist, sondern annähernd beliebig groß werden kann. Der Nachteil ist jedoch, dass die Performance aus zwei Gründen drastisch sinkt:

- einerseits, weil der Zugriff auf sekundäre Speichermedien deutlich langsamer ist als auf den Hauptspeicher und
- andererseits, weil nicht die schnellsten Algorithmen eingesetzt werden, sondern die, die am besten mit Datenblöcken umgehen können.

In den letzten Jahren sind die Preise für Hauptspeicher kontinuierlich gesunken, gleichzeitig ist die Leistungsfähigkeit der eingesetzten Netzwerkkomponenten enorm angestiegen, genauso wie das Know-how über die Verteilung von Berechnungen. Zusammen eröffneten diese Entwicklungen die Möglichkeit, große Datenmengen, verteilt auf mehrere Rechnerknoten, im Hauptspeicher zu verarbeiten.

Typen von In-Memory Datenhaltung

In-Memory spielt im Big-Data-Umfeld eine besondere Rolle, da erst durch den schnellen Zugriff auf die Daten typische Big-Data-Herangehensweisen möglich werden. So erfordert zum Beispiel ein explanatives Vorgehen Zugriffszeiten, die ein Benutzer noch als akzeptabel empfindet. Dabei gibt es unterschiedliche Typen und Szenarien für den Einsatz von In-Memory-Technologien.
Die ausschließliche In-Memory-Verarbeitung der Daten hat aber auch Nachteile. So ist es, trotz sinkender Hauptspeicherpreise, immer noch sehr teuer, alle Daten im RAM zu halten, und außerdem sind die Daten im flüchtigen Hauptspeicher nicht notwendigerweise vor einem Systemausfall geschützt. Dafür müssten sie auf der persistenten Festplatte liegen. In der Praxis haben sich daher unterschiedliche In-Memory-Varianten entwickelt. Neben den reinen In-Memory-Systemen gibt es unterschiedliche Grade an Hybrid-Systemen.
Bei den In-Memory-Systemen haben sich zwei Herangehensweisen herauskristallisiert:

- **die reinen In-Memory-Systeme (z. B. In-Memory Datenbanken und Data-Grids):**
  Sie speichern alle Daten im Hauptspeichern und nutzen die Festplatte nur als persistenten Speicher (z. B. um die Ausfallsicherheit zu erhöhen).

- **die Hybrid In-Memory-Systeme:**
  Sie speichern die Daten teils auf der Festplatte und teils im Hauptspeicher.

Reine In-Memory-Systeme

Reine oder native In-Memory-Systeme haben alle Daten im Hauptspeicher. Auf diese Weise kann sehr schnell auf alle Daten zugegriffen werden. Dies erfordert jedoch auch, dass der gesamte Datensatz in den Hauptspeicher passt und darüber hinaus noch genügend Platz für Verarbeitungsstrukturen wie zum Beispiel Indizes ist.

Hauptspeicher ist flüchtig. Das bedeutet, dass Daten, die im Hauptspeicher liegen, nach einem Neustart[186] des Systems nicht mehr zur Verfügung stehen. Für In-Memory-Systeme heißt das, dass – selbst wenn alle Daten im Hauptspeicher verarbeitet werden – es doch notwendig ist, eine Sicherung auf der Festplatte vorzunehmen bzw. die Verteilung auf ein ausfallsicheres Rechner- und Hauptspeicher-Cluster.

Hybrid In-Memory-Systeme

Eine Alternative zu reinen In-Memory-Systemen stellen Hybridsysteme dar. Sie speichern die Daten auf der Festplatte und verarbeiten einen Teil im Hauptspeicher. Dabei sind diese Systeme meist so konzipiert, dass nur die relevanten Daten im Hauptspeicher liegen. Da die relevanten Daten als Hot-Data und die weniger relevanten als Cold-Data bezeichnet werden, spricht man hier auch von einem Temperatur-Modell.

Das Temperatur-Modell hat den Vorteil, dass nur die Daten im teuren Hauptspeicher verarbeitet werden, die wirklich auch benötigt werden. Gerade bei analytischen Abfragen ist oft nur ein Bruchteil der Daten wirklich relevant. Das bedeutet aber auch, dass in einem reinen In-Memory-Modell viele Daten im Hauptspeicher liegen, die nur selten verwendet werden. Diese Herangehensweise wird von einigen Herstellern auch noch um weitere Schichten zwischen Hauptspeicher und Festplatte ergänzt. So kann zum Beispiel eine Zwischensicht aus Solid State Disks den Wechsel der Daten aus einem Cold-Zustand in einen Hot-Zustand beschleunigen.

Der Einsatz von In-Memory kann als Zugriffsbeschleuniger für Datenbanken, Data Warehouses oder Hadoop gesehen werden (vgl. Abbildung 54).

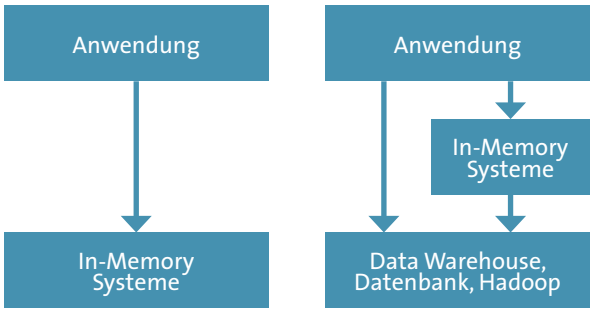


Abbildung 54: Native und hybride In-Memory-Systeme

Mit ihrem schnellen Zugriff auf große Datenmengen ermöglichen In-Memory-Systeme vollkommen neue Anwendungen.

Analytical In-Memory Computing in Datenbanken

Da der Hauptspeicher jedoch erstens flüchtig, also beim Ausschalten des Computers auch alle Daten verschwinden, und zweiten sehr teuer ist, hat man den Hauptspeicher um sekundäre Speichermedien, zuerst Bänder und dann Festplatten, erweitert. Diese dienten der Auslagerung und der persistenten Speicherung selten genutzter Daten.

Datenbanken nutzten fast ausschließlich die sekundären Speichermedien, da die Datenmengen, die mit ihnen verwaltet wurden, so groß waren, dass der Hauptspeicher sie nicht fassen konnte und ein Aufrüsten der Computer mit mehr Hauptspeicher zu teuer gewesen wäre. Aus diesem Grund wurden Datenbanken so programmiert, dass sie auf den Zugriff auf diese Speichermedien hin optimiert waren. Sie verwendeten zum Beispiel Algorithmen, die nicht auf einzelnen Datensätzen arbeiteten, sondern auf

[186] gewollten oder ungewollten

Blöcken, bestehend aus einer ganzen Gruppe von Datensätzen. Diese Blöcke konnten dann sequenziell gelesen werden, was bei Bändern und Festplatten verhältnismäßig schnell geht.

In den letzten Jahren sind die Preise für Hauptspeicher jedoch so stark gefallen, dass es möglich ist, alle Daten einer Datenbank in den Hauptspeicher eines Computers oder mehrerer[187] Computer zu laden. Das hat gleich mehrere Vorteile, zuerst einmal wird der Zugriff auf die Daten enorm beschleunigt[188] und darüber hinaus können wesentlich effizientere Algorithmen verwendet werden, die nicht auf Blöcken, sondern auf den einzelnen Datensätzen arbeiten.
Nicht nur die Verarbeitung, auch die Speicherung großer Datenmengen im Hauptspeicher hat in den letzten Jahren immer stärker an Bedeutung gewonnen. Gerade analytische Datenbanken – also Datenbanken, die hauptsächlich zur Analyse von Datenbeständen verwendet werden – benötigen schnellen Zugriff auf die zugrundeliegenden Datenbestände, da für die Analyse meist große Teile eines Datensatzes gelesen werden müssen. Die ist jedoch nur durch die Speicherung im Hauptspeicher gewährleistet. Datenbanken mit In-Memory Technologien gehen jedoch weit über die Speicherung der Daten im Hauptspeicher hinaus. Effizientere Algorithmen gepaart mit Datenstrukturen, die mit Blick auf Analysen optimiert wurden, bringen enorme Performanz-Gewinne gegenüber transaktionalen Datenbanksystemen. Einer besonderen Bedeutung kommt dabei die spaltenbasierte Speicherung zu. Sie ermöglicht eine besonders effiziente Datenspeicherung, die gezielt auf die Bedürfnisse von Analysen ausgerichtet ist. Fehler! Verweisquelle konnte nicht gefunden werden. zeigt den Unterschied zwischen spaltenbasierter und zeilenbasierter Speicherung. Bei der ersten Art erfordern Aggregationsoperationen wie zum Beispiel die Mittelwertbildung nur einen Zugriff auf einen Datensatz (die gewünschte Spalte), wobei die zweite Art jede Zeile als individuellen Datensatz betrachten muss.

Die volle Leistungsfähigkeit spielen analytische Datenbanken jedoch erst aus, wenn neben In-Memory und spaltenbasierter Speicherung noch die massiv-parallele Verarbeitung in einem Cluster hinzukommt. Erst dieser Dreiklang kombiniert die Stärken aller drei Technologien zu einem hochperformanten System, das auch mit größten Datenbanken keine Probleme hat.

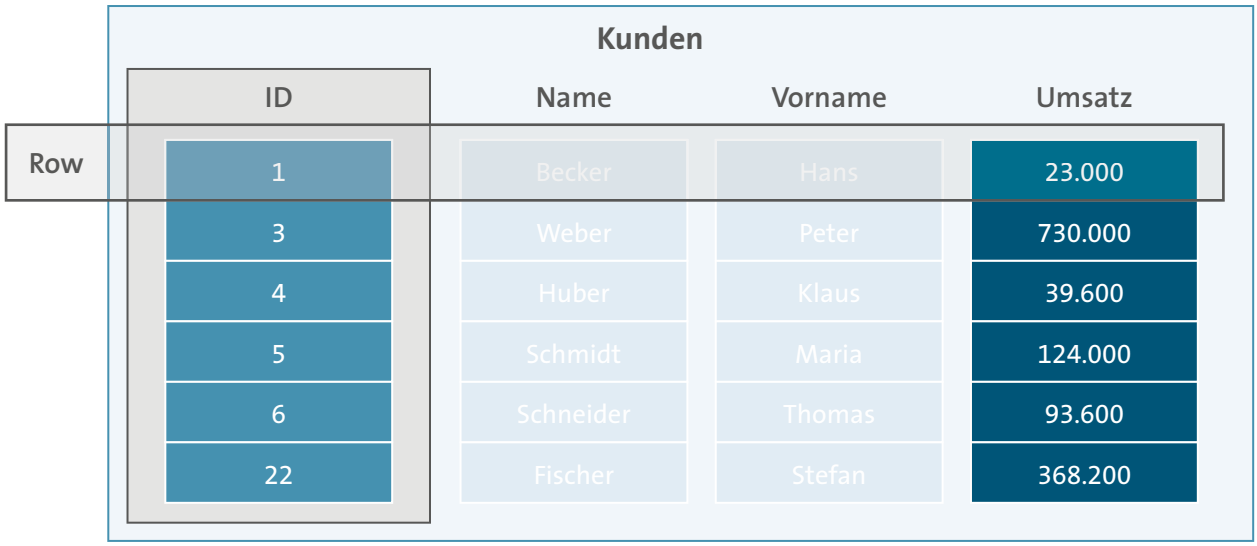


Abbildung 55: Spalten- und zeilenbasierte Speicherung

[187] im Falle einer Clusterdatenbank
[188] der Zugriff auf die Festplatte erfordert üblicherweise mehrere Größenordnungen mehr CPU-Zyklen als der Zugriff auf den Hauptspeicher

Gerade im Webumfeld oder im Wertpapierhandel haben sich Anwendungsfälle entwickelt, die besondere Ansprüche an die Antwortzeiten von transaktionalen Datenbanken haben und in kürzester Zeit sehr viele Transaktionen durchführen müssen. Ähnlich verhält es sich mit dem automatisierten oder algorithmischen Handel von Wertpapieren. Dabei kommt es auf jede Millisekunde an. Um solche Antwortzeiten bei Datenbanken zu gewährleisten, werden Operationen nur noch im Hauptspeicher durchgeführt und erst zeitversetzt persistent auf die Festplatte geschrieben. Da dieser Anwendungsfall sich grundsätzlich von den zuvor beschriebenen Anwendungsfällen der analytischen Datenbanken unterscheidet, kommen auch grundverschiedene Technologien zum Einsatz.
In transaktionalen Systemen bietet zum Beispiel die spaltenbasierte Speicherung keinerlei Vorteile, auch ist die massiv-parallele Verarbeitung eher ein Hindernis, da Transaktionen erst über mehrere Rechner synchronisiert werden müssen.

6.3 Akka und Scala

Die im Abschnitt 3.3 aufgezeigten Spannungsgebiete Verteilungsfähigkeit, Konstruktion und Anwendungsarchitektur stellen besondere Anforderungen an Software-Frameworks. Dazu gehören:

- **belastbar bzw. widerstandsfähig**
  (hohe Ausfallsicherheit bei verteilter Verarbeitung durch automatische Fehlerbehebungsmechanismen),
- **skalierbar**
  (Erhöhung der Kapazität ohne Anpassung des Programmcodes),
- **ereignisgetrieben**
  (in der Verarbeitungsebene und in Echtzeitanalyse),
- **benutzerorientiert**
  (kurze Antwortzeiten, hohe Erreichbarkeit).

Entscheidende Grundlage: Aktorenmodell

In einzelnen Fällen werden nicht Standardarchitekturen, sondern individuelle Strukturen implementiert. Das eingesetzte Framework bestimmt den Lösungsraum der Architektur.

Das Aktorenmodell erfüllt alle vier zuvor genannten Anforderungen. Schon seit 1973[189] bekannt, zeigte es 1995 seinen kommerziellen Nutzen[190]: Das erstellte Produkt[191] wies eine Verfügbarkeit von neun Neunen auf (= 99,9999999% Verfügbarkeit) – weniger als 30 Millisekunden Ausfallzeit im Jahr.

Akka

Das Open Source Framework Akka[192] bildet dieses Aktorenmodell ab und erweitert es aufgrund aktueller Erkenntnisse. Es ist seit 2011 im kommerziellen Einsatz. Die Implementierung auf der JVM (in Scala, vgl. S. 131) führt dabei zur Fähigkeit, hochintegrierbar zu sein.

[189] 1973 von Carl Hewitt vorgeschlagen
[190] Die Architektur war damals in Erlang realisiert worden.
[191] Switch AXD 301 von Ericsson
[192] www.akka.io

Die Eigenschaften, die Aktoren in all diesen Bereichen interessant machen, sind ihre inhärente Verteilbarkeit[193], Widerstandsfähigkeit[194] und Skalierbarkeit[195]. Das treibende Prinzip hinter diesen Eigenschaften ist, dass die resultierende Architektur von Aktor-Applikationen eine inhärente Robustheit aufweist[196].

Das Aktorenmodell stellt einen universellen Programmieransatz dar, der als Grundlage zur Erstellung konkreter Lösungen dient. Aufbauend auf der elementaren Kommunikation zwischen Aktoren stehen komplexere Interaktionsmuster zur Verfügung, die bei Bedarf individuell erweitert werden. So können auch existierende Programmteile oder Bibliotheken eingebunden werden.

Scala

Die objekt-funktionale Programmiersprache Scala (Name leitet sich von »scalable language« ab) für die Java Virtual Machine (JVM). Damit kann sie auf alle Java-Bibliotheken zugreifen, mit Java Frameworks zusammen arbeiten und ist somit hochgradig integrierbar. Scala erlaubt den kompakten Sprachkern durch eigene Operatoren zu erweitern und insbesondere eigene DSL[197] zu erstellen. Die funktionale Programmierung erlaubt Programme aus Funktionen aufzubauen und damit Nebenwirkungen der imperativen Programmierung zu vermeiden. In kommerziellen Anwendungen wird Scala seit Jahren genutzt. Die Fähigkeit zu skalieren führte u. a. zur Lösung Twitter[198]. Im Abschnitt 6.4 wird mit Stratosphere eine weitere Lösung aufgezeigt, die Scala in Integration im Big-Data-Umfeld anwendet.

Akka und Scala

Die funktionalen Eigenschaften von Scala[199] stellen die Basis für zielorientierte und elegante Elemente zur parallelen Datenverarbeitung. Akka als Framework stellt über das Aktorenmodell den integrativen Rahmen für die Nutzung in Big-Data-Produkten. Dies wird zum Beispiel vom Spark-Framework[200] intensiv genutzt – einem vielversprechenden Wettbewerber zu Hadoop und MapReduce.

Weitere Potenziale

Viele andere angebotenen Lösungsansätze ermöglichen eine Parallelisierung durch die Aufteilung des Datenvolumens, durch die Verteilung des Datenvolumens auf Netzwerkknoten sowie durch eine Verwaltung von Aufgaben nach dem Muster der Stapelverarbeitung: Eine Anfrage wird an alle Knoten geschickt, die relevante Daten besitzen, diese bearbeiten ihren jeweiligen Anteil und die Ergebnisse werden gesammelt oder zur weiteren Verarbeitung abgelegt. Diese Lösungsansätze jedoch sind nicht geeignet für die fortlaufende Analyse von Daten – insbesondere von Big-Data-Datenströmen – in Echtzeit.

Betrachtet man dagegen die Fähigkeit, Streams in Scala in Kombination mit Akka abzubilden, so steht hier ein flexibler Rahmen für die Bearbeitung von Datenströmen zur Verfügung. Angereichert wird dies in Akka durch die im Kern ereignisgetriebene Architektur, die feingliedrige Parallelität auf der Ebene einzelner Ereignisse bietet.

---

[193] aufgrund ihrer Kapselung und der Verwendung von rein asynchroner Kommunikation
[194] Jeder Aktor wird von seinem Supervisor überwacht.
[195] da Synchronisation aufgrund der Kapselung nicht erforderlich ist
[196] Eine einführende Betrachtung aus dieser Perspektive findet sich in dem Artikel »Creating Resilient Software with Akka«. InfoQ, 6. Juni 2013 http://www.infoq.com/articles/resilient-software-with-akka
[197] Domain Specific Languages, d.h. formale Sprachen, die für ein bestimmtes Problemfeld entworfen werden
[198] Wikipedia: Twitter (Referenz 28)
[199] Stichwort Lambda-Kalkül: http://de.wikipedia.org/wiki/Lambda-Kalkül
[200] http://spark.incubator.apache.org/

6.4 Stratosphere: Beitrag der europäischen Forschung zur Big-Data-Plattformentwicklung

Stratosphere[201] ist ein skalierbares Datenanalysesystem der nächsten Generation, welches durch eine sogenannte deklarative Spezifikation einem Data Scientist die einfache Erstellung von komplexen Datenanalyseprogrammen ermöglicht und diese Programme durch automatische Parallelisierung, Optimierung und Hardwareadaption höchstskalierbar verarbeitet.

Stratosphere als Open-Source-Plattform für Big Data Analysis auf hochparallelen Clustern ist aus einem gemeinsamen Forschungsprojekt der Technischen Universität Berlin, des Hasso-Plattner-Instituts an der Universität Potsdam und der Humboldt-Universität Berlin sowie weiterer Partner in Europa im Rahmen der Information und Communication Technology Labs (ICT Labs) des Europäischen Instituts für Innovation und Technologie (EIT) hervorgegangen.

Stratosphere integriert sich nahtlos in Big-Data-Infrastrukturen, die auf dem Hadoop-System basieren und unterstützt u.a. das Dateisystem von Hadoop (HDFS) und dessen Ressourcenmanager (YARN), ersetzt jedoch die Programmierabstraktion und die Laufzeitumgebung durch ein erweitertes MapReduce-Programmiermodell, das sogenannte PACT-Modell. In diesem Modell können in Java und Scala neben klassischen MapReduce-Programmen auch komplexere Operatoren zum Verbinden von Datenströmen sowie zur Ausführung von iterativen Algorithmen angesprochen werden, was die Erstellung von skalierbaren Data Mining und Predictive Analytics erheblich vereinfacht.

Stratosphere bietet Schnittstellen für relationale Datenbanken (JDBC) und Graph-Analyse (Pregel/Spargel) an und unterstützt im PACT-Modell komplexe Algorithmen (Maschinelles Lernen, Graph Mining, Text Mining, etc.), welche in klassischen MapReduce-Systemen oder in relationalen Datenbanken (SQL) nicht bzw. nur mit viel Programmieraufwand oder nicht skalierbar realisiert werden können.

Stratosphere als Datenanalysesystem der nächsten Generation ist als ein Software Stack aus Komponenten angelegt, welcher die in-situ Analyse von Daten aus unterschiedlichen Datenquellen ohne Erfordernis von ETL-Prozessen ermöglicht (vgl. Abbildung 56).

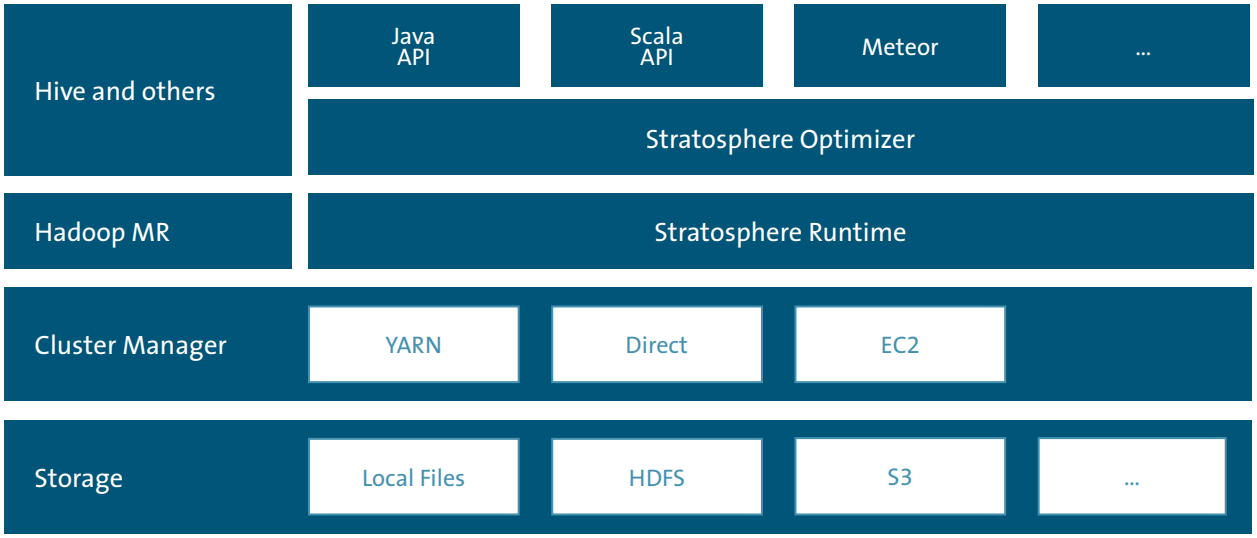


Abbildung 56: Stratosphere Software Stack

[201] www.stratosphere.eu

Auf der obersten Ebene bietet Stratosphere Programmierschnittstellen in den Programmiersprachen Java und Scala an, in denen Datenanalyseprogramme gemäß des PACT-Modells spezifiziert werden können. Durch die Verwendung dieser populären Sprachen ist eine große Menge an Entwicklern in der Lage, Datenanalyseprogramme zu erstellen. Darüber hinaus stellt Stratosphere mit Meteor und Spargel weitere Progammierschnittstellen bereit. Meteor ist eine erweiterbare Skriptsprache für einfache Datenanalyseaufgaben. Spargel bietet das populäre knoten-zentrierte Programmiermodell für graphstrukturierte Daten an. Stratosphere parallelisiert diese Analyseprogramme automatisch und optimiert ihre Ausführung in Abhängigkeit von Systemumgebung, Workload und Eigenschaften der zu analysierenden Daten.

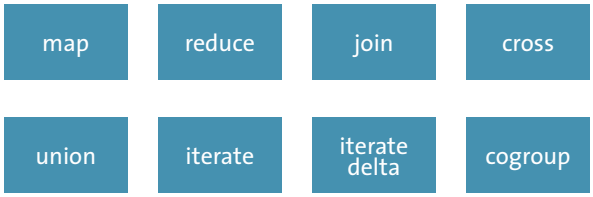


Abbildung 57: Stratosphere-Operatoren

Hierzu werden neuartige Verfahren der Codeanalyse, Codegenerierung und Datenbankoptimierung angewendet, die im Rahmen der Hochtechnologieforschung speziell für Stratosphere entwickelt wurden. Durch diese Optimierungen können auch Nutzer ohne Systemprogrammiererfahrung das System effizient nutzen. Stratospheres PACT-Programmiermodell erweitert zudem das von Hadoop bekannte MapReduce-Modell durch neue Operatoren und gibt damit Data Scientists einen umfangreichen Werkzeugkasten zur Lösung von Datenanalyseproblemen auf Big Data (vgl. Abbildung 57)

Die neben map und reduce zusätzlichen Operatoren im PACT-Modell ermöglichen die einfache Spezifikation von komplexen Datenanalyseprogrammen, wie sie im Data Mining, Maschinellen Lernen und komplexer Statistik üblich sind (z. B. Regression, Clustering, Graph-Analyse). Die Operatoren join, cross, union, und cogroup erlauben dabei die Verknüpfung, Verbindung oder Korrelation von mehreren Datenströmen, welche in map/reduce-Systemen wie Hadoop nur mit großem Aufwand möglich sind.

Insbesondere hervorzuheben sind die Operatoren iterate und iterate-delta, da diese Stratosphere von den meisten anderen kommerziellen und nichtkommerziellen Systemen unterscheiden: Algorithmen des Data Minings, des maschinellen Lernens und der Graph-Verarbeitung erfordern häufig, dass die Daten mehrfach durchlaufen werden. Stratosphere unterstützt mit iterate und iterate-delta hierzu nativ skalierbare iterative Algorithmen und kann diese im Gegensatz zu Hadoop, SQL-Datenbanken oder neueren Systemen wie Spark automatisch parallelisieren und effizient auf Rechenclustern verarbeiten.

Neben den erweiterten Operatoren ist Stratosphere auch in der Lage, komplexe Datenflüsse abzubilden. Während bei Hadoop MapReduce ein vorgegebener Datenfluss von Map nach Reduce vorgegeben ist, ist Stratosphere in der Lage, Daten in vielfältiger Weise zwischen den Operatoren zu senden.

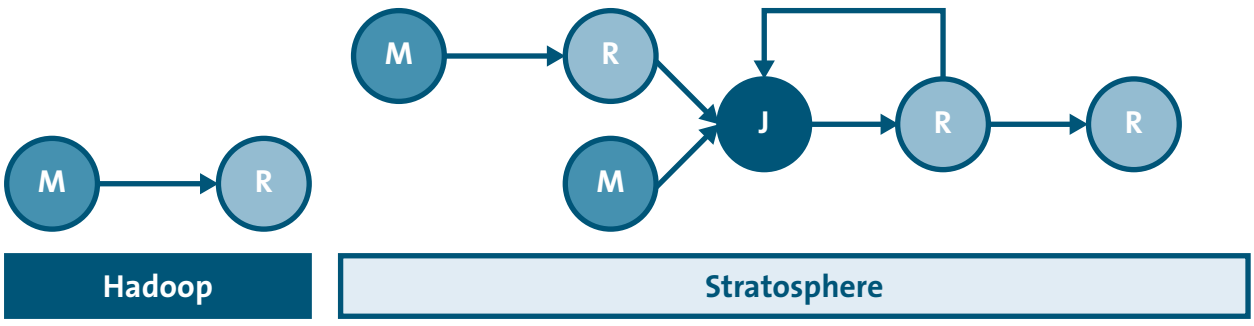


Abbildung 58: Stratosphere – Möglichkeit komplexer Datenflüsse

Abbildung 58 zeigt einen Datenfluss der aus mehreren mappern, reducern und einem join operator besteht. Diese Datenflüsse erlauben es, mit Stratosphere auch komplexe Daten-Integrationsaufgabe auszuführen.

Insgesamt sind seit 2008 mehr als 15 Mann-Jahre Forschung und Entwicklung in das Stratosphere-Open-Source-System eingeflossen. Stratosphere wird derzeit von einer aktiven europäischen Open Source Community weiterentwickelt und wird bereits in mehreren Anwendungen eingesetzt, zum einen im Kontext von Verbundprojekten gefördert durch das BMWi, das BMBF und die EU, zum anderen in direkten Partnerschaften von Unternehmen mit Stratosphere.

Ferner ist das Stratosphere-System Bestandteil des Smart Data Innovation Lab (www.sdil.de) sowie die strategische Flaggschiff-Plattform zur Datenanalyse des Europäischen Instituts für Technologie (EIT). Das Stratosphere-System ist unter der Apache 2.0 Lizenz verfügbar.[202]

6.5 Big Data und Open Source – Strategische Weichenstellungen

In drei bis fünf Jahren wird ein Großteil aller neuen Daten nicht in traditionellen, sondern in neuen Speicherlösungen wie z. B. Hadoop und Object Storage landen. Open-Source-Lösungen wie Hadoop spielen somit eine wichtige Rolle im Big-Data-Ökosystem. Für einige Bausteine einer Big-Data-Lösung sind neben den reinen Open-Source-Lösungen auch kommerzielle Lösungen am Markt, die auf Open-Source aufsetzen (vgl. Tabelle 14).

Baustein	Lösungen
Hadoop	Hortonworks, Cloudera
ETL & Analytics	Talend, Pentaho, Jaspersoft
Data Visualization	ggplot2, OpenDX

Tabelle 14: Kommerzielle Lösungen (Auswahl) auf Open-Source-Basis

Bei einem genaueren Blick auf die Chancen und Risiken von Hadoop wird ersichtlich, dass sich das konsequente Vertrauen auf Open Source durchaus lohnen kann. So lässt sich z. B. ein Vendor Lock-In – die zwingende Bindung an einen Hersteller – verhindern. Zudem läßt sich wertvoller Spielraum bei Betriebskosten, strategischen IT-Investitionen und einem Innovationstempo bewahren, welches in der Regel wesentlich höher ist als bei proprietären Lösungen. Als historisches Beispiel sei hier Linux genannt: Nachdem Linux eine kritische Masse erreicht hatte, war der Fortschritt hier viel rascher als in den von einzelnen Anbietern kontrollierten Unix-Silos. Der größte Teil der Arbeit an und um Linux wird weiterhin von der Nutzergemeinde erledigt. Wer ins Jahr 2014 springt, wird viele Parallelen zu Hadoop erkennen.
Der genaue Marktanteil der Hadoop- beziehungsweise Object-Storage-Lösungen lässt sich schwer voraussagen. Aber kaum ein Experte zweifelt am generellen Trend, dass

[202] Weitere Informationen mit Anwendungsbeispielen, umfangreiche Dokumentation und Technologie-Whitepapers, sowie ein Installations-Image für verschiedene virtuelle Maschinen, der Source-Code des Systems und Kontaktinformationen finden sich unter http://www.stratosphere.eu

dem Open-Source-Emporkömmling Hadoop glänzende Zeiten bevorstehen. Denn angesichts des Hypes rund um Big Data dürfte es nur eine Frage der Zeit sein, bis mehr und mehr Unternehmen entsprechende Lösungen einsetzen.

Bei Open Source hat der Anwender immer die Auswahl, die Entwicklung, den Betrieb und die Wartung einem anderen Hersteller anzuvertrauen, oder dies – zumindest im Prinzip – auch selber zu übernehmen.

Zur Auswahl der passenden Open-Source-Lösung ist noch eine weitere Betrachtung sinnvoll: In der Hadoop-Open-Source-Gemeinschaft arbeiten mehrere hundert Entwickler an verschiedenen Projekten, die sich modular in eine Gesamt-Lösung einfügen. Bei diesen Entwicklern muss man zwischen

- Reviewern,
- Contributoren und
- Committern

unterscheiden.

Committer sind die erfahrensten Mitglieder der Gemeinschaft, die sich um die Koordination kümmern und die Richtung und Roadmap formulieren, so dass am Ende die verschiedenen Projekte wie Bausteine zusammenpassen. Sie sind außerdem die letzte Instanz zur Qualitätssicherung neuer Software.

Wer sehen will, wie stark sich ein Hersteller in die Open-Source-Gemeinde um Hadoop einbringt, sollte sich die Zahl der Committer in der jeweiligen Belegschaft anschauen.

Die Analogie vom Markt für Server-Betriebssysteme, aber auch für relationale Datenbanken, zeigt, dass sicher auch für Hadoop-Distributionen nicht beliebig viele Vendoren gleichermaßen am Markt partizipieren werden, sondern dass sich vermutlich zwei bis drei dominante Anbieter herauskristallisieren werden. Die besten Chancen haben die Hadoop-Distributionen all jener Firmen, die ihr Engagement für dieses Big-Data-Betriebssystem auch dadurch ausdrücken, dass eine signifikante Zahl ihrer Angestellten als Committer aktiv in der Gemeinde mitarbeitet.

Als Fazit ergibt sich: Unternehmen sollten sich gut überlegen, wo und wie sie Open-Source-Technologie in ihrer Big-Data-Strategie verwenden wollen; ignorieren sollte man Open Source auf keinen Fall.

7 Risiken bei Big-Data-Anwendungen

Die Risiken, die Big-Data-Projekte mit sich bringen, sind nicht zu vernachlässigen. Mitunter sind es neue Risiken, weshalb es wichtig ist, sich die Risiken und Gefahren in der Theorie bewusst zu machen. Die ausführliche Darstellung der Risiken soll nicht von der Big-Data-Analyse abschrecken, sondern über die neuen Risiken aufklären.

Wie in allen Bereichen des täglichen Lebens, bestehen im Zusammenhang von Big-Data-Anwendungen gewisse Risiken und Gefahren. Oft sind dies Gefahren, die selbstverständlich geworden sind, beispielsweise die Möglichkeit eines Datenverlustes oder gar eines Datendiebstahls. Das Verständnis der Gefahr selbst macht diese zwar nicht minder gefährlich, doch sorgt sie dafür, dass sich Anwender und Anbieter darauf vorbereiten können.
Die Risiken, die Big-Data-Projekte mit sich bringen, sind nicht zu vernachlässigen. Mitunter sind es neue Risiken, weshalb es wichtig ist, sich die Risiken und Gefahren in der Theorie bewusst zu machen. Dieses Kapitel des Big-Data-Leitfadens soll deshalb nicht vom Umgang mit Big Data abschrecken, sondern über die neuen Risiken aufklären.
Zudem gibt es eine Anzahl von Herausforderungen auf der technischen und administrativen Seite[203], derer sich die Anwender und Anbieter bewusst sein müssen. Zu den Gefahren, die sich trotz der Perspektiven, die sich durch den Einsatz und die Verwendung von Big-Data-Technologien und -Techniken eröffnen, gehören vor allem die Schädigung der Reputation bis hin zu deren Verlust. Es kann zu einem Vertrauensbruch zwischen Geschäftspartnern sowie zwischen Unternehmen und Kunde kommen. Imageschäden sind als Folgen vorhergehenden falschen Handelns zu sehen. Sie sind mitunter das Resultat menschgemachter Fehler. Da Fehler oft aus unzureichendem Wissen entstehen, ist es umso wichtiger, die Risiken zu kennen. Bereits bekannte Risikofelder, wie die allgemeine IT-Sicherheit werden sich, im Zusammenspiel mit der weiter zunehmenden Technologie-Abhängigkeit sowie der zunehmenden Komplexität von Systemen, weiter ausweiten.
Die mit den verschiedenen Bereichen der Entwicklung und Anwendung von Big-Data-Technologien verknüpften Gefahren stellen regelrechte Stolperfallen bei der Umsetzung von Big-Data-Szenarien dar. Die Abbildung 59 veranschaulicht die verschiedenen Risiko-Bereiche.

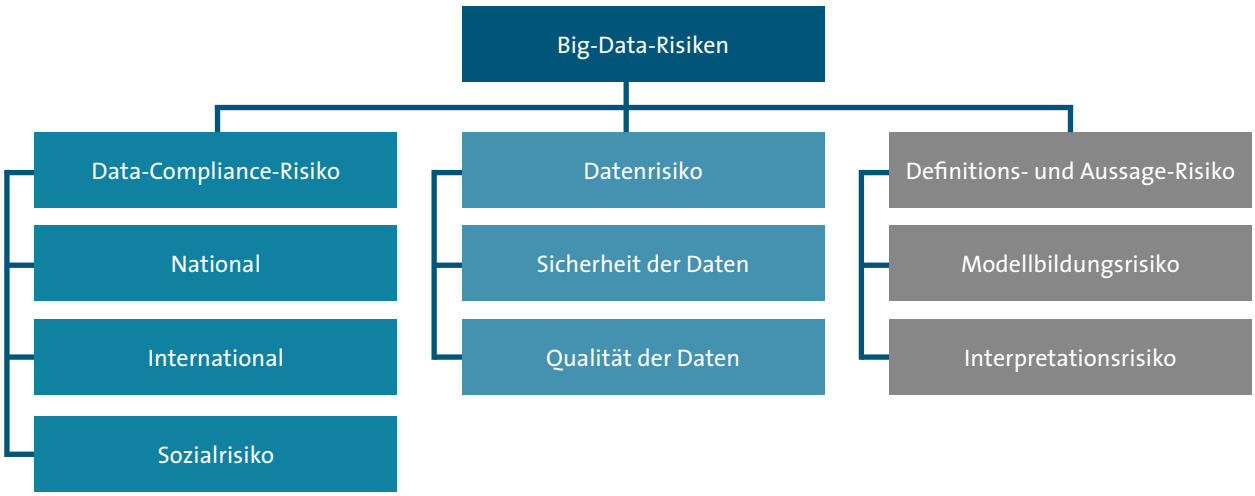


Abbildung 59: Risikobereiche bei Big Data

[203] Datenschutz, Bandbreite in der Datenübertragung, juristische Themen etc.

Die effiziente Umsetzung eines Big-Data-Projektes erfordert Kompetenzen in den Bereichen des Data Managements, des Data Warehousings, der Datenbanken, der IT-Infrastruktur, der Skriptsprachen, des Enterprise-Content- und Document-Managements sowie der Business Intelligence. Außerdem werden Fachkompetenzen im Bereich des Datenschutzes benötigt. Je besser es gelingt, diese Kompetenzen in Form eines Projektteams oder einer speziellen Data-Science-Fachabteilung abzudecken, desto geringer fallen die Risikowahrscheinlichkeiten der aufgezeigten Risiken aus. Für gewöhnlich existieren die notwendigen Kompetenzen verteilt innerhalb eines Unternehmens und müssen für Big Data aggregiert werden.

7.1 Data-Compliance-Risiken

Big-Data-Technologien bergen insbesondere ein Risiko hinsichtlich der Einhaltung von gesetzlichen Vorgaben zum Datenschutz. Die Einhaltung der gesetzlichen, unternehmensinternen und vertraglichen Regelungen stellt für die Compliance eine beträchtliche Herausforderung dar. Unter dem Begriff der Data Governance versteht man das Compliance-Risiko, welches sich aus dem Erheben, Sammeln, Speichern, Löschen und Verändern sowie der Weitergabe, dem Zugriff auf und der Auswertung von Daten ergibt.

Viele Daten und deren Server fallen nicht nur in den Interessenbereich nationaler Institutionen, sondern stehen sowohl mit europäischen, als auch mit internationalen Einrichtungen in wechselseitiger Abhängigkeit. Der geltende Rechtsrahmen muss zu jeder Zeit klar sein.

Unternehmen, die nicht nur nationalen, sondern auch internationalen Raum tätig sind, müssen diesen Aspekt berücksichtigen. Es dürfen weder Annahmen bezüglich des rechtlichen Rahmens noch Rückschlüsse von einer nationalen Regelung auf eine andere getroffen werden. Die Unwissenheit über die jeweiligen nationalen Regelwerke schützt nicht vor den Folgen eines Verstoßes. Daher ist es wichtig, genauestens informiert zu sein.

Nationale Rahmenbedingungen

Datenschutzgesetze dienen dazu, personenbezogene Daten zu schützen. Innerhalb Deutschland ist dabei das Ziel, jedem Menschen das Recht auf informationelle Selbstbestimmung zuzusichern, welches bereits im Volkszählungsurteil jedem Bürger zugesichert wurde. Innerhalb des Bundesdatenschutzgesetzes (BDSG) ist aus diesem Grund ein grundsätzliches Verbot, der Erhebung, Verarbeitung und Nutzung personenbezogener Daten definiert (§ 4). Laut Gesetz ist die Verwendung dieser Daten nur dann erlaubt, wenn ein spezieller gesetzlicher Erlaubnistatbestand existiert oder eine explizite Einwilligung der betroffenen Person vorliegt.

Neben dem Prinzip des Verbots mit Erlaubnisvorbehalt sind innerhalb des BDSG weitere Grundsätze für den Umgang mit personenbezogenen Daten definiert. Zum einen ist das der Grundsatz der Zweckbindung. Dieser besagt, dass Daten generell nicht ohne einen konkreten Zweck erhoben und verarbeitet werden dürfen und schließt somit eine wahllose Ansammlung von Daten weitestgehend aus. Ergänzend besagt der Grundsatz der Datenvermeidung und Datensparsamkeit, dass bei der Datenerhebung anlässlich eines konkreten Zwecks geprüft werden muss, welche Daten zu dessen Erfüllung tatsächlich benötigt werden, um sicher zu stellen, dass keine überflüssigen Daten erhoben werden. Daten, welche zu einem bestimmten Zweck erhoben wurden, dürfen grundsätzlich nicht ohne das Einholen weiterer Einwilligungen für andere Zwecke weiterverwendet werden. Gleiches gilt für Daten, welche zur Erfüllung einer Vertragsbeziehung erhoben wurden. Letztlich findet sich im BDSG der Grundsatz der Transparenz. Dieser besagt, dass jeder Betroffene über die Rahmenbedingungen der Datenerhebung informiert werden sollte. Der Betroffene sollte also über den Zweck, den Ort und die Dauer der Speicherung sowie alle beteiligten Parteien informiert werden.

Die unachtsame Ansammlung von Daten über den Gebrauchszeitraum hinaus birgt die Gefahr, für regelwidrige Speicherung und Handhabung von Daten belangt zu werden. Es ist daher zwingend notwendig, genau zu überprüfen, welche Arten von Daten erhoben und verarbeitet werden sollen und welche bereits vorliegen.

Sonderregelungen

Darüber hinaus sieht das BDSG in einigen Fällen Sonderregelungen vor. Für die Datenerhebung zur Erfüllung der eigenen Geschäftszwecke finden sich innerhalb des Paragraph 28 einige Ausnahme-Regelungen, welche die Erhebung in speziellen Fällen erlauben. Außerdem erlaubt ist die Nutzung von Daten aus öffentlich zugänglichen Quellen (§29). Im Sinne dieser Regelung gelten Daten als öffentlich zugänglich, wenn sie für einen beliebigen Personenkreis zugänglich sind. Allerdings muss vor der Erhebung eine Abwägung der schutzwürdigen Interessen des Betroffenen gegen die Interessen der verantwortlichen Stelle vorgenommen werden. Darüber hinaus definiert das BDSG für die Datenerhebung zu Werbezwecken spezielle Anforderungen und sieht nur wenige Ausnahmen vor, welche die Erhebung ohne explizite Einwilligung erlauben.

Verwendung personenbezogener Daten

Eine weitere Ausnahmeregelung stellt die Verwendung personenbezogener Daten in anonymisierter oder pseudonymisierter Form dar. Dabei gilt es, alle Informationen aus den Daten zu entfernen, die einen Personenbezug herstellen. Auch durch die Kombination von Daten aus unterschiedlichen Quellen kann ein Personenbezug bestehen kann.

Werden Daten zur Anonymisierung aggregiert, muss daher sichergestellt sein, dass eine ausreichende Menge an Daten gelöscht wird, so dass keine Rückschlüsse auf Einzelsachverhalte mehr möglich sind. Wird eine Pseudonymisierung durchgeführt, gilt es, alle personenbezogenen Daten durch Pseudonyme zu ersetzen. Nach dem Willen des Gesetzgebers ist die Verwendung der Daten in dieser Form zulässig, wenn auch durch die Anreicherung der Daten mit weiteren Informationen kein Personenbezug hergestellt werden kann. Von besonderer Bedeutung ist dieser Punkt gerade dann, wenn die aufbereiteten Daten an Dritte weitergegeben werden sollen.

Das ebenfalls im Zusammenhang mit Big-Data-Projekten zu beachtende Telemediengesetz (TMG) definiert für die Verwendung von bestimmten Online-Tracking-Daten die Anforderung, dass der Betroffene vorab informiert werden und ihm die Möglichkeit eingeräumt werden muss, die Erhebung abzulehnen. Laut TMG muss dies unabhängig davon geschehen, ob die Daten mit oder ohne Namen verwendet werden. Zusätzlich regelt das Telekommunikationsgesetz (TKG), die Anforderungen für die Erhebung und Verwendung von Standortdaten aus GPS, GSM und WLAN-Netzen. Laut TKG dürfen diese Daten grundsätzlich nur anonymisiert oder auf Basis einer Einwilligung des Betroffenen verwendet werden.

Urheberrechtsgesetz im Kontext von Big Data

Letztlich spielt auch das Urheberrechtsgesetz im Kontext von Big Data eine Rolle. Innerhalb des UrhG sind Regelungen und Gesetze zu Datenbanken enthalten. Es wird eine Unterteilung in die eigentliche Datenbank und das Datenbankwerk vorgenommen. Im Sinne des UrhG ist das Datenbankwerk ein »Sammelwerk, dessen Elemente systematisch oder methodisch angeordnet und einzeln mit Hilfe elektronischer Mittel oder auf andere Weise zugänglich sind« (Paragraf 4 Abs. 2 UrhG) und fällt somit zumeist unter den Urheberrechtsschutz.

Internationale Rahmenbedingungen

Erfolgt die Datenerhebung oder Übermittlung über nationale Grenzen hinaus, finden entweder die nationalen Regelungen des jeweiligen Landes oder inter- bzw. supranationale Gesetze ihre Anwendung. Grundsätzlich ist zu beachten, dass in jedem Fall vorhergehend geprüft werden muss, ob die Datenverwendung in Deutschland zulässig ist und erst im Anschluss daran, die Übermittlung in das jeweilige Empfängerland.

Innerhalb der Europäischen Union ist der Datenschutz als Grundrecht in der Charta verankert. Besondere Bedeutung kommt in der Anwendung vor allem den folgenden beiden Richtlinien zu:

1. Die Richtlinie 95/46/EG zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr
2. Die Richtlinie 2002/58/EG über die Verarbeitung personenbezogener Daten und den Schutz der Privatsphäre in der elektronischen Kommunikation.

Der Datenaustausch zwischen Mitgliedstaaten der EU kann aufgrund der Richtlinien zumeist reibungslos ablaufen. Die von der EU vorgegeben Richtlinien im Gegensatz zu Verordnungen stellen dabei Aufforderungen an die Länder dar, nationale Gesetze aufzusetzen oder anzupassen.

Der Datenschutz aller Nicht-EU-Staaten wird an dieser Vorgabe gemessen. Liegt keine nationale Regelung für den Datenschutz vor oder wird das vorliegende Datenschutzgesetz als nicht ausreichend eingestuft, muss vorhergehend zu jedem Datenaustausch der Datenschutzbedarf der betroffenen Parteien geprüft werden. Als unsichere Drittstaaten, in die keine personenbezogenen Daten aus Deutschland übertragen werden dürfen, gehören unter anderem Japan, Indien, China und die USA.

Sozialrisiko

Ferner wirft die Anwendung von Big-Data-Technologien neue wert- und moralbezogene sowie ethische Fragen auf, welche unter dem Begriff Sozialrisiko zusammengefasst werden können. Die Auswertung von Unmengen an Daten aus bisweilen oftmals kritischen Quellen, wie beispielsweise sozialen Netzwerken, vermittelt innerhalb der Bevölkerung den Eindruck, von Staaten und Unternehmen umfassend überwacht zu werden. Spätestens wenn Firmen personalisierte Werbung verschicken, welche klar macht, dass die Vorlieben und Gewohnheiten der jeweiligen Empfänger bekannt sind, wenn also der persönliche Bezug offenkundig wird, steigt das Misstrauen an, da das Gefühl bestärkt wird, unter permanenter Beobachtung zu stehen. Wenn Kunden sich Fragen stellen müssen wie: »Woher wissen die denn das?«, oder: »Was wissen die noch?« kann auch die Kundenbindung Schaden nehmen. Durch derartige Aspekte, verursacht das Thema Big Data selbstverständlich auch kontrovers geführte, öffentliche Diskussionen, welche die Notwendigkeit einer kritischen Auseinandersetzung aller Seiten mit der Materie verdeutlichen. Die Streitparteien vertreten oftmals antagonistische Interessen, doch es sollte immerzu versucht werden, einen Ausgleich und Konsens zwischen den Informationsinteressen der Unternehmen und dem Recht auf informationelle Selbstbestimmung der Betroffenen zu finden. Es ist dabei notwendig, den Nutzen, der für den Betroffenen durch die Freigabe und Verwendung der Daten entstehen kann, dagegen aufzuwiegen, wie viel er von sich preisgeben muss.

7.2 Datenrisiken

Neben den Risiken, welche aus regelwidrigem Verhalten im Sinne des juristischen Rahmens resultieren, bestehen auch auf technischer Ebene bestimmte Risiken, die zu beachten sind. Auch diese Risiken ergeben sich aus der Erhebung und Weiterverarbeitung von Daten. Zum einen besteht dabei ein Risiko, in Hinblick auf die Gewährleistung der Daten-Sicherheit. Zum anderen bergen die Daten selbst durch ihre Qualität bestimmte Risiken.

TCP/IP als Basis des Internets ist jedoch (potentieller Verlust von Datenpaketen, fehlende Echtzeit-Fähigkeit) nicht geeignet, sehr große, sensible Datenmengen zu übertragen. Komplexe Big-Data-Szenarien benötigen gegebenenfalls zusätzliche Technologien für eine schnelle zuverlässige Übertragung von Daten.

Die unternehmensweit gültigen Verfahren für Datenschutz, Anonymisierung und Pseudonymisierung müssen auf Datenvolumina im Multiterabytes-Bereich und Petabytes-Bereich angepasst werden. Die derzeit genutzten Technologien sind oftmals nicht für enorm große Datenmengen geeignet, wie sie in komplexen Big-Data-Szenarien anfallen werden.

Sicherheit der Daten

In einer Zeit, in der Unternehmen große Datenmengen automatisch ansammeln und zwischen vernetzten Speichern weltweit transportieren, sind effiziente Strategien zur Zugriffsbeschränkung unbedingt notwendig, um die große, potentielle Angriffsfläche abzusichern. Angreifer werden versuchen die Daten zu löschen und zu manipulieren sowie sie zu kopieren, um sie an Dritte weiter zu reichen oder für anderweitige Zwecke zu verwenden. Die Daten müssen also gesichert werden. Gleichzeitig muss der Zugriff durch Berechtigte gewährleistet bleiben. Bereits beim Ansammeln der Daten werden Unternehmen hierbei mit folgenden Fragen konfrontiert:

- Wozu sollen die Daten erhoben werden?
- Welche Art von Daten sollen erhoben werden?
- Wie sollen Daten erfasst werden?
- Wie können sie sicher gespeichert werden?
- Wie können sie und vor unberechtigten Zugriffen Dritter geschützt werden?
- Welche ergänzenden Datenfelder sind dazu nötig?
- Welche Schutzmechanismen und -strategien müssen angewandt werden?

Jederzeit ist sicherzustellen, dass die Daten vor Störungen und Eingriffen von außen geschützt sind. Eine tragende Rolle spielt hierbei eine effektive Verschlüsselung der Daten.
Ein weiterer essenzieller Faktor ist der Schutz der Daten vor unbefugten Zugriffen. Um dies gewährleisten zu können, sind effiziente Zugriffsbeschränkungen notwendig.

Das Resultat ineffizienter Zugriffbeschränkungen und Sicherheitsregularien können beispielsweise Geheimnisverrat, Erpressung oder der Verlust wichtiger Geschäftsdaten sein.

Qualität der Daten

Abgesehen vom Risiko der Schutzwürdigkeit von Daten, bergen diese zudem selbst ein gewisses Risiko. Innerhalb der Anwendung entscheidet die Qualität der Ausgangsdaten darüber, inwiefern nachfolgende Analysen für das Unternehmen wertvolle Ergebnisse erzeugen können. Wird die Qualität der Daten in frühen Stadien der Modellfindung falsch eingeschätzt, steigt das Risiko von Fehlinterpretationen.

Datenqualität im herkömmlichen Sinne beschreibt oft nur die Prüfung der Vollständigkeit von Datensätzen (beispielsweise »Verfügen alle Adressen über eine gültige Postleitzahl?« oder »Stimmen Postleitzahl und Postanschrift überein?« oder die reine Vermeidung von Duplikaten. Da für Big-Data-Projekte auch Daten aus externen Datenquellen und Social-Media hinzugezogen werden, müssen die Lösungen einen technischen Ansatz zur Implementierung von Expertenwissen bzw. Domänenwissen liefern.
Darüber hinaus müssen die Big-Data-Lösungen die Plausibilität von Daten prüfen. Beispielgebend sei hier die Analyse einer bestimmten Wahrnehmung oder Stimmung zu

einem Produkt oder Sachverhalt in Social-Media genannt: Stimmen die Daten oder wurden bewusst falsche Meinungen publiziert. Handelt es sich um ein nachhaltiges Stimmungsbild oder um eine Momentaufnahme? Stammen die Aussagen zu einer Stimmung von einer relevanten Benutzergruppe?

Das gilt auch für die Prüfung von Daten aus externen Quellen hinsichtlich einer möglichen Manipulation. Erforderlich sind technische und methodische Lösungen für die Verschlüsselung sowohl der lagernden als auch der in Bewegung befindlichen Daten.

Ein wichtiger Aspekt der der Datenqualität ist die Analyse-Ergebnisse können nur dann einen Mehrwert erzeugen, wenn die Analysen auf Basis korrekter Daten angewendet wurden. Die Datenintegrität.kann von verschiedenen Faktoren abhängig sein, unter anderem von der Aktualität der Daten sowie der Beschaffenheit der Quelle und der Übertragungsart. Damit verknüpft spielt auch die Authentizität der Daten eine Rolle.

Ein weiterer Aspekt ist die Konsistenz der erhobenen Daten. Verluste bei der Übertragung, der Speicherung, versehentliches oder absichtliches Löschen, können dazu führen, dass die Daten in ihrer Gesamtheit nicht mehr vollständig sind.

Darüber hinaus spielt die Verfügbarkeit der Daten eine entscheidende Rolle. Innerhalb des Systems muss daher sichergestellt sein, dass die Daten im System zu jeder Zeit, mit relativ geringem Aufwand und in korrekter Form, von den befugten Personen abgerufen und verwendet werden können.

7.3 Definitions- und Aussagerisiko

Auf der Grundlage einer zielorientierten Fragestellung, der statistisch-mathematischen Modelle und deren struktureller Vorgaben, werden aus der Menge der gesammelten Daten die zur welche zur Untersuchung geeignen ausgewählt. Auf diesem Weg – von der konkreten Fragestellung über die Auswahl des Modells, bis hin zur Implementierung der Big-Data-Anwendung – können die in Abbildung 60 gezeigten Gefahren auftreten.
Die einzelnen Prozessschritte bauen dabei aufeinander auf, und sie dürfen nicht separat betrachtet werden. Innerhalb des Prozesses sind zwei Risiken zu separieren: das Modellbildungsrisiko und das Interpretationsrisiko.

Modellbildungsrisiko

Ein übergreifendes Risiko besteht innerhalb der Modellbildung durch die unbedingt notwendige Einhaltung von Datenschutz-Richtlinien und Gesetzen. Beispielsweise sollte schon innerhalb der Formulierung der Fragestellung eine Vorstellung dafür vorliegen, welche Daten rechtskonform erhoben und verarbeitet werden dürfen, so dass das erhoffte Ergebnis letztendlich auch zu erhalten ist. Durch das Aufsetzen einer Fragestellung, welche keinen Personenbezug erwartet, kann auch dem Risiko einer unzureichenden Anonymisierung oder Pseudonymisierung der Daten entgangen werden. Die Fragestellung sollte möglichst konkret gestellt werden, um einen genauen Erwartungshorizont für das Ergebnis aufzuzeigen, und moralisch-ethische Aspekte ebenso wie unternehmerische Aspekte berücksichtigen.
Der Teilprozess der Modell-Auswahl umfasst die Auswahl geeigneter statistisch-mathematischer Modelle sowie die Auswahl der Analysetechniken und -verfahren. Für keine Big-Data-Fragestellung existiert eine Pauschallösung. Standards, welche sich mit der Zeit für klassische BI-Lösungen entwickelt und bewährt haben, können im Kontext von Big Data mitunter keine effiziente Anwendung finden. Innerhalb der Entwicklung können erste Probleme auftreten, die durch fehlerhafte Definitionen innerhalb der vorherigen Schritte induziert wurden. Wird die Entwicklung als eigenständiger Prozess gesehen, welcher nicht durch bereits getroffene Entscheidungen

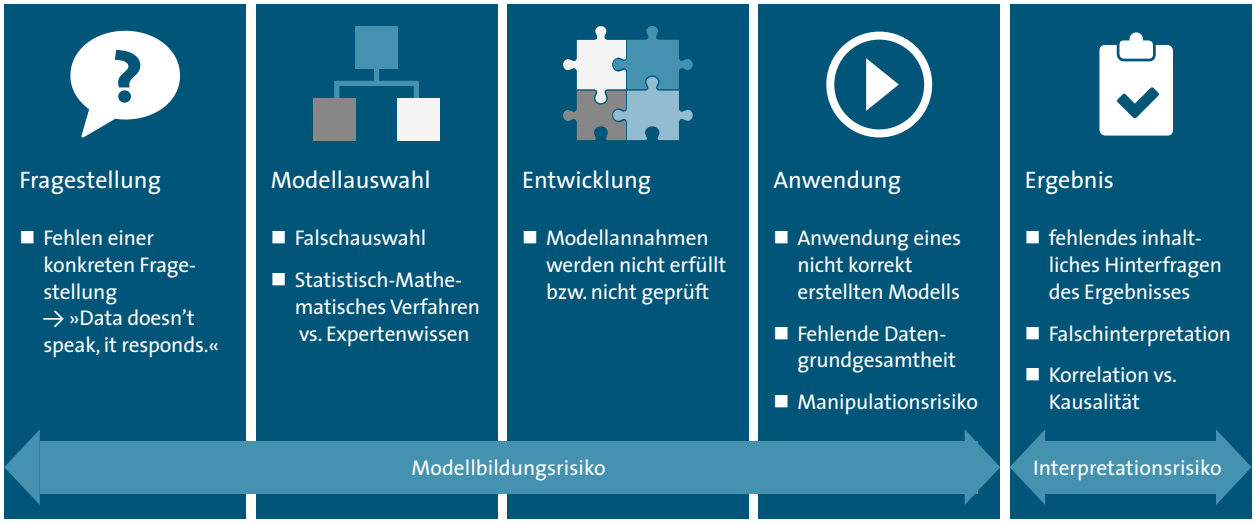


Abbildung 60: Vom Modell zur Aussage: mögliche Risiken auf diesem Weg

beeinflusst wird, kann es zudem dazu kommen, dass Modellannahmen keine Einbeziehung finden und in der späteren Anwendung nicht erfüllt werden. Inkonsistenzen und Qualitätsmängel innerhalb der zugrunde liegenden Daten, Fehler innerhalb des Modells, sowie Sicherheitslücken, welche Manipulationsrisiken erzeugen, zeigen sich spätestens innerhalb der Anwendung.
Vermehrt werden Analysen direkt in die Prozessablaufe integriert. Vor allem in diesem Zusammenhang, spielen die Risiken eine große Rolle. Die Ergebnisse derartiger Analysen werden zumeist, direkt im weiteren Prozessverlauf angewandt, ohne eine zwischengeschaltete Überprüfung oder Interpretation durch den Menschen. Stellen diese Analysen den zugrunde liegenden Sachverhalt nicht korrekt dar, kommt es unweigerlich zu Fehlern im Prozessablauf.

Interpretationsrisiko

Wurde die Wahl eines Projektes erfolgreich innerhalb der Anwendung umgesetzt oder steht die Entwicklung des Modells kurz vor einem erfolgreichen Abschluss, ist es kein weiter Weg mehr bis hin zur Ergebnisinterpretation sowie den zu erzielenden Schlussfolgerungen. Doch selbst wenn alle bisherigen Annahmen korrekt waren, das Modell zutreffend ausgewählt, entwickelt und angewendet wurde, können sich innerhalb der Auswertung und Auslegung der Ergebnisse neue Probleme und Fehler offenbaren.
Viele Modelle schaffen es zwar, komplexe Zusammenhänge fachgemäß und fehlerlos zu ermitteln, verfügen aber nicht über eine leicht verständliche Darstellung. Derartige Modelle sind in hohem Maße anfällig für Fehlinterpretationen durch Laien. Insbesondere kausale Zusammenhänge werden gerne gefunden, obwohl die zugrundeliegenden Analysen derartige Rückschlüsse möglicherweise überhaupt nicht zulassen. In der Folge entstehen fehlerhafte Aussagen. Dass solche Fehler auch über längere Zeit nicht in Vergessenheit geraten, wird beispielsweise an Hand der im Buch Freakonomics[204] dargestellten Fälle ersichtlich.
Big Data kann (wie die meisten statistischen Modelle) zunächst nur für denjenigen einen Nutzen und ein verwertbares Ergebnis erzeugen, der die Regeln der Interpretation kennt. Bei regelmäßigen Analysen müssen Definitionen für die Verwertung und Interpretation des Ergebnisses erstellt werden. Existierende Definitionen, müssen überprüft werden.
Unabhängig von unbewussten Fehlern innerhalb der Analyse und Interpretation können Ergebnisse und Aussagen bewusst verfälscht werden. Derartiger Missbrauch ist besonders kritisch, da er sowohl schwer zu erkennen ist.

[204] ISBN-10: 3442154510

7.4 Faktoren der Risikovermeidung

Um den benannten Risiken begegnen zu können, ist es in erster Linie wichtig, den potentiellen Fehlerquellen kontinuierlich Beachtung zu schenken. Viele Fehler entstehen durch unachtsames oder unüberlegtes Handeln. Bisherige Standards und bewährte Systeme, Modelle oder Entscheidungen, finden innerhalb des Big-Data-Kontextes oft keine direkte Umsetzung bzw. Anwendung mehr. Werden alle möglichen Risikofaktoren (vor allem auch das Bauchgefühl bei der Fragestellung und bei der Interpretation) angemessen beachtet und Entscheidungen und deren Auswirkungen gegeneinander abgewogen, kann ein sicherer Umgang mit Big Data gewährleistet werden.

Die Abbildung 61 zeigt einen Überblick über die Faktoren, die innerhalb der Risikovermeidung eine Rolle spielen.

Faktor: Daten

- Aus den Umständen und dem Zweck der Datenerhebung, sowie aus Verträgen oder Gesetzen ergeben sich oftmals spezielle Löschfristen für die betreffenden Daten. Es bestehen also Vorgaben über deren Aufbewahrungszeitraum. Die Einhaltung dieses Mindesthaltbarkeitsdatums sollte unbedingt gewährleistet werden.

- Die Qualität der Daten sollte gründlich, hinsichtlich ihrer Konsistenz, Aktualität und Korrektheit geprüft werden

- Durch das Einbringen fiktiver Daten in Form von Hashwerten über einzelne Datensätze oder Datenbanken, kann die eindeutige Identifizierung eines Datensatzes und dessen Integrität überprüft und somit gewährleistet werden[205]. Es lässt sich so das Herkunftssystem der Daten ermitteln sowie prüfen, ob ein Datensatz ein Original oder eine Kopie ist.

- Um die unrechtmäßige Datenweitergabe oder Datendiebstahl aufdecken zu können, ist die Anreicherung der im System erhobenen Datensätze mit gefälschten bzw. unechten Datensätzen empfehlenswert.

Abbildung 61: Faktoren der Risikovermeidung

[205] Stichwort: Daten-DNA

Faktor: Data Management

- Die Datenschutzgesetze sollten regelmäßig überprüft und entsprechende Risikovermeidungsmaßnahmen festgelegt werden. Zum Beispiel durch Hinweisgebersysteme, Anonymisierung und Aggregation, Pseudonymisierung sowie durch Schulungen der Mitarbeiter.

- Eine Datenlandkarte mit Metainformationen der Daten im System kann ersichtlich machen, ob Daten von Änderungen an äußeren Bedingungen betroffen sind oder ob kein Handlungsbedarf besteht.

- Daten sollten im System nur endgültig gelöscht werden können. Löschungen die Daten nur scheinbar aus der Datenbank entfernen und sie im Hintergrund archivieren bergen ein hohes Risiko. Löschprotokolle verbessern die Nachweisbarkeit von Löschungen.

- Um eine Big-Data-konforme Verwaltung der Daten umsetzen zu können, muss zu jeder Zeit und mit geringem Aufwand auf die eigenen Daten zugegriffen werden können.

Faktor: Organisation

- Angemessene Ressourcen, in Form von Budget, Expertise, Arbeitskraft und Zeit gewährleisten Compliance.

- Eine eigene Data-Science-Fachabteilung verbessert die Compliance und reduziert Risiken. Innerhalb dieses Daten-Gremiums sollten Personen aus dem Betriebsrat und der IT, der Datenschutzbeauftragte und der Chief Data Officer vertreten sein.

- Es sollten klare Formulierungen bezüglich der geregelten Verantwortlichkeiten und Aufgaben sowie der bestehenden Schnittstellen, Zugriffsrechten und Nutzungsregeln existieren.

- Die Nutzer sollten über die Funktionsweise der im Modell verwendeten Analyse-Algorithmen aufgeklärt werden. Ein allgemeines Verständnis der Anwendung beugt Fehlinterpretationen vor.

- Erkenntnisse der internen Überwachung sollten regelmäßig kommuniziert werden. Besondere Aufmerksamkeit sollte Hinweisen auf Verstöße zukommen. Die bestehenden Sanktionen sollten allgemein bekannt sein.

Faktor: Prozess

- Durch eine externe Überwachung des Prozesses der Datenwertschöpfung können Datenmissbrauch und Ergebnismanipulation vermieden werden.

- Beständige Begleitung des Projektprozesses durch den Betriebsrat sowie durch einen Datenschutzbeauftragten um kontinuierlich die Rechtskonformität zu überprüfen.

- Strenge Kontrollen und Prüfungen des Ablaufs können Schwachstellen innerhalb der internen Prozesse aufdecken. Aus den Ergebnissen der Prüfung können überarbeitete Prozessplanungen aufgesetzt werden.

Faktor: Kunden als Betroffene

- Der Mehrwert der für einen Anwender entsteht, welcher seine Einwilligung zur Nutzung seiner persönlichen Daten erteilt, sollte deutlich artikuliert werden. Der Kunde sollte genauestens über Art und Umfang der Verwendung informiert werden. Durch die Darstellung des persönlichen Nutzens und die genaue Auskunft, kann das Vertrauen des Kunden durch die Transparenz bestärkt werden.

- Ohne eine explizite Einwilligung, d.h. ohne Legitimation sollte keine Datenspeicherung und -verarbeitung stattfinden.

- Personenbezogene Daten sollten immer nur zweckbezogen angefordert und gespeichert werden. Laut dem BDSG gelten zum einen die Zweckbindung und zum anderen die Datensparsamkeit, weshalb niemals mehr Daten als unbedingt notwendig gespeichert werden sollten.

7.5 Methodische Herausforderungen

Big Data verkörpert vor allem eine Kombination vieler verschiedener, technologischer Architekturen. Online-Transaktionsverarbeitung und Online Analytical Processing verschmelzen. Data Warehouses und BI-Lösungen erhalten mit innovativen, dem Bereich Big Data zuzurechnenden Technologien[206] neue Aspekte. Die Herausforderung besteht in der Kombination der bisherigen mit den neuen Technologien und der Überwindung der traditionellen Trennung von transaktionaler und analytischer Online-Verarbeitung. Das erfordert neben den Technik-Investitionen vor allem organisatorische Maßnahmen und einen Kulturwandel in den Unternehmen – bis hin zur Neudefinition von Rollen und Verantwortlichkeiten. Anwender, die bereits mit Data Warehousing und Business Intelligence vertraut sind, werden auf dem Weg zur Verarbeitung sehr großer Datenmengen zunächst einzelne Analyseaufgaben durch neue Technologien ersetzen oder ergänzen. Auch die Kombination von Stapelverarbeitung und Online-Verarbeitung stellt eine Herausforderung dar, die in einem Big-Data-Projekt mit methodischen und technischen Maßnahmen adressiert werden muss.

7.6 Technische Herausforderungen

Mit den neuen Technologien ist die Skalierbarkeit der vorhandenen Systeme zu prüfen. Werden große Datenmengen aggregiert und zunächst im eigenen Rechenzentrum gespeichert und verarbeitet, erschöpft sich die Skalierbarkeit von Standardsystemen[207] im Terabytes- und im Petabytes-Bereich. Oft müssen dann weitere Systeme – mithin weitere Rechenzentren – geplant, installiert und in die Unternehmensprozesse integriert werden. Hier werden Erfahrungen und Lösungen benötigt, die bislang im Supercomputing typisch sind.

Dabei ist wiederum offen, wie weit die bereits genutzten und auch die neu hinzugekommen Software-Lösungen dieses Wachstum mittragen, denn oft ist die Parallelisierung innerhalb einzelner Systeme schon eine enorme Herausforderung für die Systembetreuer in den Anwenderunternehmen.

Echtes Disaster-Recovery (Wiederherstellung von Systemen und Daten im Katastrophenfall) kann im Petabytes-Bereich derzeit nicht mit akzeptablem finanziellen Aufwand realisiert werden. Die Anwender müssen also sehen, dass sie mit den vorhandenen Technologien eine größtmögliche Hochverfügbarkeit, auch gegen logische Fehler, erreichen. Die hardwareseitigen Hochverfügbarkeitsmechanismen (RAID, Snapshot-Technologien) müssen durch Datenspiegelungsverfahren ergänzt werden. Auch die Deduplizierung von Daten spielt eine Rolle. Kontraproduktiv im Sinne der Erschließung von derzeit noch in den großen Datenmengen verborgenen Informationen wäre es, auf eine Reduzierung der zu administrierenden und zu sichernden Daten zu setzen.

Die Erfahrungen aus dem Supercomputing zeigen auch, dass die Performance des gesamten Systems und auch die Performance zwischen den Servern, innerhalb des SANs und im Weitverkehrsnetz genauer betrachtet werden muss. Es wird künftig eine Herausforderung sein, den Datendurchsatz auch im Petabytes-Bereich nicht nur auf

[206] In-Memory Computing, Complex Event Processing, neue Datenbank-Architekturen
[207] Storage und Server

der verarbeitenden Seite, sondern auch bei der Ein- und Ausgabe zu erreichen. Dementsprechend muss auch die Leistungsfähigkeit des Netzwerks (LAN und WAN). Dieser Problematik werden sich auch die Telekommunikationsanbieter stellen müssen.

Die Netzwerkauslastung wird auch bei der Verteilung der Berechnungsergebnisse eine Rolle spielen, da viele Big-Data-Szenarien den Nutzen darin ziehen, dass die Berechnungsergebnisse (im Gegensatz bzw. in Erweiterung zu klassischen Business-Intelligence-Lösungen an größere Benutzergruppen, z. B. ganze Vertriebsmannschaften verteilt werden).

8 Technologien zur Umsetzung rechtlicher Anforderungen

Das Kapitel 8 beschäftigt sich mit Fragestellungen, bei deren Bewältigung Technologieexperten, Rechts- und Organisationswissenschaftler eng zusammenarbeiten müssen.
Zuerst wird der Frage nachgegangen, wie Garantien über den Datenschutz in die Datenanalyse integriert werden können. Ein Patentrezept kann nicht angeboten werden; es werden jedoch Ansätze diskutiert, die sich als sinnvoll herauskristallisiert haben (Abschnitt 8.1).
Eine wichtige Frage im Zusammenhang mit der Verwertung persönlicher digitaler Daten ist noch Gegenstand der Forschung: Wie kann ein Modell zur Verwertung persönlicher digitaler Daten aussehen, das Dateninhaber, Datenverwerter sowie Dienstleister der Datensicherung, Datenaufbereitung sowie des Datenschutzes zusammenführt? Als eine mögliche Antwort auf die Herausforderungen im Umgang mit persönlichen digitalen Daten wird die Etablierung einer Deutschen Daten-Treuhand vorgestellt (Abschnitt 8.2).
Diskutiert wird im Abschnitt 8.3 ein Konzept, durch Rollenverteilung den Personenbezug von Daten zu vermeiden.
Von großem Interesse sind auch erste Erfahrungen bei der Implementierung von Open-Data-Ansätzen (Abschnitt 8.4).

8.1 Privacy Preserving Data Mining

Einer der größten Risikofaktoren für Big-Data-Projekte liegt in den Anforderungen hinsichtlich des Datenschutzes. Hemmnisse liegen nicht nur in rechtlicher Anforderungen hinsichtlich personenbezogener Daten – etwa aus Bundesdatenschutzgesetz und EU-Datenschutzrichtlinie. Auch die Angst vor schlechter PR[208] oder die Angst, geschäftskritische Daten für Analysen herauszugeben – etwa für eine branchenweite Betrugserkennung – können eine Big-Data-Idee trotz klar erkennbarem Nutzen blockieren.

Während die Sicherheit von kritischen Daten beim Big Data im Sinne der reinen Speicherung von Daten durch Standard-IT-Sicherheitsansätze erfüllt werden kann, liegt die Situation beim Big Data im Sinne der Analyse und Nutzbarmachung dieser großen Datensätze anders. Hier existiert in vielen Fällen gerade das Interesse, die Ergebnisse der Analyse öffentlich zu machen, zum Beispiel indem als Ergebnis einer klinischen Studie neue Behandlungsmöglichkeiten identifiziert werden, Erkenntnisse zu Risikofaktoren in Versicherungsverträgen zur Preiskalkulation genutzt werden oder in der Fraud Detection neue Betrugsmuster zur Betrugsabwehr eingesetzt werden sollen.

Durch immer mehr und immer hochdimensionalere Daten wird es hier immer einfacher möglich, sehr individuelle Muster zu finden, die auf kleine Untergruppen von Fällen angepasst sind. Dadurch steigt die Gefahr, dass sich unabsichtlich aus publizierten Mustern und Ergebnissen Informationen über einzelne Personen zurückschließen lassen. Als Beispiel: das Muster »junge Kunden

[208] Stichwort Datenkrake

verursachen höhere Schäden in der KFZ-Versicherung« ist sicherlich völlig unkritisch und publizierbar. Aber aus dem Muster »Porschefahrer unter 20 Jahren mit Wohnsitz in der PLZ 53727 verursachen häufiger Schäden über 1 Mio. Euro« lassen sich mit hoher Wahrscheinlichkeit personenbezogene Informationen zurückschließen – gerade wenn andere Informationsquellen wie Lokalnachrichten aus dem Internet zur Verfügung stehen.

Das Privacy-Preserving Data Mining beschäftigt sich mit der Frage, wie Garantien über den Datenschutz in die Datenanalyse integriert werden können. Aufgrund der Komplexität der Fragestellung gibt es dabei kein Patentrezept, verschiedene Ansätze haben sich aber als sinnvoll herauskristallisiert.

Ansatz Anonymize-and-Mine

Beim Ansatz Anonymize-and-Mine werden Daten zuerst anonymisiert (bzw. pseudonymisiert). Dies geschieht, indem gezielt Informationen weggelassen werden, bis klar definierte Anonymitätskriterien erfüllt sind. Die gebräuchlichsten Anonymitätsmaße sind hier die k-Anonymität, l-Diversität und t-Closeness. Geeignete Tools dafür sind frei verfügbar. Der Vorteil dieses Ansatzes ist, dass einmal anonymisierte Daten bedenkenlos weiterverarbeitet werden können, die kritischen Informationen sind ein für allemal zerstört. Der Nachteil ist, dass dies ungerichtet erfolgt und dabei auch Informationen, die für eine Analyse sehr relevant sein könnten, zerstört werden können. Gerade bei sehr hochdimensionalen Daten – typisch für Big Data – ist dies ein sehr schwieriges Problem. Als Beispiel: wenn das Data Mining auf sehr hochdimensionalen Versicherungsdaten herausfindet, dass nur Alter und Wohnort für das Risiko relevant sind, ist eine Anonymisierung einfach. Wird allerdings vorher anonymisiert ist es sehr einfach möglich, dass gerade Alter und Wohnort aus den Daten entfernt werden, da sie zusammen mit wenigen anderen Daten eine Identifikation erlauben.

Ansatz Mine-and-Anonymize

Der Ansatz Mine-and-Anonymize geht die entgegengesetzte Richtung: die Datenanalyse erfolgt auf nichtanonymisierten Daten, erst für das Ergebnis werden Datenschutzgarantien gegeben. Dies erfolgt entweder durch ein geeignetes Post-Processing – Ergebnisse, die vorgegebenen Privacy-Kriterien widersprechen, werden herausgefiltert – oder durch den direkten Einbau der Kriterien in den Data-Mining-Algorithmus. Dadurch ist die Umsetzung dieses Ansatzes eher kompliziert – für jedes Data-Mining-Verfahren und jede Datenschutzanforderung muss eine eigene Implementierung erfolgen – aber dadurch sind hier auch die besten Ergebnisse zu erwarten.

Secure Distributed Computing

Ein Ansatz, der sich gerade bei verteilten Daten eignet, ist das Secure Distributed Computing. Typische Einsatzfälle sind, wenn mehrere Unternehmen bei der Datenanalyse kooperieren wollen – etwa um Betrugsmuster zu finden – ohne ihre eigenen Daten herauszugeben oder die verschiedenen Informationen über dieselben Personen an mehreren Stellen getrennt gespeichert sind und aus Sicherheitsgründen keine kombinierte Datenbank in Betracht kommt. Mittels spezieller kryptographischer Techniken lassen sich Data-Mining-Algorithmen umsetzen, die dieselben Ergebnisse erzeugen wie bei einer klassischen Analyse auf einem kombinierten Datensatz, ohne dass die einzelnen Daten exportiert werden müssen oder erschließbar sind. Auch hier sind hochqualitative Ergebnisse zu erwarten, bei der Umsetzung handelt es sich aber wiederum um Speziallösungen, die zudem aufgrund der eingesetzten kryptographischen Verfahren sehr laufzeitintensiv sind.

Zusammengefasst lässt sich sagen, dass das Privacy-Preserving Data Mining sehr gute Ansätze liefert, Datenschutzanforderungen mit mathematischen Garantien zu erfüllen. Aufgrund der Komplexität der Fragestellungen sollten diese Fragestellungen aber auf jeden Fall direkt zu Beginn eines Big-Data-Projektes adressiert werden, um effektive Lösungen zu finden.

8.2 Custodian Gateways und ihre Einsatzmöglichkeiten bei Big-Data-Anwendungen

Die Verwertung und Vermarktung persönlicher Daten in digitaler Form nimmt stetig zu. Doch ebenso rasant werden die daraus resultierenden Spannungen und Probleme wachsen, sofern versäumt wird, den ordnungspolitischen Rahmen in Bezug auf die Nutzung dieser Daten entsprechend vorzugeben. Denn die großen Nutznießer der persönlichen Daten sind nicht die Individuen selbst, sondern jene, die diese Daten mit Hilfe von User Analytic Tools aufbereiten und gegen Entgelt zur wirtschaftlichen Verwertung anbieten.

Aktuell ist die werbetreibende Industrie der größte Adressat derartiger Datenverwertungsmodelle. Das Individuum partizipiert an der Verwertung seiner Daten hingegen in der Regel nur mittelbar durch unentgeltlich zur Verfügung gestellte Informationen oder Anwendungen (Dienste).
In der Bevölkerung zeichnet sich jedoch ein wachsendes Bewusstsein für den Wert persönlicher digitaler Daten ab, so dass davon auszugehen ist, dass die Bürger in Zukunft verstärkt nur dann eine wirtschaftliche Verwertung ihrer persönlichen Daten gestatten werden, wenn sie an den Erträgen angemessen beteiligt werden. Eine große Herausforderung liegt daher u.a. in der differenzierten monetären Bewertung persönlicher digitaler Daten. Hinzu kommen Aspekte der Daten-Sicherheit und des Verbraucherschutzes. Ebenso sind Aspekte und Potentiale der Steuer- und Wirtschaftspolitik zu berücksichtigen; so kann ein gezielter Aufbau von Verwertungsstrukturen zur Steigerung der nationalen Wertschöpfung beitragen, indem Individuen und Inhaber persönlicher digitaler Daten dabei unterstützt werden, ihre Rechte an deren Verwertung auszuüben, die Verwertung selber zu optimieren sowie ihr wirtschaftliches Potential gezielt und individuell zu nutzen.

Folglich ist ein Modell zur Verwertung persönlicher digitaler Daten, an dem sowohl Dateninhaber, Datenverwerter als auch Dienstleister der Datensicherung, Datenaufbereitung sowie des Datenschutzes beteiligt sind, zu entwickeln. Die Ziele der neuen Datenverwertung sind jedoch nicht allein mit neuen technischen Lösungen zu erreichen. Voraussetzung für die Gestaltung bzw. Steuerung einer solchen persönlichen digitalen Datenwirtschaft (PDD) ist es, zwischen den diversen Stakeholdern ein »Level Playing Field«[209] auszutarieren. Eine mögliche Antwort auf die Herausforderungen im Umgang mit persönlichen digitalen Daten ist die Etablierung einer Deutschen Daten-Treuhand (DEDATE) in Form einer öffentlich-rechtlichen Körperschaft (vgl. Abbildung 62). Sie fungiert zum einen als Steuerungseinheit, welche Spielregeln für die Nutzung und Verwertung der Daten, unter Berücksichtigung der Bedürfnisse der Marktteilnehmer, festlegt. Zum anderen gewährleistet sie die Einhaltung der vom Individuum gewährten Nutzungsrechte und nimmt ggf. die Nutzungsentgelte entgegen, welche an die Individuen ausgeschüttet werden. Vorteil eines solchen Modells ist die codierte Speicherung und kontrollierte Nutzung der persönlichen Daten. Der Datentreuhänder (Custodian) verfolgt keine wirtschaftlichen Interessen durch die Verwertung der Daten, sondern muss allein seiner Aufgabe der Datenspeicherung und sicherung gerecht werden. Dadurch kann auch folgenden Herausforderungen im Zuge der Verwertung persönlicher digitaler Daten begegnet werden:

- Aushöhlung und Missbrauch ziviler und kommerzieller Rechte der Dateneigentümer,
- Etablierung von unseriösen oder kriminellen Akteuren auf dem Markt der Datenerfassung und -verwaltung,
- Verhinderung eines Marktgleichgewichts auf dem Markt für persönliche digitale Daten und der Ausschöpfung der Innovations- und Wertschöpfungspotentiale dieser Daten bei Behinderung oder Blockade des Zugriffs auf freigegebene persönliche digitale

[209] Der Begriff bezieht sich auf den sich gegenwärtig intensivierenden Kampf um die Gewinne aus persönlichen Daten (»battle for share«), der durch Marktteilnehmer ohne marktbeherrschende Position hervorgerufen wird. Diese Tendenzen gehen von Individuen, Nutzer-Communities und Konzernen aus allen möglichen Branchen sowie auch der Regierung aus. Damit erlangt die persönliche digitale Datenwirtschaft eine erhebliche gesamtwirtschaftliche Bedeutung, die bei der Erstellung der volkswirtschaftlichen Gesamtrechnung in Betracht gezogen werden muss. Als logische Folge dieses Prozesses müssen Änderungen in der Besteuerung, bei den rechtlichen und ordnungspolitischen Rahmenbedingungen usw. vorgenommen werden.

Daten aus wirtschaftlichen Interessen, wenn der Datenverwalter gleichzeitig Verwertungsinteressen verfolgt und

- Einführung einheitlicher Verwertungsregeln für alle persönlichen digitalen Daten ohne Berücksichtigung von deren Sensitivität, des Verwertungskontextes und des Verwerters (öffentlich oder privat).

Zudem kann ein Treuhandmodell der codierten Datenspeicherung und sicheren Datenverwaltung

- die Anonymität der Nutzer gewährleisten,
- der unterschiedlichen Sensitivität der einzelnen Datenarten gerecht werden,
- den Kontext der Datennutzung kontrollieren und
- zwischen der Datenverwertung durch öffentliche Instanzen und kommerzielle Nachfrager unterscheiden.

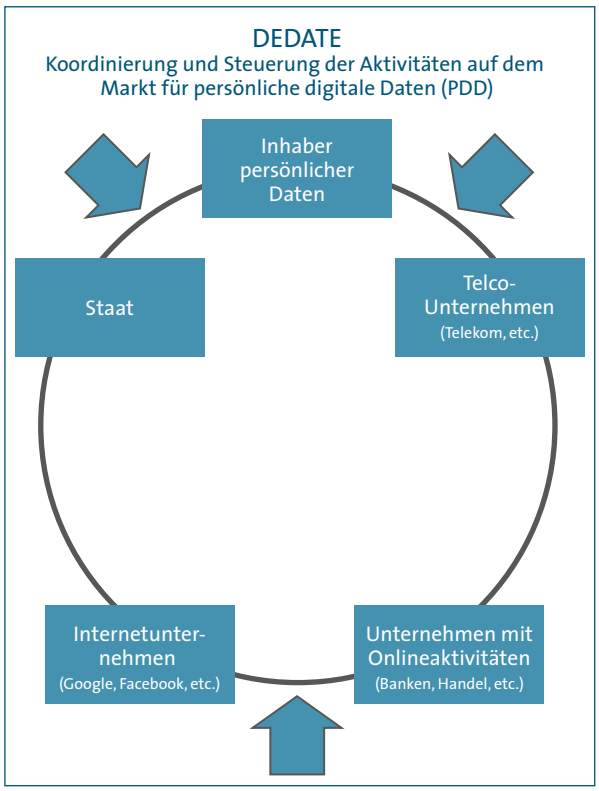


Abbildung 62: DEDATE als Koordinations- und Steuerungseinheit des Marktes für persönliche digitale Daten

Die Ausgestaltung des Marktes für persönliche digitale Daten erfordert die Definition von Rahmenbedingungen und Strukturen sowie den Aufbau einer technischen Infrastruktur und entsprechender Dienste für den Handel von Daten samt zugehörigen Geschäftsmodellen für die Marktteilnehmer. Die Bandbreite der bei der Initiative zu berücksichtigenden Forschungsfelder und Stakeholder[210] sowie die Auswirkungen der Entscheidungen auf andere Forschungsfelder legt einen multidisziplinären und integrativen Forschungsansatz nahe, bei dem in einem iterativen Prozess unter Berücksichtigung aller Stakeholder die – nach derzeitigem Stand fünf – relevanten Forschungsbereiche für die Initiative definiert werden:

1. Der erste Forschungsbereich umfasst das Themengebiet Technik und Sicherheit. In diesem Bereich ist u.a. zu klären, mittels welcher Technologien sich eine sichere, stabile und zu bisherigen Systemen kompatible Plattform für die Speicherung und den Handel von Daten realisieren lässt.

2. Ein weiterer Forschungsbereich betrifft den Rechtsrahmen. Hierunter werden rechtliche Fragestellungen für den entstehenden Markt für persönliche digitale Daten zusammengefasst. Zu klären ist in diesem Zusammenhang u.a. die steuerrechtliche Behandlung von Transaktionen mit persönlichen Daten im Netz, die rechtliche Fixierung des »Individuum Continuum«[211], die handelsrechtlichen Rahmenbedingungen für den Marktplatz für persönliche Daten, die Rechtsform und rechtliche Grundlagen für die Treuhandgesellschaft sowie Aspekte des Verbraucherschutzes.

3. Im Forschungsbereich Finanzen bzw. volkswirtschaftliche Auswirkungen werden die volkswirtschaftlichen Effekte einer entstehenden persönlichen digitalen Datenwirtschaft betrachtet. Hierunter fällt insbesondere der Einfluss auf Bestandsindustrien sowie sich ergebende Potenziale für neue Industrien, die Wirkung der neuen persönlichen digitalen Datenwirtschaft auf

[210] Politik, Wirtschaft, Individuen und Verbänden, Foren und Communities etc.
[211] Als »Individuum Continuum« wird die Gleichstellung der Rechte für das Individuum innerhalb und außerhalb des Netzes bezeichnet..

die volkswirtschaftliche Gesamtrechnung, Auswirkungen auf die Geldmenge und das Steueraufkommen und Fragen zur Konvertierung der neuen Währung (persönliche Daten) auf dem Handelsplatz.

4. Neben der volkswirtschaftlichen Betrachtung ist auch die Erarbeitung betriebswirtschaftlicher Potenziale in Form von möglichen Geschäftsmodellen für die Marktteilnehmer vorgesehen. In dem Forschungsbereich Betriebswirtschaftliche Auswirkungen/ Geschäftsmodelle werden potenzielle Geschäftsmodelle für Telekommunikationsanbieter, Anbieter von Inhalten und andere Endnutzerdienste, heutige Oligarchen wie Google, Facebook etc. und Chancen für Start-Ups ausgearbeitet. Hierzu sollen Forschungskooperationen mit Unternehmen aus den jeweiligen Bereichen gebildet werden.

5. Letztendlich ist der Kommunikation und Aufklärung zum Themenfeld persönliche digitale Daten eine besondere Bedeutung beizumessen. Daher ist es sinnvoll, einen weiteren Forschungsbereich der Analyse der Sozialverträglichkeit und der Aufklärung der Bevölkerung zu widmen.

Am Beispiel des Smart Metering[212] soll aufgezeigt werden, dass die Entwicklung einer Treuhänder-Plattform für persönliche (Energie-)Daten (Custodian Gateway Administrator) nicht nur Auswirkungen für die Internetwirtschaft mit sich bringt, sondern auch für traditionelle Industrie- und Dienstleistungsunternehmen.

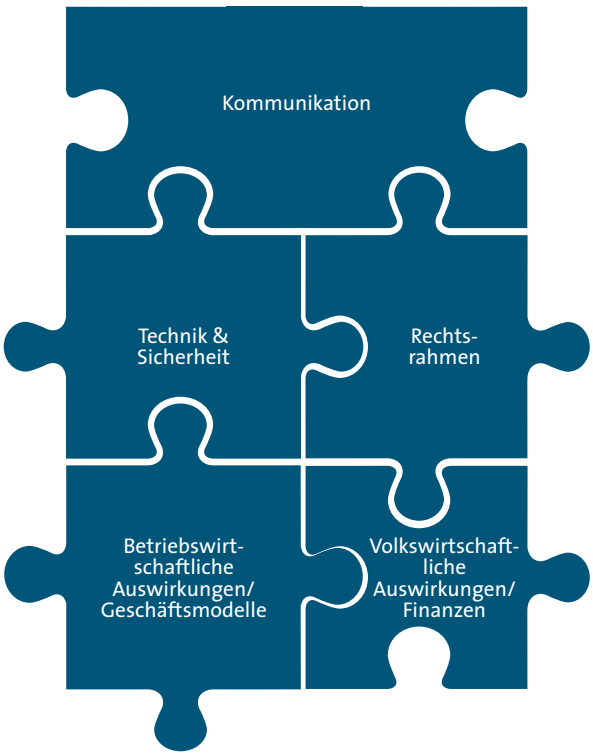


Abbildung 63: Forschungsbereiche des integrierten Forschungsansatzes

Durch die Erfassung von umfangreichen Verbrauchsdaten der Energiekunden bei Smart Metering verfügen die Energieversorger über Daten, die auch für andere Unternehmen von Interesse sein können. Viele Aktivitäten im Haushalt sind mit dem Verbrauch von Energie verbunden. So können durch die Auswertung der Energieverbrauchsdaten Kundenprofile generiert werden. Neben der primär angestrebten Optimierung der Netzauslastung und Energieversorgung mit Hilfe der Energieverbrauchsdaten ergibt sich für Energieversorgungsunternehmen ein mögliches neues Geschäftsmodell im Bereich des Handels mit anfallenden Kundendaten.

[212] Smart Meter (Intelligente Zähler) sind an das Kommunikationsnetz angebunden und ermöglichen es Energieversorgungsunternehmen, die Zählerstände von Kunden für Strom, Gas, Wasser und Wärme aus der Ferne abzurufen. Zusätzlich bietet Smart Metering in der Regel weitere Funktionalitäten wie die Erfassung von dezentral eingespeister Energie (z. B. durch die häusliche Photovoltaik-Anlage), die automatisierte Weiterverarbeitung der Daten, die Möglichkeit der Fernsperrung oder Leistungsbegrenzungen. Seit 2010 sind die Netzbetreiber in Deutschland gesetzlich verpflichtet, in allen Neubauten, bei Totalsanierungen und Letztverbrauchern mit einem Jahresverbrauch größer 6 000 Kilowattstunden Smart Meter einzubauen.

8.3 Datenschutzfreundliche Technologien: Verteilte Rollen

Das Datenschutzrecht regelt den Umgang mit personenbezogenen Daten, d.h. Einzelangaben über eine bestimmte oder bestimmbare Person. Nach Ansicht der deutschen Datenschutzbehörden sollen auch IP-Adressen personenbezogene Daten sein. Diese Rechtsansicht hat weitreichende Folgen für die Analyse des Surfverhaltens im Internet, denn im Datenschutzrecht gilt das sogenannte Verbotsprinzip. Daten dürfen nur verarbeitet werden, wenn ein Gesetz dies erlaubt oder der Betroffene zugestimmt hat. Webanalytics sind dann nur unter engen Voraussetzungen zulässig.

Die Befugnisse zur Analyse der Daten lassen sich erweitern, wenn man verhindert, dass die Daten einen Personenbezug erhalten. Dazu kann man die Daten anonymisieren, was aber die Analyse erschwert. Eine andere Möglichkeit besteht darin, die Informationen auf verschiedene Personen zu verteilen. Wenn ein Unternehmen nur über einen Teil der Informationen verfügt und damit die Daten nicht einer Person zuordnen kann, sind die Daten für dieses Unternehmen nicht personenbezogen.

Wie sieht so etwas in der Praxis aus? Web Analytics funktioniert üblicherweise so, dass der Betreiber einer Website eine Analytics-Anwendung einsetzt oder mit einem Analytics-Anbieter zusammenarbeitet und ihm die Daten zuleitet. Wenn Website-Betreiber und Analytics-Anbieter vollen Zugriff auf die Nutzerdaten haben, insbesondere die IP-Adresse, liegt eine Verarbeitung personenbezogener Daten vor. Dann bestehen hohe rechtliche Anforderungen an die Zulässigkeit.

Arbeitet man mit verteilten Rollen, erweitern sich die Möglichkeiten. Ein Beispiel ist die Anwendung PT 2.0 der nugg.ad AG, die sogar von der Datenschutzaufsicht Schleswig-Holstein zertifiziert worden ist. Der Website-Betreiber und der Analytics-Anbieter, nugg.ad, setzen eine unabhängige dritte Stelle ein, den sogenannten Anonymizer. Dieser Anonymizer erhält zwar die IP-Adresse des Nutzers der Website, er gibt diese Information aber nicht an nugg.ad und den Betreiber der Website weiter. Stattdessen liefert er auf Grund statistischer Auswertungen nur Empfehlungen über nutzerbezogene Werbung. Der Anonymizer protokolliert auch keine Nutzungsdaten, auch im Nachhinein ist es daher nicht möglich, Informationen mit IP-Adressen zu verknüpfen.

Diese Rollenverteilung führt dazu, dass jeder Beteiligte immer nur einen Teil der Informationen kennt. Wenn kein Beteiligter in der Lage ist, die Daten einer bestimmten Person zuzuordnen, liegen keine personenbezogenen Daten vor. Das Datenschutzrecht ist nicht anwendbar.

Die technischen Maßnahmen müssen durch vertragliche Vereinbarungen begleitet werden. Der Anonymizer ist Auftragsdatenverarbeiter von nugg.ad und Unterauftragnehmer des Betreibers der Website. Zwischen den Beteiligten bestehen vertragliche Vorkehrungen, die einen Zugriff auf die Datenverarbeitung des Anonymizers verhindern.

Dieses Konzept, durch Rollenverteilung den Personenbezug von Daten zu vermeiden, lässt sich auch auf andere Konstellationen übertragen. Gerade im Bereich der medizinischen Forschung gibt es mehrere Anwendungsbeispiele. Das Konzept steht und fällt damit, eine unabhängige Stelle mit der Verwaltung der Daten zu beauftragen und sichere vertragliche Regelungen zu schaffen, die auch einer Überprüfung durch die Datenschutzbehörde standhalten.

8.4 Transparenz gegenüber Betroffenen: Best Practices aus Open-Data-Projekten

Das Berliner Open-Data-Portal wird von Experten als die führende Anwendung in Deutschland eingeschätzt. 2010 gestartet, verzeichnet es mittlerweile bis zu 20.000 Zugriffe pro Monat.

Basierend auf den ersten Erfahrungen der Verwaltung und der Bürger bei der Implementierung von Open-Data-Ansätzen, wie z.B. in Berlin, sind folgende Schlüsse zu ziehen:

- Das politische Bekenntnis zu offenen Daten sowie Bestimmung eines Gesamtverantwortlichen für offene Daten auf Seiten der Verwaltung ist ein wesentlicher Ankerpunkt zur erfolgreichen Einführung eines Open-Data-Portals.
- Entscheidend für das Funktionieren von Open-Data-Portalen sind vor allem auch die Definition von Richtlinien zur Auswahl der Daten sowie wohldefinierte Prozesse für das Datenportal.
- Die öffentliche Verwaltung sollte mit Start des Projektes eine generelle Informationsstrategie festlegen.
- Einheitliche und klare Lizenz- und Nutzungsbestimmungen und der direkte Zugang zu den AGBs sind ein weitere Hauptbestandteil für das Aufsetzen eines solchen Projektes. Hauptaugenmerk bei der Nutzung der Daten sollte auf der Nachvollziehbarkeit der Daten gelegt werden.
- Die stabile und langfristige Bereitstellung und kontinuierliche Qualitätssicherung der Daten muss ein Schwerpunkt bei der Kommunikation mit den öffentlichen Datenlieferanten sein.
- Nutzung von klaren Standards für das Datenportal und dessen Sicherheit (Kompatibilität der Daten zu anderen Portalen, etc.) sind Voraussetzung für ein erfolgreiches Projekt.

Auf der technischen Seite liegen die Herausforderungen insbesondere bei den historisch gewachsenen Strukturen der unterschiedlichen Datenbestände. So gilt, weit fortgeschrittene Harmonisierungsprojekte, wie bei den Geodaten, mit anderen Fachgebieten zusammenzuführen. Datenkonverter sind zu entwickeln, Metadatensätze abzugleichen, Attribute festzulegen und die notwendigen Aufbereitungs- und Verarbeitungsprozesse zu definieren. In der Praxis bedeutet das, dass Datensätze, die bisher nicht maschinenlesbar waren, veröffentlicht werden müssen. Datensätze, die bereits online sind, werden zumeist in XML zur Verfügung gestellt, damit sie von Dritten verwendet, bei Bedarf angereichert und veröffentlicht werden können.

Die einzelnen Fachportale werden weiter bestehen und die Datensätze pflegen. Nur dort können die Daten gut gepflegt und bearbeitet werden, so dass eine Open-Data-Portal sich am besten in einem Content-Management-System (CMS) abbilden lässt.

9 Herausforderungen im Betrieb von Big-Data-Lösungen

Im Betrieb geht es darum, Big-Data-Lösungen effizient und zuverlässig zu installieren, verwalten, erweitern und verändern. Es gilt, das Zusammenspiel der verschiedenen Technologien über alle Ebenen einer Lösung hinweg (physische Infrastruktur, Daten-Haltung und -bereitstellung, analytische Verarbeitung, Visualisierung, Daten-Integration, Governance und Daten-Sicherheit) zu beherrschen.

Der Betrieb von Big-Data-Lösungen konfrontiert Unternehmen mit einer ganzen Reihe von Herausforderungen. Das Ökosystem aus Big-Data-Technologien und -Anbietern ist fragmentiert, entwickelt sich aber mit hoher Geschwindigkeit. Out-of-the-Box-Lösungen sind rar, da Lösungen meist auf die speziellen Anforderungen des Unternehmensumfeldes zugeschnitten werden. Die Situation verstärkt sich außerdem durch den Mangel an Big-Data-Spezialisten und Architekten in vielen Unternehmen. Aus diesem Grunde richten viele Anwender auch ihre Blicke auf die spezialisierten Service Provider, die ihre Big- Data-Lösungen managen und betreiben.

Das Kapitel 9 bietet Unternehmen, die ihre Big-Data-Lösungen im eigenen IT-Umfeld betreiben wollen, Empfehlungen und Check-Listen an.

Entscheidend für das Lösungsdesign und die Auswahl von Technologie-Partnern sind die Geschäftsanforderungen und die Einsatzgebiete, die eine große Auswirkung auf das operationale Modell der Lösung haben.

Unternehmen stehen heute vor der Wahl,

- zusammen mit einem Spezialanbieter und seiner Ende-zu-Ende-Technologie für ein begrenztes Spektrum von Einsatzgebieten zu beginnen oder alternativ
- eine gesamtheitliche Big-Data-Plattform zu etablieren, die vom Leistungsspektrum her verschiedene Mandanten und Use Case-Anforderungen unterschiedlichster Art auf Basis einer Universal-Plattform bedient.

Der Abschnitt 9.1 betrachtet die operationalen Implikationen einer universellen, unternehmensweiten Big-Data-Plattform auf Basis von Hadoop.

Auf dem Weg zu einer Big-Data-Plattform sind zahlreiche Entscheidungen zu fällen und Herausforderungen zu bewältigen. Dafür sollen Anregungen vermittelt werden, wie man durch die Komplexität der Architektur, Technologie und der operationalen Implikationen für die gegebenen Geschäftsanforderungen navigieren kann.

9.1 Betrieb einer unternehmensweiten Hadoop-Plattform

Jede Komponente des Big-Data-Technologie-Baukastens (vgl. Abbildung 2) hat spezifische, operationale und betriebliche Implikationen, die einzeln betrachtet werden sollen.

Physische Infrastruktur

Aus Sicht einer operationalen Architektur sind Überlegungen zur physischen Infrastruktur anzustellen und Fragen zu klären, die es ermöglichen, eine passende Umgebung im IT-Umfeld des Unternehmens zu definieren.

Deployment-Modell

Welches grundlegende Deployment-Modell ist aus Sicht des Unternehmens das richtige?
Heute bietet sich ein breites Spektrum an:

- Offpremise-Option in einer Public-Cloud-Umgebung auf Basis eines mandantenfähigen Hadoop-as-a-Service Modelles. Die operationale Verantwortung liegt beim Cloud- bzw. Big-Data-Service-Anbieter.
- Onpremise Option auf Basis einer auf das Unternehmen zugeschnittenen Hadoop-Umgebung im eigenen Rechenzentrum.

Für ein Deployment im eigenen Rechenzentrum gibt es mehrere Möglichkeiten, die Infrastruktur im Rechenzentrum zu gestalten:

- Zum einen werden im Markt werden heute Appliance-basierte Lösungen für Hadoop angeboten, bei denen in sich optimierte Hardware- und Software-Komponenten eine standardisierte Plattform bilden.

alternativ

- Rechenzentren können auf Basis von kommerziellen oder Open-Source-basierten Hadoop-Distributionen, für die heute geeignete Hardware-Referenzarchitekturen und Deployment-Modelle vorhanden sind, eine Abbildung auf die standardisierte Infrastruktur des eigenen Rechenzentrums vornehmen und dort schon im RZ befindliche Hardware-Bausteine[213] nutzen.

Auf jeden Fall sollten folgende Fragen überprüft bzw. auch architekturell entschieden werden:

- Läuft die Big-Data-Plattform in virtualisierten Umgebungen (z. B. VMware)?
- Läuft die Plattform ebenfalls in Cloud-Instanzen wie z. B. Amazon, vCloud, Azure?
- Welche Appliance-Konfigurationen sind sinnvoll einsetzbar (speziell für BI Use Cases)?
- Welche Betriebssystem-Umgebung ist im RZ relevant?
- Welche modernen Netzwerkarchitekturen werden unterstützt?
- Wie unterstützt die Umgebung Netzwerk-Isolation im VLAN/VXLAN Kontext?
- Wie sieht es mit dem IPV6-Support aus und welche zukünftigen Software-Defined Networking Standards sind zu etablieren?

Daten-Lokationsanforderungen

Welche Daten-Lokationsanforderungen sind mit Blick auf die rechtlichen und regulatorischen Anforderungen in der Plattform zu erfüllen?
Zwei wichtige Fragen stellen sich vermehrt beim Deployment im Rechenzentrum, sowohl bei kommerzieller als auch bei Open-Source-Software nämlich die Frage der Virtualisierung und die Auswahl der richtigen Speicherarchitektur. Beide sind elementar für die Ausprägung der physischen Infrastruktur einer Hadoop Umgebung.

Betrieb des Hadoop-Clusters

Wird der Hadoop Cluster direkt auf der Server-Infrastruktur (bare-metal) oder virtualisiert betrieben?
Vor allen Dingen in produktiven Umgebungen laufen die meisten Hadoop-Implementierungen heutzutage

[213] Enterprise Grade oder Commodity

direkt auf den Servern Virtualisierung wird aber mehr und mehr eingesetzt, um flexibler und agiler zu werden, unterschiedliche Fehler-Domänen besser in den Griff zu bekommen sowie unterschiedliche Hardware-Pools zur Separierung von Entwicklungs- und Produktionsclustern zu gestalten. Ausserdem eignet sich der Virtualisierungsansatz für die explorativen Analysen der Data Scientists, bei denen ständig veränderte Algorithmen die Daten analysieren und hier die Hadoop-Cluster sehr dynamisch bereitgestellt werden können.

Virtualisierung generiert einen geringen Performance-Overhead und zusätzliche finanzielle Kosten, hilft aber bei drei Fragestellungen und Anforderungen:

- Isolation von RZ-Ressourcen, um Kapazitätsreserven bereitzustellen und die unterschiedlichen Workload-Anforderungen von mehreren Mandanten und Abteilungen besser abzubilden.[214]
- Isolation von unterschiedlichen Versionen, die dem Unternehmen die Möglichkeit einräumt, parallel unterschiedliche Betriebssysteme, Anwendungen und Distributionsversionen zu betreiben. [215]
- Security Isolation – strikte Datenisolation und Privacy Compliance zwischen Usern bzw. Gruppen.

Speicher

Sollten Unternehmen ihren Hadoop-Cluster auf teuren Enterprise-Grade Speicher-Netzen [216] oder eher auf Commodity Servern mit Direct Attached Storage[217] abbilden?

Viele Hadoop-Cluster-Implementierungen laufen derzeit auf Commodity Servern. In diesem Bereich verfolgen die Vendoren verschiedene Ansätze, um mit etwas höheren Investitions- die Betriebskosten der Infrastruktur zu vermindern.

Beim Network Storage handelt es sich um spezialisierte Speicher-Arrays, die für Hadoop Workloads und deren spezielle Bandbreitenanforderungen optimiert sind[218]. Die Anschaffungskosten pro Terabyte sind höher als bei White Box Servern[219], aber die Total Cost of Ownership kann niedriger sein, wenn die Einkaufsabteilung geschickt agiert, da für Enterprise-Grade Speicher gilt:

- Sie liefern eine bessere Datennutzung und geringere Replikationsausprägungen im Hadoop Distributed File System[220].
- Sie zeichnen sich durch eine verbesserte Manageabarkeit und Performance aus. Außerdem erlaubt der Einsatz von präventiven Maintenance-Verfahren eine bessere Balancierung des Clusters.
- Es eröffnen sich Einsparungsmöglichkeiten im gesamten Software-Stack einschließlich der Hadoop-Lizensierungen, so dass die Cluster effizienter aufgebaut werden können.
- Es bieten sich bessere Möglichkeiten der Re-Balancierung des Compute- und Storage-Verhältnisses[222].

Daten-Haltung

Um die richtigen Entscheidungen für das Datenmanagement zu treffen, sind in Abhängigkeit von der Unternehmenssituation folgende Fragen zu klären:

- Welche Hadoop-Distribution deckt die heutigen und zukünftigen Anforderungen bestmöglich ab?

---

214 So kann verhindert werden, dass sich Ressourcen-intensive Jobs negative über Mandantengrenzen hinweg bemerkbar machen. (Hadoop hat heute nur ein paar eingeschränkte Möglichkeiten dieses auch ohne Virtualisierung abzubilden).

215 Das ist vor allen Dingen in Test-/Entwicklungsumgebungen oder in der Produktion bei unterschiedlichen Anforderungen (High-Performance- oder Low-Cost-Betrieb) wichtig.

216 Network Attached Storage

217 integrierte Direct Access Storage Devices (DASD)

218 Beispiel: NetApp's Engenio E Series Storage.

219 White Box Server werden aus standardisierten Komponenten von Integratoren oder Providern assembliert und nicht unter einem Marken-Namen vertrieben.

220 Abbildung von 1.3 – 2 Kopien der Daten im Gegensatz zum im Standard Hadoop üblichen Default von 3 Kopien der Daten im Direct Attached Storage

221 Plattenfehlerrate < 1% gegenüber 2-3% im Direct Attached Storage

222 Je nach Workload gibt es unterschiedliche Anforderungen der Compute- und IO-Kapazität.

- Welche Hadoop-Distribution erschließt am besten die auch weiter zu erwartenden Open-Source-Innovationen und kompensiert die Vorzüge proprietärer Lösungen?

Eine detaillierte Erörterung dieser Fragen würde den Leitfaden sprengen – es sei jedoch auf den Abschnitt 6.4 verwiesen.

Daten-Zugriff

Seinen traditionellen Fokus hat Hadoop im Bereich des effizienten Batch-Processings von MapReduce-Workloads. Hadoop entwickelt sich in Richtung interaktiver und Near-Time-orientierter Einsatzbereiche sehr schnell weiter. Für Aufgaben aus dem Bereich Big Data Analytics auf Basis strukturierter Daten kommen heute häufig In-Memory-Lösungen zum Einsatz.

Dagegen empfiehlt sich für Aufgaben der Big Data Analytics auf unstrukturierten und semi-strukturierten Daten das Hadoop-Modell mit seiner optimierten Scale-out-Architektur. Die Erweiterung von Hadoop um SQL-Zugriffsmöglichkeiten ermöglicht es mittlerweile, auch Aufgaben mit strukturierten Daten sehr effizient und mit hoher Geschwindigkeit im Daten-Integrationsbereich der Big-Data-Architektur zu verarbeiten.

Folgende Fragen sind zu beantworten, um Anforderungen aus der Datenverarbeitung an die Plattform abzubilden:

- Welche Typen von Analytics[223] werden für die Verarbeitung benötigt?
- Wie schnell müssen Entscheidungen umgesetzt werden?[224]
- Ist eine Multi-Step-Verarbeitung[225] der Daten notwendig, bevor sie gespeichert werden?
- Sind Stream Computing und CEP notwendig? Sind spezifische zeitkritische SLAs einzuhalten? Ist ein partieller Datenverlust akzeptabel?
- Wie häufig werden Daten geändert und abgefragt[226]?
- Wie eng sind die Daten im Hadoop Cluster mit existierenden, relationalen Daten verknüpft und gekoppelt?
- Welche nicht-relationale Datenbank[227] passt zu den Unternehmensanforderungen?

Daten-Integration

Im Bereich Daten-Integration sind folgende Fragestellungen wichtig:

- Welche Datenquellen[228] bieten eine Wertschöpfung für das Unternehmen und die Einsatzbereiche?
- Welche Datenschutz-Vorschriften gelten für die mit Social-Media-Werkzeugen generierten bzw. personenbezogenen Daten?
- Welche Datenvolumina sind zu bearbeiten und welche Datenstrukturen sind relevant?
- Welche Latenzanforderungen bestehen für die Daten?

Zur Integration von Datei-basierten Applikationen eignet sich NFS, um darüber die Daten in einen Hadoop-Cluster zu laden. Hingegen nutzen Web-basierte Applikationen und Services eher einen Zugriff auf Hadoop über REST API's und WebHDFS-Mechanismen. Für die Integration von Hadoop-Clustern in die BI-Landschaften der Unternehmen stehen SQL-basierte Zugriffs-Schnittstellen zur Verfügung. Auf diesem Wege können BI-Tools über ODBC/JDBC-Zugriffe die in Hadoop gespeicherten Daten nutzen.

Bei der SQL-Schnittstelle ist zu beachten, welche SQL-ANSI-Standards unterstützt werden, damit die BI-Tools effizient eingesetzt werden können.

---

[223] Machine Learning, Statistical Analysis. Predictive Analytics

[224] Latenz der Entscheidung

[225] Multi-Step-Verarbeitung steht für die mehrfache Analyse und Auswertung der Datenströme innerhalb eines Verarbeitungsjobs – so kann z.B. nach einer Text-Analyse noch eine Social-Media-Analyse und dann eine GPS/Wetter-Datenanalyse stattfinden.

[226] Realtime vs. Batch

[227] Hbase and Cassandra arbeiten nativ auf dem HDFS, während Couchbase und MongoDB auf eigenen Datenkopien arbeiten.

[228] intern vs. extern, Social/People vs. Maschinen-generierter Daten

IT-Sicherheit

Im Bereich IT-Sicherheit sind folgende Fragestellungen relevant:

- **Daten-Isolation:**
  Wie werden Daten-Nodes in einer mandantenfähigen Hadoop-Struktur voneinander isoliert?
- **Access Management:**
  Welche Zugriffskontrollmechanismen werden von den Hadoop Systemen unterstützt und sind relevant – Kerberos oder LDAP Support
- **Security Auditing und Monitoring:**
  Wer hat Änderungen auf den Hadoop-Filesets oder in der System-Plattform vorgenommen? Welche Softwarelösung unterstützt eine durchgehende Security-Audit-Funktionalität über Hadoop und die anderen Unternehmens-Datenbanken hinweg?
- **Datenverschlüsselung:**
  Welche Verschlüsselungsmöglichkeiten sind auf den Data Nodes verfügbar? Wird eine transparente Entschlüsselung on-the-fly ermöglicht?

Weitere Betriebskriterien

In diesem Bereich sind folgende Fragen zu klären:

- Wie kritisch sind Backup und Recovery[229]?
- Ist Disaster Recovery auf den Rohdaten erforderlich?
- Wie kann die Total Cost of Ownership über den Lebenszyklus des Clusters weiter optimiert werden?
- Wie stellt der Betrieb sicher, dass der Cluster hinsichtlich der Performanz ausbalanciert bleibt, auch wenn über die Zeit der Hardware-Pool aus heterogenen Infrastrukturelementen bestehen wird?
- Welche Implikationen sind bei Migrationen über verschiedene Distributionen oder Versionen einer Distribution zu erwarten? Erlaubt die Umgebung Rolling Upgrades zur Minimierung von Unterbrechungen?
- Welche Ausprägung an Mandantenfähigkeit ist zu implementieren?[230]
- Wie werden im Unternehmen die Fähigkeiten und Talente der Mitarbeiter entwickelt , um eine Big-Data-Plattform und Hadoop auch operativ zu unterstützen?

Bei der Auswahl einer Hadoop-Distribution müssen zahlreiche Kriterien bewertet und mit den Unternehmensanforderungen abgeglichen werden.

Zu den wirtschaftlichen Kriterien gehören u.a.:

- Commitment des Herstellers zur Distribution und ihrer Enterprise-Fähigkeit
- Anzahl der Entwickler beim Hersteller in den relevanten Projekten des Hadoop-Ökosystems
- Wettbewerbsfähigkeit und Flexibilität des Preismodelles[232]
- Grad des Vendor-Lock-Ins im Vergleich zu proprietären Innovationen
- Breite des Trainings- und des Zertifizierungsangebotes
- 24/7-Support
- Relevantes Software-Anbieter-Ökosystem, das für die Distribution zertifiziert ist[233].

Die hier dargestellten Kriterien dienen der Entscheidungsfindung und technologischen Betrachtung für eine Hadoop-Plattform und ihre Charakteristika, die aus Sicht des Betriebes notwendig sind.

---

[229] Anzahl Kopien im HDFS-Cluster

[230] Selbst innerhalb eines Unternehmens muss ein Universal Hadoop-Cluster unterschiedliche Geschäftsbereiche und möglicherweise unterschiedliche Tochterunternehmen bedienen.

[231] Beispiele: Entwicklertraining für Pig, Datenbankadministratoren-Training für Hive, IT-Operations-Training für die gesamte Hadoop-Plattform.

[232] Physische Nodes, Speicher, virtualisierte Umgebungen und Lizenzgrenzen, Pay-as-you-go, Perpetual-Lizenzen

[233] relevant auch aus Sicht von EDW- & BI-Umgebungen

Bereich	Kriterium	Fähigkeiten von Open Source Hadoop 2.0
Data Management	Unterstützung von Batch, interaktiv, Online & Streaming Use Cases	YARN ■ erlaubt gemischte Workloads, auch non-MapReduce Verarbeitung (viele interaktive, einige Real-time, sowie Graph-Datenbank Verarbeitung) ■ Application Master Konzept isoliert Mandanten, damit bessere Security ■ Resource Manager verbessert Resourcennutzung im Cluster ■ Rolling upgrades für alle Versionen ab Hadoop 2.0. Möglichkeit für Benutzer zur Steuerung von Upgrades und Versionierung ■ Nutzung von HIVE Diensten auf Basis Apache TEZ- Interactive Queries ■ STINGER Phase 2 soll mehr SQL Funktionen liefern und und 20-80x schneller werden
	Volles Daten Lifecycle Management	FALCON Data Lifecycle Management Framework – ermöglicht Tool Orchestrierung über XML Beschreibungssprache wie z.B.: ■ Datenimport Replikation ■ Scheduling und Koordination ■ Data lifecycle Policies = Datenauditierung und Lineage ■ Multi-cluster und SLA Management. SLA wird je nach Workload durch Scheduler in YARN gemanaged und kontrolliert ■ Data Lake Handling (Data Lake Konzept – Alle Daten werden in einem Pool gesammelt und diese dann unterschiedlichsten Bereichen für Analysen zur Verfügung gestellt, um die Daten-Silos in den Unternehmen aufzuheben und damit die Wertschöpfungskette der Analysen zu erhöhen)
	Datenintegration/Real-time Ingestion	STORM ist derzeit kein Bestandteil von Hadoop 2.0. Yahoo hat Verbesserungen zur besseren Ablauffähigkeit der Messaging Technologie in YARN implementiert.
Reliability	High availability	Hohe Verfügbarkeit durch integrierte Hochverfügbarkeitsfunktionen in den Software-Komponenten
	Disaster recovery	Multi Data Center Disaster Recovery. AMBARI managed heute nur einen einzelne Cluster. Multi-Cluster Management muss durch Eigenentwicklung auf Basis der API Calls in die einzelnen Clustern implementiert werden
	Rolling upgrades ohne Service Unterbrechung	Höchstverfügbarkeit der Verarbeitung auch bei rollierenden Versionswechseln in der Software-Infrastruktur
	Fallback capability nach Versions/ Release Wechsel	Alle Konfigurationen werden in Ambari DB gespeichert. Heute funktioniert ein Zurückfallen auf die vorherige Version durch manuellen Prozess
	HDFS Snapshots	Point in time recovery über Snapshots
	Backup & recovery procedures	Dokumentiert und verfügbar über Falcon

Bereich	Kriterium	Fähigkeiten von Open Source Hadoop 2.0
Managebarkeit	Automation der Initial Installation & Upgrades	Nach der Installation des Ambari Servers (automatisierbar), Automation der Registration/Installation auf verschiedene Hosts über die Ambari API.
	Support für Automation Frameworks (Puppet, Chef)	Automation eines vollen Deployments eines Hadoop Clusters auf der Basis eines unterstützten Betriebssystemes mit der Ambari API. Ambari nutzt Puppet (standalone mode)
	End-to-end Management Framework über das Hadoop Ecosystem	Ambari föderiert das Hadoop Ecosystem in ein universelles Management Interface. REST API wurde durch ein gemeinsames Design von Red Hat, Teradata, Microsoft und Hortonworks entwickelt. REST API ermöglicht ISVs oder Unternehmen das Ersetzen von Teilen von Ambari durch ihre spezielle Lösung. Wurde bereits getan durch Teradata für Viewpoint Integration und Microsoft für die System Center Integration. Wird auch genutzt im Savannah Projekt zum Deployment von Hadoop auf OpenStack.
	Integriertes Performance Monitoring	Performance Monitoring wird derzeit über Ganglia realisiert. Es wurde schon ersetzt durch HP OpenView und mit ähnlicher Strategie kann es mit anderen Lösungen wie Tivoli, BMC, ...ersetzt werden.
	Release & Patch Management Support & Reporting	Ambari erzeugt Report über verschiedene Versionen der jeweiligen Komponenten auf den Hosts und im Cluster
	CMDB Support	Durch die Registrierung auf den Hosts hat Ambari die Information über Hardware/Operating System/Rollen etc. Diese können über API'S ausgelesen und in einer CMDB Lösung abgespeichert werden.
Security	Granulare Rollen-basierte Zugriffskontrolle via Active Directory , LDAP, Kerberos	KNOX ■ Vereinfachte Security für User und Administratoren ■ ermöglicht durchgehenden Zugriff und Single Application Feel ■ -Abstrahiert Benutzer von der Lokation ihrer Dienste
	Mandanten, Daten, Netzwerk und Namensraum-Trennung in allen Diensten	YARN ermöglicht Kunden das Anfordern und Ablehnen von Containern auf den spezifischen Hosts. Daten residieren auf jedem Node des Clusters, diese können auf den jeweiligen Hosts verschlüsselt werden. Namenode Föderierung kann konfiguriert werden – »chroot« Umgebung.
	Unterstützung für Datenverschlüsselung	Verfügbar über Filesystem Verschlüsselung oder auf OS Level oder über 3rd Party Voltage
	Auditierbarkeit	Alle Konfigurationsänderungen werden über Ambari in der Ambari DB gespeichert.

Tabelle 15: Bewertung von Betriebskriterien für Hadoop, basierend auf Hadoop 2.0

9.2 Betrieb einer unternehmensweiten Stream-basierten Real-time-Analytics-Plattform

Neben den Big-Data-Architektur-Elementen für Data at Rest, die die wichtigen Data-Store- und Analytics-Plattformen auf der Basis von Hadoop und die EDW-Plattformen umfassen, kommen in Big-Data-Einsatzfällen vermehrt Anforderungen zum Tragen, bei denen es um Data in Motion geht. Hier geht es um immensen Datenmengen, Real-time-Verarbeitung und -Analytics.

Hierbei kommen Streaming-Technologien zum Einsatz, die es ermöglichen, im Low-Latency-Bereich (im µs Bereich) auf Daten-Events zu reagieren, diese miteinander zu korrelieren, zu aggregieren, CEP sowie analytische Operationen gegen strukturierte, semi- und unstrukturierte Daten vorzunehmen, z. B.:

- Textdateien, Tabellenkalkulationen, Grafiken, Video- und Audioaufzeichnungen
- E-Mail, Chat und Instant Messaging, Webdatenverkehr, Blogs und Social Networking-Websites
- Finanztransaktionen, Daten aus dem Kundenservice, Daten aus polizeilich eingesetzter Suchsoftware, System- und Anwendungsprotokolle
- Satellitendaten, GPS-Daten, Sensorprotokolle, Daten aus Kartenlesegeräten und Zugriffsdaten.

Stream-Computing-Plattformen sind von ihrer Eigenschaft und Struktur her Applikationsserver-Container mit hoher In-Memory-Compute- und -Analyse-Fähigkeit[234].
In den Runtime-Containern der Stream-Computing-Plattform werden Daten über standardisierte Konnektoren direkt aus dem Netzwerk, über Message Queues, über direkte Connectivity mit den API-Services der Social Networks, Anbindungen an Data Warehouses oder auch durch File Ingestion in die operative Auswertungslogik eingebracht.

Die immer weiter steigenden Anforderungen an die Auswertung von Events , die z. B. aus der steigenden Anzahl von Sensoren (Internet of Things), Mobile Apps sowie GPS-Informationen und Instrumentierung von Fahrzeugen und Maschinen stammen, machen es notwendig, diese Datenvolumina in Echtzeit zu analysieren und nur solche Daten in die Data-Store-Technologien zu übertragen, die eine zeitlich längere Relevanz oder weitere Verarbeitungs- und Analytics-Funktionen benötigen.
Aus diesem Grunde werden Streaming-Technologien zum einen als High-Volume Data Ingest Service und zur Vorverarbeitung zu den Big Data Stores eingesetzt. Zum anderen ermöglichen sie Real-time-Analysen, wenn im Einsatz Low-Latency-Anforderungen zu erfüllen sind.

Typische Anwendungsbeispiele bilden:

- **Financial Services:**
  Einsatz im Bereich High Volume Trading, Real-time Trade Monitoring und Fraud Detection.
- **Telekommunikation:**
  Einsatz im Bereich Real-time Call Detail Record Auswertung mit Mobile Advertisement, Fraud Detection, dynamische Netzwerk-Optimierung.
- **Security:**
  Einsatz im Bereich Real-time Video/Audio Überwachung

Ergebnisdaten, die zur Speicherung oder Weiterverarbeitung anstehen, werden über Standard-Konnektoren und Adapter in Richtung Enterprise Service Bus, Data Warehouse oder in ein Filesystem geschrieben.
Die Streaming Runtime Container selbst enthalten keine eigenen Persistenz-Layer über ihre In-Memory Speicherbereiche hinaus.

An dieser Stelle sollen die operationalen Implikationen und Themenstellungen beispielhaft für die IBM InfoSphere Streams-Plattform dargestellt werden, um die wesentlichen Optionen und Randbedingungen für den Einsatz einer Real-time- Analytics-Plattform zu skizzieren.

[234] z. B. durch Einsatz von Text Analytics, statistischen Analysen, R-basierter Analytics und Operatoren zum Parsen, Filtern und Aggregieren von Daten

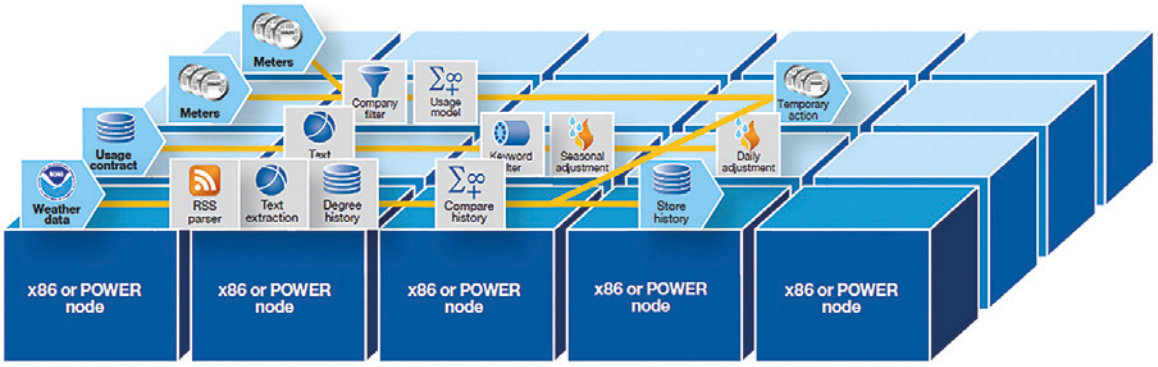


Abbildung 64: Typische Laufzeit-Umgebung einer Streams-Applikation

Physische Infrastruktur

Die leistungsstarke Verarbeitungsplattform, die die komplexen, kontinuierlichen Datenanalysen mit höchstem Durchsatz und schnellsten Reaktionszeiten ermöglicht, erlaubt einen Infrastruktur-Einstieg auf Basis von Einzelservern, die auf eine praktisch unbegrenzte Anzahl von Rechner-Knoten hochskalieren kann.

Als Hardware-Infrastruktur für die Streams-Plattform dienen standardisierte x86 oder IBM Power Linux-basierte Systeme (vgl. Abbildung 64).

Innerhalb des Clusters können je nach Nutzeranforderungen unterschiedliche Hardware-Ausprägungen betrieben werden. Streams verfügt über ein automatisiertes Deployment der Applikationen über den Cluster unter der Berücksichtigung der Kapazitätsanforderungen während der Ausführung.

Streams entscheidet automatisch zur Laufzeit, wo die einzelnen Streams-Applikationen und Operatoren ausgeführt werden. Hierfür werden Load-Balancing-Mechanismen und Verfügbarkeitsmetriken der Ausführungsplattform auf die sogenannten Processing-Elemente angewandt. Auf diese Weise kann z. B. die Ausführungsumgebung dynamisch rekonfiguriert werden, um die kontinuierliche Verarbeitung der Datenströme im Falle von Server- oder Softwarefehlern aufrecht zu erhalten.

Daten-Haltung

Plattenspeicher in den Rechenknoten wird nur für Konfigurationsdaten, die Softwareinstallation und für das Shared Filesystem[235] zur Konfiguration des Clusters benötigt.

Die Streams-Umgebung benötigt ein Daten-Management- und Sicherungskonzept, wie man es für Standard-Applikationsserver und ihre Konfigurations- und Deployment-Umgebungen heutzutage in den meisten Rechenzentren etabliert hat.

Für die Herstellung einer Hochverfügbarkeitskonstellation nutzt Streams eine Recovery-Datenbank, um Failover von Services und Instanzen reibungslos und automatisch durchzuführen.

Daten-Zugriff

Streams-Instanzen und die Streams-Applikationen halten alle zur Ausführung notwendigen Daten In-Memory vor. Externe Daten werden über Konnektoren und Adapter zugeführt bzw. aus der Streams-Applikation heraus in z. B. Datenbanken, Message Queues oder Files persistiert.

Beim Design einer Streams-Applikation ist eine Ende-zu-Ende-Betrachtung der Laufzeitarchitektur zu beachten. Informations-Feeds und die ausgehenden

[235] NFS oder GPFS

Daten-Volumina, die in Message Queues, Hadoop oder Data Warehouses gespeichert werden, müssen von ihrer Leistungsfähigkeit und ihrem IO-Verhalten an die zu erwartende Verarbeitungsgeschwindigkeit angepasst sein.

Die analytische Verarbeitungskapazität wird infrastrukturell als Compute-Kapazität bereitgestellt und skaliert linear.

Streams-Processing-Anwendungen erzeugen sehr leicht Volumina im Bereich von Zehntausenden Events pro Sekunde. Hier bietet die flexible Architektur der Streams-Plattform einen sehr einfachen Plattformbetrieb.

Daten-Integration

Die Daten-Integration der Plattform geschieht durch standardisierte Adapter und Konnektoren[236]. Für spezielle Anforderungen können mit der Streams-Entwicklungsumgebung eigene Adapter entwickelt werden.

Die Event-Verarbeitung erlaubt die Nutzung der folgenden Toolkits für Analyse-Zwecke: Advanced Text Analytics, SPSS, R analytics, TimeSeries, Geospatial, OpenCV (Video), CEP, Industry toolkit (FIX).

IT-Sicherheit

InfoSphere Streams Runtime kann auf Security-Enhanced Linux (SELinux) Umgebungen mit den entsprechenden Security Policies ablaufen.

Für die Authentisierung von Benutzern auf der Plattform kann das Pluggable Authentication Module (PAM) oder ein externer Lightweight Directory Access Protocol (LDAP) Server eingebunden werden. Die Zugriffe der Benutzer und Administratoren auf die Applikationen und ihre Konfigurationen werden durch ACLs gewährleistet.

Streams verfügt über ein integriertes Audit Logging, um alle Aktivitäten der Benutzer und Applikation nachzuweisen und nachvollziehbar zu machen.

Weitere Betriebskriterien

Die Operations Console der Streams-Plattform ermöglicht mehrere Optionen für das Monitoring und das Management der Applikationen.

Über die Console werden die Streams-Jobs gestart, die dann kontinuierliche Datenanalysen durchführen. Alle Applikationen sind über das integrierte Applikations-Repository sichtbar und können dort der Jobverarbeitung zugeführt werden. Eine Integration in andere externe Scheduler (UC4 ...) ist über Schnittstellen möglich.

Die Integration der Streams-Umgebung in IT-Monitoring-Systeme kann über die Streamstool-Command-Funktionalität einfach implementiert werden.

Die Operations Console erlaubt die Visualisierung der laufenden Applikationen und die Verbindungen zu den aktiven Datenquellen in der Verarbeitung.

Aus Sicht eines produktiven Einsatzes von Streaming-Technologien kann eingeschätzt werden, dass diese Umgebungen für den Betrieb sehr einfach und flexibel verwaltet werden können. Die Streaming-Systeme leben von der Parallelisierung der Compute- und Memory-Power heutiger Server-Umgebungen und lassen sich einfach in Cloud-basierte Umgebungen integrieren.

[236] Beispiele sind JMS, REST, JDBC/ODBC zu allen gängigen relationalen DB-Systemen, TCPIP, Files, HTTP-, HTTPS-, FTP-, FTPS, RSS- Feeds etc.

10 Big-Data-Expertise und -Know-how

Um das Potenzial von Big Data zu erschließen, ist Wissen aus Analytik, IT und dem jeweiligen Fachbereich gefragt. Bislang gibt es nur wenige Fachkräfte, die diese Kompetenzen kombinieren. Solche Data Scientists werden jedoch dringend gesucht. In den USA gehören sie schon zu den meistgesuchten technisch-wissenschaftlichen IT-Fachleuten[237], und eine Studie von McKinsey sagt für die USA eine Lücke von über 50% für die 2018 voraus. In einer Fraunhofer-Potenzialstudie wünschen sich 95% der Befragten Best Practices und Schulung als Fördermaßnahme.
Data Scientists vereinen die Rollen als Impulsgeber, Ideengeber, Lösungsarchitekt, Umsetzer, Analyst, Kommunikator, Überzeuger. Es existieren verschiedenste Schulungskonzepte für Data Scientists. Das Kapitel 10 gibt eine Übersicht über Gemeinsamkeiten und Mindestanforderungen.

Mit den Möglichkeiten zur Speicherung und Verarbeitung von Big Data wächst die Nachfrage nach Fachleuten, die solche Daten analysieren, um sie in Wert zu setzen. Sie analysieren große Datenmengen jenseits von Excel, Business Intelligence Tools und gängigen Statistikpaketen. Mit wissenschaftlichen Datenanalysemethoden entwickeln sie Modelle zur Informationsextraktion und Prognose für Big-Data-Anwendungen. Sie sind Analysten mit IT-Kompetenzen und Fachleute in ihrem Anwendungsbereich

In den USA prägten Patil und Hammerbacher von LinkedIn für solche Experten die Bezeichnung Data Scientists und im Oktober 2012 bezeichnete der Harvard Business Review den Beruf als »The sexiest job of the 21st century«[238]. In den USA gehören Data Scientists bereits zu den meistgesuchten technisch-wissenschaftlichen IT-Fachleuten, und 2012 ist dort die Zahl der Masterstudiengänge für Datenanalytik sprunghaft gestiegen. Bis die Absolventen die Hochschulen verlassen, rekrutieren amerikanische Unternehmen Mathematiker, Wirtschaftswissenschaftler, Ingenieure, Informatiker, Physiker und sogar Psychologen.

Aufgabe der Big-Data-Analysten ist es, geschäftsrelevante statistische Erkenntnisse aus den Daten zu gewinnen. Zunächst allerdings werden sie sich mit den Daten vertraut machen und sie mit explorativen Methoden untersuchen. Datenschutz und Datenqualität stellen wichtige Ansprüche, für die sie geeignete Maßnahmen finden müssen. An Business-Analysten werden aus dem Betrieb laufend neue Fragen herangetragen werden, die sie mit ad-hoc-Anfragen und weiteren Analysemethoden zu beantworten suchen. Für sich wiederholende Anfragen, Reports und Dashboards werden sie Skripte erstellen. Ein weiteres Einsatzgebiet ist die Entwicklung von statistischen Modellen für die automatisierte Datenanalyse. Solche Modelle dienen dazu, Informationen, Relationen und Metadaten zu extrahieren, irrelevante Daten herauszufiltern, Bewertungen zu berechnen, Prognosen zu erstellen oder Entscheidungen zu treffen.

Selbstverständlich müssen Data Scientists die klassischen Methoden von Statistik und Data-Mining beherrschen. Ihre besondere Kompetenz erlangen sie dadurch, dass sie weitere Verfahren anwenden können, wo die klassischen versagen: Im Umgang mit hohen Datenvolumina (verteilte Speicherung), strukturell komplexen Daten (NoSQL-Datenbanken) und der realzeitnahen Verarbeitung (Parallelverarbeitung, Streaming und in-memory Processing).

Im Gegensatz zu den USA gibt es in Deutschland noch keine Studiengänge für Data Scientists. Auch eine Websuche nach Seminaren zum Stichwort »Data Science« lieferte keine Ergebnisse für deutschsprachige

[237] Vgl. CNN Money: Best new jobs in America
[238] http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

Universitäten. Dagegen gibt es eine beachtliche Menge von Uni-Seminaren zum Suchbegriff »Big Data«. Die nächsten zwei Tabellen zeigen das Ergebnis einer Websuche am 1.10.2013, die keinen Anspruch auf Vollständigkeit oder Repräsentativität erheben kann, aber dennoch interessant ist. Tabelle 16 zeigt, dass die meisten Seminare in der Informatik angeboten werden, einige auch für die Wirtschaftsinformatik und weitere Studiengänge. Das ist eine gute Nachricht: In etwa 2 Jahren werden dem Arbeitsmarkt mehr anwendungsorientierte Informatiker mit Data-Science-Know-how zur Verfügung stehen.

	KIT Karlsruhe: Performanz	TU München	HU Berlin	Fernuni Hagen: Management	RWTH Aachen	KIT Karlsruhe: Tools	KIT Karlsruhe: Hadoop	Uni Konstanz	Uni Heidelberg	TU Dortmund	Uni Karlsruhe: Physik	Uni Augsburg	Uni Leipzig	Uni Stuttgart	LMU München: Analysis	Uni Kaiserslautern	Uni Freiburg	Uni Hildesheim	LMU München
Studiengang																			
Informatik	x	x	x	x	x	x	x	x	x			x	x	x	x	x	x		
Wirtschaftsinformatik				x						x				x					x
andere								Informationswirtschaft			Physik							Informationswissenschaft	Medieninformatik
Infrastruktur																			
Cloud						x						x	x			x			
Performanz	x	x				x							x						
Datenmanagement																			
Graphverarbeitung																			
NoSQL, In-Memory	x			x		x	x	x			x	x	x	x	x	x	x		
Anfrageverarbeitung?							x		x						x				
MapReduce	x		x	x	x	x	x		x		x	x		x	x	x	x		
Hadoop Ecosystem			x	x	x	x	x	x				x				x			
Architektur	x															x			
Approximative Algorithmen															x				
Parallelisierung		x									x								
Streaming		x	x			x			x					x					

	KIT Karlsruhe: Performanz	TU München	HU Berlin	Fernuni Hagen: Management	RWTH Aachen	KIT Karlsruhe: Tools	KIT Karlsruhe: Hadoop	Uni Konstanz	Uni Heidelberg	TU Dortmund	Uni Karlsruhe: Physik	Uni Augsburg	Uni Leipzig	Uni Stuttgart	LMU München: Analysis	Uni Kaiserslautern	Uni Freiburg	Uni Hildesheim	LMU München
Analytik																			
Mahout							x	x				x							
»Machine learning, Statistik, Analytik«									x	x	x		x						
Visualisierung														x					
Bezug zu BI												x	x	x					
Anwendung																			
Datensicherheit, Datenschutz														x					x
Wirtschaftlichkeit	x													x	x				
Soziale Netzwerke												x	x						
Mobilfunkdaten										x									
Anwendungsmöglichkeiten											x		x	x		x	x	x	
Geschäftsmodelle																			x
	Gruppe 1: Big Data Engineering								**Gruppe 2: Data Science**							**Gruppe 3: Big Data Anwendungen**			

Tabelle 16: Seminarangebote an deutschen Universitäten mit dem Stichwort »Big Data« im Titel.

Da die Beschreibungen der Seminare im Netz nicht standardisiert und unterschiedlich detailliert sind, ist ein genauer Vergleich nicht möglich. Für die Tabellen wurden deshalb wichtige Stichwörter extrahiert und den Themen Infrastruktur, Datenmanagement, Analytik und Anwendungen zugeordnet. Sortiert man die Seminare nach diesen Themen, erhält man die in den Tabellen gekennzeichneten drei Gruppen.

Alle Uni-Seminare in Tabelle 16 beschäftigen sich mit Datenmanagement/Infrastruktur, aber Gruppe 1 ausschließlich. Sie richtet sich an Big-Data-Architekten und -Entwickler. Gruppe 3 behandelt auch Anwendungsaspekte. Hier finden sich vermehrt Seminare aus Anwendungsdisziplinen der Informatik. Gruppe 2 beschäftigt sich zusätzlich mit Analytik und deckt alle drei Dimensionen der Data Science ab: IT, Analytik und Fachwissen.

Tabelle 17 zeigt ebenfalls überraschend viele Treffer mit Weiterbildungsangeboten für Berufstätige. Modulare Angebote vom selben Anbieter wurden zusammengefasst. Die Veranstaltungen reichen von 1-10 Tagen und setzen teilweise Vorkenntnisse in Statistik, Business Intelligence, Datenmanagement oder IT voraus. Die 1-2-tägigen Seminare im linken Teil der Tabelle bezwecken meist einen Überblick, praktische Übungen gibt es hier nicht. Eine Einteilung der Stichwörter in Infrastruktur, Datenmanagement, Analytik und Anwendung lässt auch in dieser Tabelle die gleichen drei Gruppen wie bei den Uni-Seminaren erkennen.

	Integrata	SCN	Oracle	Deutsche Informatik Akademie	StatSoft	DWH academy	Avantgarde Labs	Eduvision	metafinanz	EMC	Fraunhofer IAIS	IT-Schulungen.com	Management Circle	Actian	SAS
Dauer	2	2	1	1	1	1	2-4	5	5	5	8	10	2	0,5	10
Überblick/Konzepte	x	x	x	x	x	x									
Übungen									x	x	x	x			
Modular											x	x			
Voraussetzungen															
Datenmanagement	x		x						x			x			
IT	x		x						x		x	x			
Statistik										x	x				x
BI						x						x			
Infrastruktur															
Produkte und Hersteller	x	x										x			
Fertiglösungen (Appliances)			x		x									x	x

	Integrata	SCN	Oracle	Deutsche Informatik Akademie	StatSoft	DWH academy	Avantgarde Labs	Eduvision	metafinanz	EMC	Fraunhofer IAIS	IT-Schulungen.com	Management Circle	Actian	SAS
Datenmanagement															
Verteilte Systeme	x	x				x					x				
Hadoop ecosystem	x	x	x					x	x	x	x	x			
Stream Processing, Event Processing		x		x							x	x			
NoSQL		x	x			x		x			x				
Architekturen				x							x	x			
Referenzsystem			x			x					x	x			x
Bezug zur klassischen Datenhaltung	x	x									x	x		x	
Datenintegration, Import, Export				x	x		x		x			x			x
Analytik															
Datenqualität, Datenaufbereitung					x		x			x	x		x		x
Bezug zu BI						x						x		x	
»Textanalyse, Entity Recognition«				x	x		x			x	x				x
Visualisierung				x						x	x				
Data Mining, fortgeschrittene Analytik Modellierung					x		x	x		x	x		x		x
Anwendung															
Vorgehensweise im Unternehmen				x						x			x		
Anwendungsmöglichkeiten			x			x	x	x		x			x	x	
»Fallbeispiele, Praxisberichte«				x			x	x					x		
Data Governance				x									x		
	Gruppe 1: Big Data Engineering			**Gruppe 2: Data Science**									**Gruppe 3 Big Data Anwendungen**		

Tabelle 17: Seminarangebote für Berufstätige mit dem Stichwort »Big Data« oder »Data Science« im Titel

Insbesondere im Bereich der Analytik kann man in Zukunft differenziertere Angebote erwarten, wie in Tabelle 18 angedeutet. Während Basic Analytics die wichtigsten Data-Mining und maschinellen Lernverfahren enthält, können Spezialkurse Methoden der Batch-Analytik versus Stream-Analytik behandeln, eingebettete Methoden in großen Sensornetzen, die semantische Anreicherung zu Smart Data, Textanalytik und Methoden zur Analyse von Sprach-, Bild-, und Videodaten, und die visuelle Analytik.

Big Data Analytik: Potenzial, Roadmaps, Strategien				
Anwendungen in einzelnen Branchen	...	...	...	...
Visual Analytics				
Text Analytics	Multimedia Analytics	Batch Analytics	Stream Analytics	...
Embedded Analytics			Smart Data	
Basic Analytics				
Big Data Management				
Big Data Infrastruktur				

Tabelle 18: Vorschlag zur Differenzierung des Analytikangebots

An den Universitäten wären Sommerschulen und Graduiertenkollegs ein gutes Angebot für Studenten, die jetzt die Big-Data-Seminare besuchen.

11 Big Data – Ausgewählte Anbieter von Technologien, Lösungen und Know-how

Das Kapitel 11 stellt Anbieter von Technologien und Lösungen sowie von Consulting-, Engineering- und Integrations-Services im Bereich Big Data vor. Aufgenommen wurden Unternehmen und Organisationen, deren Experten an der Entwicklung dieses Leitfadens mitgewirkt haben.

11.1 Atos IT Solutions and Services

Kontaktinformationen
Atos IT Solutions and Services GmbH
Otto-Hahn-Ring 6 | 81739 München
Tel. +49 (0) 211 399 0 | de-info@atos.net
http://de.atos.net/de-de/home/unsere-leistung/business-integration-solutions/data-and-information-management.html

Über Atos
Atos SE (Societas europaea) ist ein internationaler Anbieter von IT-Dienstleistungen mit einem Jahresumsatz für 2012 von 8,8 Milliarden Euro und 77.000 Mitarbeitern in 47 Ländern. Der globale Kundenstamm des Unternehmens profitiert von einem umfangreichen Portfolio, das drei Bereiche umfasst: Beratung und Technologie-Services, Systemintegration und Managed Services sowie BPO- und transaktionsbasierte Hightech-Services durch Worldline. Mit seiner umfassenden Technologie-Expertise und tiefgreifendem Branchenwissen unterstützt der IT-Dienstleister Kunden in folgenden Marktsegmenten: Produzierendes Gewerbe, Handel, Dienstleistungen; Öffentliche Verwaltung, Gesundheits- und Transportwesen; Banken und Versicherungen; Telekommunikation, Medien und Versorgungsunternehmen.

Big-Data-Lösungen
Big Data ist ein Schwerpunkt Thema bei Atos. Atos hilft Unternehmen effiziente Strategien für einen erfolgreichen Einstieg in das Thema zu finden, die richtigen technologischen Entscheidungen zu treffen und existierende Geschäftsprozesse zu modernisieren oder neue, innovative Geschäftsmodelle zu entwickeln. Das Portfolio umfasst zudem Beratung und Technologie-Services, Systemintegration sowie Outsourcing-Dienstleistungen. Als weltweiter IT-Partner des Internationalen Olympischen Komitees ist Atos für die Gesamtleitung der Technologie-Partner verantwortlich, welche die riesige kritische IT-Infrastruktur für die Olympischen Spiele 2012 in London und 2014 in Sotschi entwickeln und betreiben. Big Data ist dabei eines der wichtigen Themen bei der Bewältigung der Herausforderungen einer Olympia-Veranstaltung.

Einsatz-Szenarien
Das IT-Unternehmen Atos hat die Leitung des »Big Data Public Private Forum«-Projekts, kurz BIG, übernommen. Im Zentrum der Diskussion soll dabei die Verarbeitung und Bedeutung von Big Data für die Wirtschaft stehen. Das BIG-Projekt hat weitreichende Bedeutung für die Wirtschaft, Wissenschaft, Politik und Öffentlichkeit. Es soll einen Maßnahmenplan für den geeigneten Einsatz von Big Data-Technologien liefern. Die Verarbeitung großer Datenmengen spielt in unserer Gesellschaft und im aktuellen Wirtschaftsumfeld eine zunehmend wichtige Rolle. Generiert werden die hohen Datenvolumina beispielsweise durch Zähler und Sensoren, die etwa Klima-, Verkehrs- oder Schadstoffbelastungsdaten einer Stadt erfassen, aber auch durch Online-Anfragen bei Reisebüros oder Positionierungs-Diensten wie OpenStreetMaps.

Die Integration, Analyse und Nutzung dieser Daten wiederum kann zur Entwicklung neuer, innovativer Produkte und Services beitragen – etwa eines Verkehrsumleitungssystems, das aktiv wird, wenn bei bestimmten Wetterbedingungen erhöhte Luftschadstoffwerte auftreten.

Big-Data-Technologien
Atos setzt als herstellerunabhängiger Integrationsanbieter auf aktuelle Big-Data-Technologien, die sich vor allem aus BI-Systemen, veränderten NoSQL-Datenbanksystemen und Speichersystemen rekrutieren.

11.2 Empolis Information Management

EMPOLIS
INFORMATION MANAGEMENT

Kontaktinformationen
Empolis Information Management GmbH
Europaallee 10 | 67657 Kaiserslautern
Tel.: +49 (0) 631 68037-0 | info@empolis.com
www.empolis.com

Big-Data-Lösungen
Empolis Smart Information Management® Lösungen befähigen Unternehmen und Organisationen, die exponentiell wachsende Menge strukturierter und unstrukturierter Daten zu analysieren, zu interpretieren und automatisiert zu verarbeiten. Sie nutzen damit ihr Wissenskapital, um unternehmenskritische Geschäftsprozesse zu optimieren. Entscheider, Mitarbeiter und Kunden erhalten so stets situations- und aufgabengerecht genau die Information, die für sie relevant ist.
Dabei werden die in einem Component Content Management System erstellten und verwalteten Inhalte mit dem in einem Knowledge Management System hinterlegten oder generierten Wissen über Produkte, Kunden, deren Profile, Lieferanten uvm. zu intelligenten, smarten Inhalten kombiniert, um so einen Mehrwert aus Information zu schaffen.
Empolis stellt seine bewährten Lösungen auch als Software as a Service (SaaS) zur Verfügung. Alle notwendigen Komponenten der jeweiligen Applikation – ob Datenbank, ausgefeilte Suchverfahren oder spezifische Applikationslogik – sind in die Empolis Smart Cloud ausgelagert und werden in einem hochmodernen Data Center von Empolis betrieben – gemäß den strengen Bestimmungen des deutschen Datenschutzgesetzes.

Einsatz-Szenarien
- Smart Documentation: Effiziente Technische Dokumentation
- Smart Publishing: Intelligentes Publizieren über sämtliche Kanäle

- Experience Management: Systematische Wiederverwendung und Erweiterung des Unternehmenswissens
- Service Resolution Management: Optimale Wissensversorgung des Service Center zur schnellen Problemlösung
- Smart Diagnostics: Effiziente Diagnose von Gerätestörungen und Reparatur
- Competitive Intelligence: Automatisierte Wettbewerbsbeobachtung
- Decision Intelligence: Entscheidungsunterstützung durch umfassendes Wissen

Big-Data-Technologien

Empolis verfügt über mehr als 25 Jahre Erfahrung im Information Management. Ein zentraler Anteil dieser Erfahrung ist die kontinuierliche Produktentwicklung. Das Empolis Content Lifecycle System (CLS) und das Empolis Information Access System (IAS) blicken auf mehr als zwei Jahrzehnte ihrer Versionshistorie zurück. Ein wichtiger Teil dieses Weges führt über rund 20 Verbundvorhaben aus der vorwettbewerblichen Forschung. In diesen Projekten wurden immer wieder neue Ideen und Technologien gemeinsam mit Anwendern, Forschern und auch Wettbewerbern entwickelt, die die Produkte voran gebracht haben – Partnerschaft mit Innovatoren schafft Innovation. Empolis bietet Lösungen, die sowohl technologisch als auch operational über den State-of-the-Art hinausgehen – egal ob es dabei um Skalierbarkeit, semantische Verfahren, Text Mining oder Informationsextraktion geht. Auf der Basis des hochskalierenden IAS verfügt Empolis für eine Vielzahl von Anwendungsfällen über optimal angepasste Analyseverfahren für unstrukturierte Inhalte. Beispielsweise ist Empolis-Technologie in der Lage, einen Tag Traffic auf Twitter in weniger als 20 Minuten oder die deutsche Version der Wikipedia in drei Minuten semantisch zu annotieren und zu verarbeiten.
Neben statistischen Algorithmen umfasst dies auch die massiv-parallele Auswertung mit linguistischen Verfahren zur Informationsextraktion. Diese wiederum bilden die Grundlage für Empolis Smart Information Management® Lösungen, die mit Hilfe der inhaltlichen Analyse die unstrukturierten Inhalte in maschinell verarbeitbare strukturierte Information verwandeln.

11.3 EXASOL

Kontaktinformationen

EXASOL AG
Neumeyerstraße 48 | 90411 Nürnberg
Tel.: +49 (0) 911 23991 0 | info@exasol.com
www.exasol.com

Big-Data-Lösungen

EXASOL ist der Hersteller der relationalen In-Memory-Datenbank EXASolution. Sie wurde speziell für Enterprise-Data-Warehouse-Anwendungen, Big Data und umfangreiche Analytics-Prozesse entwickelt.

Die auf In-Memory-Technologie basierende Datenbank wird für zeitkritische komplexe Analysen großer Datenmengen, umfassende Datenrecherchen, Planungen oder Reportings eingesetzt.

Durch die Integration von Geodaten, Big-Data-Quellen und unstrukturierten Daten eröffnet EXASolution zusätzliche Auswertungsdimensionen, die noch effizientere und Ad-hoc-Analysen zulassen. Die Easy-to-manage-Datenbank lässt sich einfach in bestehende IT-Infrastrukturen integrieren und erfordert geringeren Administrationsaufwand bei niedrigeren Investitions- und Betriebskosten.

Einsatz-Szenarien

Die Hochleistungsdatenbank EXASolution ist in unterschiedlichsten Branchen einsetzbar: im Einzelhandel, für Webanalysen-Anbieter, Versicherungen, E-Commerce-Unternehmen, Telekommunikationskonzerne oder auch im Energiesektor. Alle Organisationen mit einem großen Datenaufkommen können von den Vorteilen einer massiv-parallel arbeitenden In-Memory-Datenbank profitieren. CRM-Auswertungen, strategische Simulationen, Scoring- und Rankingberechnungen, prozessorientierte Datenaufbereitung sowie Real-time – und Click-Stream-Analysen bilden dabei die typischen Anwendungsszenarien. Auch Geoinformationen und unstrukturierte Daten lassen sich schnell auswerten.

Big-Data-Technologien

EXASolution durchbricht mit seinen effizienten Lösungsmodellen aus dem Bereich des High Performance Cluster Computings die bisherigen Leistungsbarrieren eines Data Warehouse. Die einzigartige Kombination der In-Memory-Technologie mit einer Shared-Nothing-Architektur unter Einsatz von innovativen Kompressionsalgorithmen hilft Unternehmen neue analytische Herausforderungen zu meistern. Intelligente Algorithmen verteilen die Daten selbstständig innerhalb eines Clusters und führen automatisch die notwendigen Optimierungsschritte on-the-fly durch.

Die automatische Anpassung des selbstlernenden Systems an Nutzungsgewohnheiten der Anwender und die Verwendung von gängigen Business-Intelligence- und Data-Mining-Anwendungen erhöhen die Akzeptanz bei den Fachabteilungen.

Eigens erstellte Algorithmen, Methoden – u. a. aus den Programmiersprachen R, Python und Lua – oder MapReduce-Algorithmen können zudem flexibel und hochperformant in einem Cluster ausgeführt werden, um beliebige Anforderungen umsetzen zu können. Große Datensammlungen werden so zu wertvollem Unternehmenswissen.

11.4 Experton Group

Kontaktinformationen
Experton Group AG
Carl-Zeiss-Ring 4 | 85737 Ismaning
Tel.: +49 (0) 89 923331-0 | Fax: +49 (0) 89 923331-11
info@experton-group.com
www.experton-group.de

Über Experton Group

Die Experton Group ist das führende, voll integrierte Research-, Advisory- und Consulting-Haus für mittelständische und große Unternehmen, das seine Kunden durch innovative, neutrale und unabhängige Expertenberatung bei der Maximierung des Geschäftsnutzen aus ihren ICT Investitionen maßgeblich unterstützt. Die Experton Group erbringt Marktuntersuchungen, Beratungsleistungen, Assessments, Benchmarking, Konferenzen, Seminare und Publikationen im Umfeld der Informations- und Kommunikationstechnologie. Das Leistungsspektrum umfasst hierbei Technologie, Geschäftsprozesse, Management sowie M&A.

Auch rund um Big Data unterstützt die Experton Group Anbieter und Anwender von Informations- und Kommunikationstechnologien mit Analysen, Workshops und Beratungsdienstleistungen.

Die Experton Group AG wurde am 01. Juli 2005 von sehr erfahrenen Marktforschungs- und Beratungsexperten gegründet. Die Experton Group Gesellschaften arbeiten mit über 80 festen und freien Mitarbeitern zusammen. Diese bringen Erfahrungen aus ihrer Beschäftigungszeit bei IDC, Input, Techconsult, Forrester, Gartner und META Group mit. Der Vorstand der Experton Group AG setzt sich aus Jürgen Brettel (Vorsitzender) und Andreas Zilch zusammen. Der globale Research und die Gesellschaften im Mittleren Osten werden von Luis Praxmarer geleitet. Research Partner der Experton Group sind Experture (USA), Everest Group (USA) und Evalueserve (Indien). Die Experton Group AG hat ihren Sitz in Ismaning und Niederlassungen in Frankfurt, Kassel und St. Gallen/Schweiz.

11.5 Forrester Research

Kontaktinformationen
Forrester Germany GmbH
Eschersheimer Landstraße 10 | 60322 Frankfurt am Main
Tel.: +49 (0)69 959 298 0 | hkisker@forrester.com
www.forrester.com

Über Forrester Research
Forrester Research (Nasdaq: FORR) ist ein globales Beratungs- und Marktforschungsunternehmen, das Marketing- Strategie- und Technologie Management-Experten in 13 strategischen Schlüsselrollen unterstützt. Diese stehen regelmäßig vor komplexen Geschäfts- und Technologie-Entscheidungen durch das sich stark verändernde Verhalten von Kunden, Partnern und Konsumenten. Um die Chancen der Veränderung besser verstehen und strategisch nutzen zu können, stellt Forrester herstellerunabhängige Beratung basierend auf proprietärem Research, Consumer-und Business-Daten sowie Veranstaltungen, Online-Communities und Peer-to-Peer-Executive-Programme zur Verfügung. Dies sichert Entscheidern und Unternehmen heute und in der Zukunft ihren Geschäftserfolg

Big Data Lösungen
Big Data ist ein Schwerpunkt Thema bei Forrester Research. Forrester hilft Unternehmen effiziente Strategien für einen erfolgreichen Einstieg in das Thema zu finden, die richtigen technologischen Entscheidungen zu treffen und existierende Geschäftsprozesse signifikant zu verbessern oder neue, innovative Geschäftsmodelle zu entwickeln.

11.6 Fraunhofer-IAIS

Kontaktinformationen
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven | 53757 Sankt Augustin
bigdata@iais.fraunhofer.de
www.iais.fraunhofer.de/bigdata.html
Ansprechpartner:
Dr. Stefan Rüping, Tel.: +49 (0) 2241 14 3512
Dr. Dirk Hecker, Tel.: +49 (0) 2241 14 1509

Big-Data-Lösungen
Das Fraunhofer IAIS zählt zu den führenden anwendungsorientierten Forschungsinstituten in den Bereichen Data Mining, Maschinelles Lernen, Information Retrieval und Semantische Technologien. Mit unserem umfangreichen Know-how aus Forschung und praktischer Anwendung begleiten wir Ihr Unternehmen auf dem Weg zum optimalen Einsatz von Big Data Analytics.
Wir identifizieren das Potenzial von Big-Data in Ihrem Unternehmen so detailliert, dass Sie anschließend direkt mit konkreten Projekten starten können. Gemeinsam mit Ihren Fachabteilungen entwickeln wir neue Nutzungs- und Geschäftsideen für Ihre Datenbestände. Dazu analysieren wir Ihre Prozesse und ermitteln genutzte und ungenutzte Daten. Ergänzend liefern wir einen Überblick über am Markt verfügbare und offene Datenquellen.
Mit unseren Schulungen und Coachings für Data Scientists helfen wir Ihnen beim Aufbau von Know-how. Wir begleiten Sie bei der konzeptionellen Ausarbeitung und technischen Umsetzung Ihrer Ideen. Machbarkeit und Wirtschaftlichkeit sind uns gleichermaßen wichtig wie Datenschutz und Sicherheit.

Einsatzszenarien
Wir führen Projekte in verschiedenen Branchen und Unternehmensbereichen durch:

- Automatisierte Erkennung von Betrugsmustern in Kreditkartentransaktionen und Abrechnungsdaten
- Prognosen für Absatz-, Sortimentsplanung und Personaleinsatzplanung im Handel
- Analyse von Kundendaten für die individuelle Ansprache
- Überwachung von Social- und Online-Medien für die Marktforschung
- Analyse von Mobilitätsdaten für die Standortplanung und -bewertung
- Anonymisierung und Erschließung von Mobilfunkdaten für neue Geschäftsmodelle
- Qualitätssicherung und Ursachenanalyse von Funkabbrüchen in Telekommunikationsnetzen
- Präventive Wartung und Überwachung von vernetzten Geräten
- Kurzfristige Bedarfsprognosen und vorausschauende Steuerung in der Energiebranche
- Prognosen für die logistische Planung
- Analyse von medizinischen Daten für die individuelle Behandlung
- Integration von Unternehmensdaten für Corporate Intelligence

Big-Data-Technologien
Wir setzen hochleistungsfähige Verfahren des maschinellen Lernens und semantische Technologien in skalierbaren Big-Data-Architekturen ein:

- Machine Learning und Data Mining: Datenvorverarbeitung, Modellentwicklung und -validierung, Lernverfahren auf Datenströmen in Echtzeit, Privacy Preserving Data Mining
- Semantic Technologies: Inhaltliche Strukturierung von großen Dokumentenmengen, semantische Klassifikation, Informationsextraktion, Textanalyse, Linked Data, Knowledge Representation
- Mobility Analytics: Analyse von Trajektorien, Ermittlung von Kontakt und Besuchsfrequenzen, Ereignissen und Wegemustern in Mobilitätsdaten, datenschutzkonforme Methoden
- Multimedia Pattern Recognition: Dokumentanalyse, Sprach- und Audioanalyse, Bild- und Videoanalyse
- Visual Analytics: Interaktive visuelle Exploration, visuelles Debugging und visuelles Reporting für komplexe Datenbestände
- Big Data Architecture: Skalierbare Systeme, parallele Workflows, MapReduce, No-SQL-Datenbanken, Datenströme.

11.7 Fujitsu

Kontaktinformationen
FUJITSU Technology Solutions GmbH
Mies-van-der-Rohe-Strasse 8 | 80807 München
Tel.: 01805 372 100 | cic@ts.fujitsu.com
www.fujitsu.com/de/about/local/contacts-de/index.html
www.fujitsu.com/fts/solutions/high-tech/bigdata/

Über Fujitsu
Fujitsu ist der führende japanische Anbieter von Informations- und Telekommunikations-basierten (ITK) Geschäftslösungen und bietet eine breite Palette an Technologieprodukten, -lösungen und -Dienstleistungen. Mit über 170.000 Mitarbeitern betreut das Unternehmen Kunden in mehr als 100 Ländern. Fujitsu nutzt seine ITK-Expertise, um die Zukunft der Gesellschaft gemeinsam mit ihren Kunden zu gestalten. Im Geschäftsjahr 2012 (zum 31. März 2013) erzielte Fujitsu Limited (TSE:6702) mit Hauptsitz in Tokio, Japan, einen konsolidierten Jahresumsatz von 4,4 Billionen Yen (47 Milliarden US-Dollar). Weitere Informationen unter http://www.fujitsu.com/de/

Big-Data-Lösungen
Fujitsu unterstützt sämtliche für Big Data relevante Infrastrukturkonzepte: Hadoop Cluster mit Fujitsu-spezifischen Erweiterungen, eine CEP-Engine, IMDB-Appliances auf Basis SAP HANA und integrierte IMDG-Lösungen auf Basis Terracotta BigMemory. Somit kann abhängig von Situation und Anforderungen stets der geeignete Technologiemix eingesetzt werden, um für den Kunden die optimale Lösung zu finden. Die für die Infrastruktur erforderlichen Produkte, wie Server, Speichersysteme, Netzkomponenten und Zugangsgeräte werden ebenfalls von Fujitsu bereitgestellt. Big-Data-Infrastrukturen von Fujitsu beinhalten Software und Middleware von Fujitsu selbst, aus der Open-Source-Welt und von Partnern wie SAP oder Software AG. Sie sind aber auch offen für Produkte führender ISVs. Ebenso wichtig wie Produkte und Infrastrukturkonzepte ist das Serviceangebot von der Prozess- und Infrastrukturberatung, über das Design der optimalen Infrastruktur, die Implementierung, die Integration in die bestehende IT-Landschaft bis hin zum ganzheitlichen Support. Für all diese Leistungen werden auch attraktive Finanzierungsoptionen angeboten. Darüber hinaus ist Fujitsu bekannt für seine flexiblen Sourcing-Modelle. Kundenspezifische Lösungen können vom Kunden entweder in Eigenregie oder von Fujitsu betrieben und verwaltet werden. Big-Data-Services werden auch aus der Fujitsu-Cloud bereitgestellt.

Einsatz-Szenarien
Fujitsu bietet Big-Data-Infrastrukturlösungen, die unabhängig von Branche und Unternehmensgröße eingesetzt werden können. Beispiele von erfolgreich realisierten Projekten sind bei Energieversorgern die frühzeitige Erkennung und Überbrückung von Versorgungsengpässen. Im Einzelhandel wurden Lösungen zur Optimierung der Anordnung der Artikel in den Regalen umgesetzt, die den Verkaufserfolg steigern. Im öffentlichen Sektor werden Big-Data-Lösungen zur Verbrechensbekämpfung eingesetzt, in der Landwirtschaft bieten unsere Lösungen Unterstützung zur verbesserten Erntemenge und -qualität.

Big-Data-Technologien
Fujitsu bietet für Big Data die komplette Bandbreite der Technologien – z. B. verteilte Parallelverarbeitung, Complex Event Processing, In-Memory Datenbanken, In-Memory Data Grid – und sorgt mit Consulting und Services für kundenoptimierte Umsetzungen.

11.8 Graf von Westphalen

Kontaktinformationen
Graf von Westphalen Rechtsanwälte Steuerberater Partnerschaft
Ansprechpartner: Arnd Böken, Partner
Potsdamer Platz 8 | 10117 Berlin
Tel.: +49 (0) 30 726111-0 | Fax: +49 (0) 30 726111-333
a.boeken@gvw.com
www.gvw.com

Kanzlei
Graf von Westphalen ist eine Partnerschaft von Rechtsanwälten mit derzeit mehr als 130 Rechtsanwälten und Büros in Berlin, Düsseldorf, Frankfurt, Hamburg und München sowie in Shanghai, Istanbul, Brüssel und Alicante. Als Full-Service-Kanzlei bietet Graf von Westphalen Beratung aus einer Hand in allen Bereichen des Wirtschaftsrechts.

Tätigkeitsprofil
Im Bereich des IT- und Datenschutzrechts berät Graf von Westphalen mit einem Team um die Partner Arnd Böken (empfohlener Anwalt für IT, Legal 500 Deutschland, 2014) und Stephan Menzemer (Leading IT-Lawyer laut JUVE-Handbuch Wirtschaftskanzleien, Legal 500 Deutschland) umfassend, national und international, in folgenden Bereichen:

- Gestaltung von IT-Verträgen für Anbieter und Kunden: Hardware- und Software-Beschaffung und -Management, -Erstellung und -Überlassung (Lizenzen, Urheber-rechte, Nutzung und Verwertung), Pflege und Wartung, Hosting, IT-Vergabe
- Einführung von Softwareanwendungen in Unternehmen und Behörden einschließlich Datenschutz und IT Compliance
- Komplexe und umfangreiche IT-Projekte wie Outsourcing, Cloud Computing (ein-schließlich Gestaltung von SaaS-Projekten), Softwareeinführungen und Roll Outs, Supplier-Management
- Datenschutz und IT-Compliance und IT-Security im Unternehmen: Datentransfer in-nerhalb von Konzernen, Auftragsdatenverarbeitung, rechtliche Aspekte von IT-Schutzrichtlinien, Bring your own device etc.
- Internetrecht: Internetauftritt und Webmarketing/ Analysetools, E- und M-Commerce (inkl. Mobile Apps), affiliate-Marketing
- IT-Litigation und -Mediation

Big-Data-Anwendungen
Graf von Westphalen berät im Bereich Big Data, Schwerpunkt sind folgende Tätigkeiten:

- Prüfung von Big Data Anwendungen vor und während der Konzeption hinsichtlich Datenschutzrecht und anderen rechtlichen Anforderungen
- Beratung bei Einführung und Weiterentwicklung von Big-Data-Lösungen im Unternehmen
- Gestaltung und Verhandlung von IT-Verträgen zu Big-Data-Projekten einschließlich Rechtsfragen Open Source (Apache License 2.0 etc.)
- Rechtsgutachten zu Datenschutz und anderen Rechtsfragen im Zusammenhang mit Big Data
- Schulungen zu Datenschutzrecht bei Big Data. Vorträge zum Thema: Arnd Böken: »Was ist erlaubt? Leitlinien für den Datenschutz bei Big Data Projekten«, Big Data Summit, 2013; »Big Data Analytics rechtssicher einsetzen«, Big Data Summit 2014, »Hadoop-Verträge«, BITKOM Arbeitskreis Big Data, Januar 2014.

11.9 Hewlett-Packard

Kontaktinformationen
Hewlett-Packard GmbH
Herrenberger Str. 140 | 71034 Böblingen
Tel.: +49 (0) 7031 – 14-0 | firmen.kunden@hp.com
www.hp.com/de

Big-Data-Lösungen
HP bietet den Kunden ein komplettes End-to-End Portfolio inklusive Services zur Definition, Beratung, Implementierung und Betrieb von kompletten Big-Data-Lösungen als auch Hardware, Software und Appliances an.
HP Big-Data-Services beinhalten die Definition der Big Data Use Cases, die Lösungskonzeption, die Definition der Kernparameter, die Pilotierung der Big-Data-Lösung, die Implementierung und die Überführung vom Pilotprojekt in den Produktivbetrieb – entweder beim Kunden oder über ein Cloud Modell. Über den HP Big Data Discovery Experience (HP BDDE)-Service bietet HP den Unternehmen einen äußerst attraktiven Einstieg für die Verprobung von Big-Data-Anwendungsfällen an. Über einen sofort verfügbaren Pilotierungsservice lässt sich testen, ob eine Investition in Big Data zur Steigerung des Unternehmensumsatzes aussichtsreich ist. Die durchschnittliche Dauer einer Pilotimplementierung ist 12 Wochen. Falls sich Anwendungsfälle erhärten lassen, können diese extrem schnell in die Produktion umgesetzt werden. Test- und Produktionsumgebung haben denselben Aufbau. Weiterhin gehören vordefinierte Services wie z. B. »HP Information Fabric for Risk, Compliance, and Insight« , eine Kombination aus Hardware, Software und Service für das End-to-End Management von Informationen jeglicher Datentypen (strukturiert, unstrukturiert) im Unternehmen zum Angebot von HP.

Einsatz-Szenarien
HP unterstützt Kunden bei der Definition der Big Data Use Cases und bei der Bewertung des Nutzenaspektes der Big-Data-Anwendungsfälle für das Business. HP's Angebot unterstützt Unternehmen in Industrien jeglicher Art. Beispiele sind u.a. Warranty Analytics, Analysen über das Reiseverhalten, Big Data Analytics für die Landwirtschaft, Brand Awareness, Fanverhalten bei Sportveranstaltungen. Die Big-Data-Innovationen gehen auch einher mit anderen Innovationen von HP – beispielsweise der von HP LABs entwickelten Sensoren-Technologie (HP CeNSE – Central Nervous System of the Earth) für den Aufbau eines Sensoren-Netzwerks. Dieses misst mit extremer Empfindlichkeit seismologische Daten bzw. Erschütterungen und wird u.a. in der Ölindustrie für die bessere Datenerfassung- und Bewertung der Ölreserven bzw. möglicher Ölvorkommen genutzt.
Weiterer Anwendungsfall: Ein Netzwerk von Sensoren kann Daten über Erschütterungen an Straßen oder großer Brücken liefern. Eine Million Sensoren, die 24 Stunden am Tag laufen, liefern mehr als 20 Petabytes Daten innerhalb von 6 Monaten, die es bei Bedarf schnell auszuwerten und sinnvoll einzusetzen gilt. Wofür die HP Big-Data-Analytics-Lösungen prädestiniert sind.

Big-Data-Technologien
Die HP Big Data Analytics Plattform (HAVEn) für die Erfassung, Bewertung (Meaning Based Computing), Analyse und die jeglicher Datentypen – ob strukturiert, unstrukturiert, Video, Audio, Fotos, Text, Daten aus sozialen Netzen etc. können verarbeitet werden. Die Big Data Analytics Platform besteht aus einer vorintegrierten Kombination aus Hadoop, Vertica, Autonomy und HP Security Technologien (Arcsight Logger). Mehr als 700 Adapter stehen für die Einspeisung der Daten in die Big-Data-Analytics-Plattform zur Verfügung. Basierend auf der Plattform werden Zusatzlösungen und Dienstleistungen von HP und HP Partnern angeboten, z. B. HP BDDE oder HP Operations Analytics zur Analyse IT-spezifischer Daten. Die Softwarelösungen von HP Vertica und HP Autonomy für die Analyse und schnelle Verarbeitung großer Datenmengen werden auch als Einzellösungen angeboten.
HP Hardware und Applicances: Zum HP-Hardware- und Applicances-Angebot für den Betrieb von Big-Data-Umgebungen gehören Server, Speichersysteme, Netzwerke inklusive Systemen für die In-Memory-Verarbeitung als auch vordefinierte Lösungen wie z. B. HP Appliance Systems for Hadoop, SAP HANA oder Microsoft PDW.

11.10 Hortonworks

Kontaktinformationen
Hortonworks
Maximilianstrasse 35A | 80539 München
Tel.: +49 89 24218 0 | fniedermaier@hortonworks.com
www.hortonworks.com

Big-Data-Lösung
Hortonworks Data Platform (HDP) ermöglicht es Unternehmen, Daten in jedem beliebigen Format und in voller Größe kostengünstig zu speichern und auf vielfältige Weise zu verarbeiten. HDP ist die einzige 100%ige Open-Source-Distribution von Apache Hadoop im Markt. Somit vermeiden Kunden jedwede Bindung an einen Hersteller. Alle Innovationen von Hortonworks werden zu 100% als Open Source Software bereitgestellt. Unsere Roadmap ist jederzeit unter http://hortonworks.com/labs öffentlich. Hortonworks treibt die Entwicklung von Apache Hadoop ganz wesentlich. Mehr als die Hälfte des Codes der verschiedenen Apache Hadoop Module stammt von Hortonworks. Das Kernteam bei Hortonworks umfasst die ursprünglichen Hadoop Architekten und Entwickler von Yahoo. Hortonworks beschäftigt zusammen mit unserem Investor Yahoo mehr als die Hälfte aller Apache Hadoop Committer. Die Hortonworks Data Platform ist die stabilste und zuverlässigste Distribution von Apache Hadoop auf dem Markt. Jedes Release wird getestet und zertifiziert auf über 45.000 Servern im Wirkbetrieb bei Yahoo. Hortonworks bietet Kunden Schulungen, Beratung sowie Support für Hadoop.

Einsatz-Szenarien:
Hortonworks Data Platform (HDP) findet sowohl für strukturierte als auch für unstrukturierte Daten Anwendung, sowohl für Stapelverabeitung als auch für interaktive Analysen. Häufig benutzen Kunden HDP im Zusammenhang mit neuen Datentypen wie zum Beispiel Clickstream-Daten, Social Media Stimmungsdaten, Server-Logdaten, Sensordaten, Maschine-zu-Maschine-Kommunikation, Standortdaten und Freitext-Daten.

Technologien:
Hortonworks unterstützt alle in diesem Leitfaden erwähnten Apache Hadoop Module. Besonders stolz sind wir darauf, Innovationen in den Bereichen zu treiben, die es Kunden ermöglichen, Hadoop als unternehmensweiten Shared Service einzusetzen. Hierzu zählen insbesondere:

- Yarn als »Betriebssystem« für Hadoop, mit dem verschiedenste analytische Anwendunge gleichzeitig laufen können
- Ambari als Betriebsmodell, womit Hadoop nativ oder aus Tools wie Microsoft Systems Center provisioniert und gesteuert werden kann
- Knox für Sicherheit, gerade auch in Multimandanten-fähigen Umgebungen
- Tez für interaktive Verarbeitung
- Storm für die schnelle Verarbeitung von großen Datenströmen in Echtzeit

Dank unserer engen technischen Integration mit Microsoft, Teradata, SAP, Rackspace, SAS und anderen Herstellern eignet sich die Hortonworks Data Platform gut für den Einsatz zusammen mit existierenden Lösungen. Hiermit werden bestehende Investitionen weiter genutzt.

11.11 IBM

Kontaktinformationen
IBM Deutschland GmbH
IBM-Allee 1 | 71139 Ehningen
www.ibm.com
Ansprechpartner:
Dr. Mark Mattingley-Scott, scott@de.ibm.com

Big-Data-Lösungen
IBM hat eine unternehmensorientierte Big-Data-Plattform entwickelt, auf der Sie das gesamte Spektrum der geschäftlichen Herausforderungen im Bereich Big Data in Angriff nehmen können. Die Plattform vereint konventionelle Technologien, die sich gut für strukturierte Routineaufgaben eignen, mit neuen Technologien, die auf hohe Geschwindigkeit und Flexibilität ausgerichtet sind und sich hervorragend für die Ad-hoc-Datenuntersuchung, -erkennung und die unstrukturierte Analyse anbieten.
Die integrierte IBM-Plattform umfasst vier zentrale Funktionen:

- Hadoop-basierte Analyse,
- Stream-Computing,
- Data Warehousing sowie
- Informationsintegration und –governance

sowie unterstützende Plattformservices wie

- Visualisierung und Erkennung zur Untersuchung großer, komplexer Datasets
- Anwendungsentwicklung
- Systemmanagement: Überwachung und Management von Big-Data-Systemen, um sichere, optimierte Leistung zu erreichen
- Beschleuniger: Schnellere Wertschöpfung mit analyseorientierten und branchenspezifischen Modulen.

Einsatzszenarien
Die IBM Big-Data-Plattform und ihre Komponenten werden heute zur Umsetzung von innovativen, analytischen Lösungen in Industrie- und Cross-Industrie Use Cases eingesetzt. Das Lösungsspektrum deckt die komplette Bandbreite der Verarbeitung und Analyse von strukturierten, semi- und unstrukturierten Daten ab. Es werden damit Anwendungsbereiche adressiert wie Netzwerk-Analysen, Betrugserkennung, Security Intelligence, Informations-Discovery, Suche und intelligente Daten-Exploration, Real-time Event- und Datenstrom-Analysen, Social Media Analysen, DWH Offload /Archiv-Szenarien im Data-Warehouse-Umfeld, Video Surveillance, Smart-Grid-/Smart-Meter- Steuerung und Optimierung sowie in IT-zentrischen Szenarien im Bereich der Log und Cloud Analytics.

Big-Data-Technologien
IBM BigInsights ist die Enterprise-fähige Hadoop-Distribution incl. integrierter Analyse-Bausteine zur Text-Analyse, Machine Learning und für statistische Analysen.
IBM InfoSphere Streams ist die Real-time Streams Computing Plattform incl. seiner integrierten, analytischen Funktionen.
SPSS bietet Predictive Analytics Funktionen für Streaming, Hadoop und DWH-Umgebungen an.
Mit DB2 BLU verfügt IBM über eine hoch-skalierbare In-Memory DB als Erweiterung der DB2 Plattform.
Der InfoSphere Data Explorer ist eine Lösung zur effizienten Datenexploration und semantischen Suche.
IBM Cognos steht für Business Analytics- und Performance Management-Lösungen, mit denen Unternehmen ihre Geschäftsleistung planen, überwachen, analysieren und besser steuern können.
Die Puredata-System-Familie bietet optimierte Appliances für Hadoop, DWH-basierte Umgebungen und Analytics.

11.12 Microsoft

Kontaktinformationen
Microsoft GmbH
Konrad-Zuse-Straße 1 | 85716 München
Tel.: +49 (0) 89 3176 0 | hans.wieser@microsoft.com
www.microsoft.com/bigdata

Big-Data-Lösungen
Microsoft schöpft aus der eigenen Erfahrung als Betreiber von Big-Data-Plattformen wie Skype, Bing und Xbox. Die integrierte Plattform vereinfacht den Einsatz von Big Data und die intuitive Erstellung fundierter Prognosen in Echtzeit.

- Mit Power BI für Office 365 können Anwender mit vertrauten Werkzeugen wie Excel Daten aus eigenem Bestand oder der Cloud intuitiv erfassen, visualisieren und verteilen, Abfragen in natürlicher Sprache formulieren und Echtzeitdaten auf jedem mobilen Endgerät nutzen.
- Parallel Data Warehouse erlaubt die parallele Verarbeitung massiver Datenbestände in modernen Echtzeit-ROLAP Szenarien. Dabei bindet Polybase transparent verteilte Hadoop-Daten in die SQL-Abfragen ein.
- Mit Azure und speziell dem HD Insight Service können Anwender kostengünstig in Minutenschnelle massive Rechenkapazität für Hadoop nutzen. Die einfache Nutzung ermöglicht die Beschleunigung von Innovationszyklen zum Beispiel bei Entwicklung und Test neuer Produkte.
- Über den Azure Marketplace bietet Microsoft darüber hinaus Anwendern die Möglichkeit, ihre Daten mit frei verfügbaren und kommerziellen Datenquellen anzureichern

Jede der genannten Lösungen harmoniert mit dem Microsoft Produktportfolio, ergänzt aber auch ideal Ihre organisch gewachsene IT in Ihrem Rechenzentrum oder in der Cloud.

Einsatz-Szenarien
Das umfassende Lösungsportfolio ermöglicht es Microsoft, jeden der in diesem Leitfaden beschriebenen Anwendungsfälle für Big Data zu adressieren, z. B.

- Erfassung von Echtzeitdaten, zum Beispiel in Industrieanlagen zu vorausschauenden Wartung,
- Verdichtung und Verknüpfung in hybriden Data Warehouses, zum Beispiel zur Historisierung von Daten und zur Bedienung regulativer Anforderungen,
- bis zur kooperativen Visualisierung und Mustererkennung sowie Verdichtung von Daten, z. B. zur Analyse sozialer Netzwerke.

Eine Auswahl von Anwendungsfällen finden Sie unter www.microsoft.de/sql bzw. www.microsoft.com/de-de/server/sql-server/2012/kundenreferenzen.aspx

Big-Data-Technologien
Mit dem SQL Server verfügt Microsoft über die meistverbreitete relationale Datenbank am Markt. Die aktuelle Version 2014 beschleunigt durch die Nutzung von In-Memory-Technologie, Spaltenspeicherung, Kompression und innovativen Lockverfahren sowohl schreibende als auch lesende Zugriffe um das 10-100fache auf der gleichen Hardware. SQL Server beinhaltet bereits alle klassischen Data-Warehouse-Technologien für ETL, Datenqualität, Berichtswesen, OLAP.
Die MPP-Technologie des Parallel Data Warehouse (PDW) richtet sich an Anwendungsfälle mit extrem hohen Datenmengen oder geringer Toleranz für Antwortzeiten. PDW bietet gleichzeitig die Möglichkeit, Hadoop Knoten im eigenen Rechenzentrum zu betreiben, zum Beispiel für sensible Daten.
Power BI ergänzt die Plattform um intuitive analytische Funktionen durch die Module Power Query zur Exploration von Daten, Power Pivot zur Verknüpfung und Modellierung sowie Power View und Power Map für Visualisierung und Interpretation.
Azure bietet eine flexible, elastische Plattform zur Bereitstellung von Rechenleistung und Speicherkapazität für jede der beschriebenen Technologien, permanent, saisonal oder im K-Fall.

11.13 SAP

Kontaktinformationen
SAP Deutschland AG & Co. KG
Hasso-Plattner-Ring 7 | 69190 Walldorf
Tel.: +49 (0) 6227 7-77206 | mark.von.kopp@sap.com
www.sapbigdata.com

Big-Data-Lösungen
SAP kombiniert neue und innovative Big Data Geschäftsanwendungen mit Echtzeit Analysen und fundierten Vorhersagen. Dies basiert auf einer offenen und voll integrierten führenden technischen Plattform, der SAP HANA Plattform für Big Data. Flankierend und ergänzend bietet SAP dazu ein komplettes Big Data Lösungsportfolio:

SAP HANA stellt eine wegweisende Echtzeitplattform für Analytik und Anwendungen dar. Während der IT-Bereich vereinfacht wird, stellt die Plattform leistungsstarke Funktionen bereit. Dazu gehören eine deutlich verbesserte Verarbeitungsgeschwindigkeit, die mögliche Verarbeitung großer Datenmengen sowie Prognose- und Text-Mining-Funktionen. Diese Echtzeitplattform kombiniert hohe Transaktionsvolumina mit Analysefunktionen, um so Lösungen zu schaffen, die Ihre Unternehmensleistung erhöht.

Bei der Lösung SAP Predictive Analysis handelt es sich um ein Tool für die statistische Analyse und das Data Mining, mit dem sich Vorhersagemodelle erstellen lassen, durch die Sie verborgene Einsichten gewinnen und Abhängigkeiten in Ihren Daten erkennen und so möglicherweise Voraussagen über zukünftige Ereignisse machen können.

Zudem bietet der SAP Event Stream Processor (SAP ESP) kontinuierliche Business Intelligence für eine schnelle und intelligente Entscheidungsfindung und Geschäftsführung. Auch wenn die Flut eingehender Daten manchmal überwältigend ist – SAP Event Stream Processor verwandelt diese auf jedes Unternehmen einströmenden Geschäftsmitteilungen in aussagekräftige Informationen. Dabei werden irrelevante Daten verworfen, Daten aus verschiedenen Quellen kombiniert und Ereignisse im Kontext anderer Ereignisse untersucht, um festzustellen, was wichtig ist. Diese Aufgaben werden bei sehr hohen Geschwindigkeiten, in Echtzeit und für große Mengen eingehender Daten durchgeführt. So können Unternehmen schneller auf sich ändernde Bedingungen reagieren, Bedrohungen und Opportunities erkennen, sobald sie auftauchen und fundiertere Entscheidungen durch umfassende und zeitnahe Informationen treffen.

SAP Data Services bietet für Sie abschließend eine verlässliche Informationsgrundlage, um operative und analytische datenbezogene Vorhaben zu unterstützen. Sie erzielen wesentliche Vorteile durch folgende Möglichkeiten:

- Schöpfen Sie das Potenzial Ihrer Daten voll aus, indem Sie unabhängig von Datentyp, Datendomäne oder Datenquelle den Zugriff auf entscheidende Daten für alle wichtigen Geschäftsprozesse ermöglichen.
- Stellen Sie möglichst verlässliche Informationen mit höherer Genauigkeit und Zuverlässigkeit der Daten für Entscheidungs- und Geschäftsprozesse bereit.
- Optimieren Sie die Betriebsabläufe und senken Sie die Gesamtbetriebskosten, indem Sie nur eine Anwendung für mehrere Datenverwaltungsprozesse unterhalten.

Einsatz-Szenarien
Modernste Geschäftsanwendungen wie die SAP Customer Engagement Intelligence Lösungen helfen Ihnen, Marketing und Vertrieb über alle Kanäle zu konsolidieren, zu optimieren und in Echtzeit zu überwachen und zu steuern. Dabei werden insbesondere die in Zukunft stark wachsenden Daten aus sozialen Medien, aus Kassensysteme, aus dem Web und aus Geo-Daten auf einzigartige Weise mit intuitiven Benutzeroberflächen und schlanken sowie auch mobil nutzbaren Transaktionen kombiniert.

Nutzen Sie die neuen Möglichkeiten von Big Data auch im Bereich Betrugserkenung und –vermeidung mit SAP Fraud

Management. Erkennen Sie Betrug, decken ihn auf, und nutzen Sie Muster und Korrelationen aus den verschiedenen Datenquellen, um in Zukunft Betrug zu verhindern.

Mit der SAP Big Data Lösung SAP Demand Signal Management werden Sie zudem ein wirklich angebots- und nachfrageorientiertes Unternehmen. Erfassen Sie externe Markt- und Verbrauchsdaten und kombinieren Sie diese mit Ihren internen Daten und Analysen, um in Echtzeit auf Ihre Lieferkette einzuwirken und diese mit prädikativen Analysen in vorher nicht gekanntem Ausmaße zu steuern.

Durch die neuartige Kombination von Produktionsdaten mit zahlreichen weiteren Geschäftsdaten aus den Bereichen Controlling, Finanzen und Service haben Sie über die SAP Operational Equipment Effectiveness Lösung nun die Möglichkeit, Gründe für mangelnde Anlagenauslastung, Qualitätsschwankungen und Produktionsfehler frühzeitig zu erkennen und zu analysieren. Damit erreichen Sie eine höhere Produkt- und Produktionsqualität und eine bessere Nutzung Ihrer Produktionsanlagen.

Über die herausragenden Eigenschaften der SAP HANA Plattform können Sie zudem die intelligente Auflösung von Daten und Informationssilos erreichen sowie in Echtzeit externe Daten wie Maschinendaten berechnen, filtern und analysieren.

Die umfangreichen Möglichkeiten zur Entwicklung eigener Anwendungen auf der SAP HANA Plattform und die Bereitstellung von professionalen Beratungsdienstleistungen, die von der Datenmodellierung bis hin zur Entwicklung neuer Programme reichen, helfen Ihnen zudem, individuelle und auf Ihre Bedürfnisse zugeschnittene Lösungen zu entwickeln, die Ihnen auch in Zukunft einen Vorsprung im Wettbewerb gewährleisten.

Big-Data-Technologien
Das SAP Big Data Portfolio verknüpft die innovative SAP HANA Plattform mit den modernsten Werkzeugen zu Daten-Sicherheit, zur Daten-Integration, zur Visualisierung und zur Echtzeit Einbettung von Daten und Analysen in Geschäftsprozesse- und -anwendungen.

SAP Lumira und die SAP Business Objects BI Suite stellen konsistente und für die verschiedenen Zielgruppen angepasste Werkzeuge für die Anzeige und Auswertung von Big Data zur Verfügung.

Der SAP Event Stream Processor ermöglicht ein Bearbeiten und Bewerten von Ereignissen aus Maschinendaten, sozialen Medien und weiteren Sensoren in Echtzeit.

SAP Data Services stellen neben der Integration von Daten aus Umsystemen auch ein zuverlässiges Datenqualitätsmanagement und Informationsmanagement sicher.

Ihren Erfolg stellen wir zudem neben der Integration und Kombination von frei am Markt verfügbarer und erprobter Open Source Software für statistische Berechnungen und wirtschaftliche Speicherlösungen von sehr großen Datenmengen zudem über ein weltweit agierendes und führendes Partnernetzwerk sicher.

SAP bietet somit eine komplettes Portfolio an Technologien für Ihren Geschäftserfolg im Umgang und der Nutzung von Big Data: Sichere und zertifizierte Hardware und Infrastruktur der zahlreichen Technologiepartner, eine hauptspeicher-basierte offene technologische Plattform, Integration verschiedenster Datenquellen und Datentypen, modernste analytische Bibliotheken, intuitive Benutzeroberflächen und Werkzeuge zur Darstellung und Analyse bis hin zu konkreten Anwendungen für die unterschiedlichsten Fachbereiche und Industrien.

11.14 SAS

Kontaktinformationen
SAS Institute GmbH
In der Neckarhelle 162 | 69118 Heidelberg
Tel.: +49 (0) 6221 415 – 123
info@ger.sas.com
Ansprechpartner:
Dr. Dirk Mahnkopf

Unternehmensinformationen
SAS ist Marktführer bei Business-Analytics-Software und der weltweit größte unabhängige Anbieter im Business-Intelligence-Markt. Nach einer aktuellen Studie des Marktforschungsinstituts Lünendonk steht SAS auch in Deutschland deutlich an erster Stelle des Business Intelligence-Marktes.

Der weltweite Umsatz von SAS lag im Jahr 2012 bei 2,87 Milliarden US-Dollar, in Deutschland konnte SAS einen Umsatz von 134,6 Millionen Euro verzeichnen. An über 60.000 Standorten in 135 Ländern wird die SAS Software eingesetzt – darunter in 90 der Top-100 der Fortune-500-Unternehmen. 25 Prozent seines Jahresumsatzes hat SAS letztes Jahr in Forschung und Entwicklung investiert.

SAS beschäftigt weltweit ca. 13.400 Mitarbeiter, in Deutschland sind 550 Mitarbeiter tätig. Die deutsche Niederlassung wurde 1982 in Heidelberg gegründet. Für die optimale Betreuung der Kunden in Deutschland befinden sich weitere regionale Standorte in Berlin, Frankfurt am Main, Hamburg, Köln und München. Die internationale Zentrale des Unternehmens befindet sich in Cary, North Carolina (USA).

Einsatz-Szenarien
Die Softwarelösungen von SAS unterstützen Unternehmen, aus ihren vielfältigen Geschäftsdaten eine konkrete Informationsbasis für strategische Entscheidungen zu gewinnen. In Zusammenarbeit mit seinen Kunden und aus den langjährigen Projekterfahrungen hat SAS Softwarelösungen für eine integrierte Unternehmenssteuerung entwickelt. Diese Lösungen kommen im Bereich Kundenbeziehungsmanagement, Risikosteuerung, strategisches Personalmanagement, Finanzcontrolling und IT-Gesamtsteuerung erfolgreich zum Einsatz.

Big-Data-Technologien
SAS Schlüsseltechnologien unterstützen Kunden bei Big Data Projekten.

- Datenmanagement. SAS bietet umfassende Integrations- und Management-Funktionalitäten zu Hadoop und anderen Big-Data-Datenbanken. Für SAS ist Big Data darüber hinaus mehr als eine Diskussion in Verbindung mit Technologien wie Hadoop, NoSQL usw. SAS arbeitet mit einem umfassenderen Ansatz für Datenmanagement/Data Governance und bietet eine Strategie und Lösungen an, mit denen beliebige Datenmengen effektiv verwaltet und genutzt werden können.
- High-Performance Analytics. Datenanalyse mit Hilfe leistungsfähiger mathematisch-statistischer Verfahren gehört seit je her zu den besonderen Stärken von SAS.
- High-Performance Datenvisualisierung. Die Kombination aus Datenvisualisierung und In-Memory-Verarbeitung mit Features wie Prognosen on-the-fly und Szenarioanalysen, automatischer Diagrammerstellung und Bedienung per Drag-and-Drop ermöglicht einen intuitiven Zugang zu den Daten.
- Flexible Bereitstellungsoptionen. Nutzen Sie SAS Lösungen in Ihrer bestehenden IT-Infrastruktur oder als Service aus der Cloud.

Bei der Konzeptionierung, Entwicklung, Implementierung und Schulung leisten die SAS Professional Services Unterstützung. Sie vereinen Consulting, Customer Support und Education unter einem gemeinsamen Dach und stehen Kunden in allen Phasen ihrer Projekte zur Seite: vom Start-up-Gespräch über die Beratung für das konkrete Projekt bis hin zur Softwareimplementierung sowie Schulung und SAS Zertifizierung der Mitarbeiter.

11.15 SEMANTIS

SEMANTIS®

Kontaktinformationen
SEMANTIS GmbH
Oliver Roser
Postfach 120548 | 69067 Heidelberg
Tel.: +49 6221 6560484 | ro@semantis.de
www.semantis.de

Big-Data-Lösung(en)
SEMANTIS stellt Big Data Analysis in den Fokus, indem vorhandene Ressourcen und neue Lösungen effizient gekoppelt werden.

Besser und schneller zu Entscheidungen finden? Wir beraten Sie, bauen Ihre Kompetenz im Bereich visuelle Datenanalyse auf und realisieren Projekte.

Wir nutzen für

- Visualisierung, Analyse und Präsentation: Tableau Software Produkte.
- Datenbestände Ihrer SAP-Anwendung (ECC/ERP) oder Ihren individuell via Akka angereicherten Big Data Datenbestände: SAP HANA In-Memory-Datenbank.
- Individuelle Lösungen Akka & Scala.

SEMANTIS bietet hierfür u. a.

- als einziger Partner in Europa Geodatenbanken für Tableau zur Visualisierung Ihrer Daten in der Karte an (z. B. Nielsen-Bezirke, PLZ- Gebiete, Baublöcke und viele andere).
- als Partner von Typesafe Inc. die Typesafe Subscription an, welche Entwicklung und Produktivbetrieb abdeckt, sowie Schulungen und Beratung rund um die Reactive Platform.

Einsatz-Szenarien
Tableau Software ist die Lösung, wenn es um schnellstmögliche visuelle Analyse und kollaborative Entscheidungsfindung geht. Angereichert mit Kartenmaterial von SEMANTIS sind Sie in der Lage alle Facetten von Tableau Software zu nutzen. Kunden von Tableau Software sind u. a. Siemens Energy Sector (Deutschland), Exxon, Ferrari, Merck und eBay.
SAP HANA in Memory wird als ERP/ECC-(OLTP) Datenbank der SAP Kunden oder alternativ in vielfältiger Art als dedizierte Big Data Datenbank genutzt.
Der Typesafe Stack (Akka, Scala, etc.) von Typesafe Inc. implementiert die Prinzipien des Reactive Manifesto. Akka ist ein Framework mit den Schwerpunkten Skalierbarkeit und Robustheit sowie der Fähigkeit, Berechnungen transparent in einem Cluster zu verteilen. Einsatzgebiete reichen von Twitter (skaliert durch Scala) über LinkedIn (nutzt Akka) bis zu Klout (dt.).

Big-Data-Technologien
SEMANTIS verknüpft die innovativen Technologien von Tableau, SAP und Typesafe, damit Sie schon heute die Standards von Morgen nutzen können.

Hierbei setzen wir auf hohe Expertise, z. B. beim ersten weltweiten SAP HANA Online-Training der SAP AG haben unsere teilnehmenden Mitarbeiter unter den Top 500 Absolventen von über 40.000 IT-Professionals abgeschlossen.

11.16 Software AG

Kontaktinformationen
Software AG
Uhlandstr. 12 | 64297 Darmstadt
Tel.: +49 (0)6151 92-0 | webinfo@softwareag.com
www.softwareag.com

Die Software AG (FRA: SOW) hilft Unternehmen, ihre Geschäftsziele schneller zu erreichen. Mit den Technologien des Unternehmens für Big Data, Integration und Geschäftsprozessmanagement. Seit mehr als 40 Jahren steht das Unternehmen für Innovationen, die sich am Nutzen für den Kunden ausrichten. Mit den Produktfamilien Adabas und Natural, webMethods, Terracotta, Apama, ARIS und Alfabet ist das Unternehmen führend in 15 Marktsektoren. Die Software AG beschäftigt ca. 5.300 Mitarbeiter in 70 Ländern und erzielte 2012 einen Umsatz von 1,05 Milliarden Euro.

Big-Data-Lösungen
Die Software AG unterstützt mit ihren Lösungen für Intelligente Geschäftsoperationen (Intelligent Business Operations – IBO) und Big-Data-Management Unternehmen dabei, operative Geschäftsabläufe in Echtzeit zu analysieren und zu optimieren sowie neue Geschäftsmodelle zu erschließen.

Die Grundlage für die branchenspezifischen Lösungen bildet eine hochperformante und skalierbare In-Memory-Computing-Plattform, die es ermöglicht, große Datenmengen:

- im Hauptspeicher mit schnellem Zugriff zu verwalten,
- effizient und flexibel zu importieren, exportieren und zwischen Systemen auszutauschen,
- in Echtzeit zu analysieren, sowohl ad-hoc (data at rest) als auch kontinuierlich (data in motion), um wertvolle Informationen abzuleiten,
- bedarfsgerecht und zeitnah zu visualisieren und schließlich gezielt Maßnahmen einzuleiten.

Die Daten können dabei den unterschiedlichsten Datenquellen entspringen, seien es transaktionale oder analytische Daten, historische Daten aus Datenbanken, BI-Systemen oder Hadoop, Prozessinformationen, oder etwa Live-Datenströme von Sensoren, Tweets oder mobilen Endgeräten. Der Kern der Softwareplattform integriert und erweitert bereits im Markt führende, hochgradig innovative Produkte für das In-Memory-Datenmanagement (Terracotta BigMemory), das High-Speed Messaging (Terracotta Universal Messaging) sowie die Echtzeitdatenanalyse und -visualisierung (Apama Complex Event Processing und Presto).

Einsatz-Szenarien
Die IBO- und Big-Data-Management-Plattform der Software AG unterstützt die schnelle Entwicklung und Bereitstellung von Big-Data-Lösungen in verschiedenen Branchen und Industrien. Derzeit finden sich Anwendungen vorrangig in den folgenden Bereichen:

- Elektronischer Handel: Algorithmic Trading, Trade Surveillance, Anti-Money Laundering
- Customer Experience Management: Analyse des Kundenverhaltens mit personalisierten Echtzeitinteraktionen
- Betrugserkennung, – vermeidung, und Risikomanagement in Echtzeit
- Gewinnung von Live-Einblicken in operative Geschäftsabläufe, z. B. zur Supply Chain Visibility oder dem End-zu-End-Monitoring von Geschäftsprozessen
- Beschleunigung und Skalierung transaktionaler Anwendungen: Verbesserung von Zugriffszeiten auf Datenbanken, Mainframe Offloading
- Sensordatenmanagement: z. B. im Umfeld von Industrie 4.0, Preventive Maintenance, Smart Grids/Energy Management

Big-Data-Technologien
Hinter der Terracotta In-Memory-Datenmanagement-Technologie der Software AG, die basierend auf der Open Source Bibliothek ehcache den De-facto-Standard für Java Caching definiert, steht eine große Gemeinde von mehr als 2,1 Millionen Entwicklern in über 190 Ländern und mit über 500.000 Enterprise-Installationen. Kombiniert wird diese Technologie zum Datenmanagement mit marktführenden Technologien für das Complex Event Processing sowie Universal Messaging, deren Nutzen und Überlegenheit unter anderem durch die große Kundenbasis im anspruchsvollen Finanzumfeld nachgewiesen wurde. Durch die enge Integration dieser innovativen Technologien, die teilweise das Ergebnis jahrelanger Forschung sind, entsteht eine hocheffiziente, robuste, aber trotzdem flexible Softwareplattform, mit der Big-Data-Lösungen schnell und zuverlässig entwickelt werden können.

Neben dem skalierbaren Hochleistungskern verfügt die IBO- und Big-Data-Plattform über eine Anbindung an die webMethods-Suite, um einerseits Daten aus dem webMethods-Umfeld wie z. B. dem Enterprise Service Bus oder der Geschäftsprozessmanagement-Suite effizient analysieren und andererseits Geschäftsprozesse dynamischer und intelligenter gestalten zu können.

Der Fokus der Plattform liegt im Bereich der Echtzeit-Informationsgewinnung, um im richtigen Moment die richtige Entscheidung einleiten zu können. Das Hadoop-Ökosystem, Werkzeuge zur nachgelagerten Datenanalyse sowie Data-Mining-Ansätze sind komplementär zu sehen, wobei die Softwareplattform bereits über verschiedenste Anbindungs- und Integrationsmöglichkeiten wie etwa Hadoop-Konnektoren verfügt. Sie ist unabhängig von spezieller Hardware einsetzbar, allerdings kooperiert die Software AG mit strategischen Partnern wie etwa Fujitsu zur Bereitstellung entsprechender Hardwareinfrastrukturen bei Bedarf.

11.17 Talend Germany

Kontaktinformationen
Talend Germany GmbH
Stefan Lipp
Servatiusstrasse 53 | 53175 Bonn
Tel.: +49 (0) 228 76 37 76 0 | info@talend.com
www.talend.com

Big-Data-Lösungen
Mit den marktführenden Lösungen von Talend können Kunden die »Time-to-Value« eines jeden Integrationsprojektes durch einen einfachen Prozess beschleunigen und die Konsistenz des Integrationsprozesses über alle Projekte und Integrationsarten hinweg sicherstellen. Talend for Big Data läuft als einzige Integrationslösung vollständig innerhalb von Hadoop und nutzt die Möglichkeiten der durch Open Source initiierten Innovationen, um die neuesten Versionen der Big-Data-Plattformen wie Hadoop 2.0/YARN und NoSQL zu unterstützen.

Von kleinen Projekten bis hin zu unternehmensweiten Installationen maximiert die hochgradig skalierbare Talend-Plattform für Daten-, Anwendungs- und Geschäftsprozessintegration den Wert von Informationen in einer Organisation und optimiert über ein nutzenbasiertes Subskriptionsmodell den Return on Investment. Die flexible Architektur von Talend eignet sich für Big-Data-Umgebungen und lässt sich leicht an zukünftige IT-Plattformen anpassen. Alle Talend-Produkte teilen sich einen Satz leicht bedienbarer Werkzeuge, so dass sich auch die Fähigkeiten der Entwickler in den Teams skalieren lassen. Mehr als 4.000 Unternehmen weltweit nutzen Lösungen und Services von Talend. Das Unternehmen hat Hauptniederlassungen in Nordamerika, Europa und Asien und betreibt ein globales Netz aus Technik- und Servicepartnern.

11.18 Teradata

Kontaktinformationen
Teradata GmbH
Dachauer Straße 63 | 80335 München
Tel.: +49 (0)89 12009-694
Marketing.CentralEurope@Teradata.com
www.teradadata.de

Teradata – Big Data Analysen seit 35 Jahren
Teradata (NYSE: TDC) – gegründet 1979 – ist ein weltweit führender Anbieter von analytischen Datenplattformen, Marketing- und Analyseanwendungen sowie Beratungsleistungen. Die innovativen Lösungen von Teradata unterstützen Unternehmen dabei, ihre Daten so zu integrieren und zu analysieren, dass sie mehr Wissen über ihre Kunden erlangen, bessere Entscheidungen treffen und wettbewerbsfähiger werden. Mit rund 10.000 Mitarbeitern in 77 Ländern betreut Teradata mehr als 2.500 Kunden, zu denen Top-Unternehmen aus allen wichtigen Branchen gehören. Teradata zeichnet sich durch ethisches Handeln und zukunftsweisendes Denken aus und wird von Medien und Analysten wegen seiner Technologiekompetenz, Stabilität und Kundenorientierung anerkannt. Weitere Informationen unter www.teradadata.de.

Mit den Übernahmen von Aprimo (2011) und eCircle (2012), und deren Integration, hat Teradata seine Position als ein führender Anbieter auf dem Markt für Integriertes Marketing Management, Marketing Ressource Management, digitales Marketing und Media Services weiter ausgebaut. Die Anwendungen von Teradata versetzen Marketingexperten damit in die Lage, konsistente Kampagnen personalisiert und zugleich über alle Kanäle integriert durchzuführen.

Features
Umsatz 2012: $ 2,67 Milliarden

Schwerpunkte
Integriertes Data Warehousing, Big Data-Analysen, integriertes Marketing Management und weitere Marketing- und Analyseanwendungen sowie Beratungsleistungen.

Einsatz-Szenarien
- Operations: Lösungen für Master Data Management, SAP Integration, Supply-Chain Management & Logistics
- Risk & Finance: Finance & Performance Management, Enterprise Risk Management, Tax & Revenue Management
- Business Strategy & Analytics: Lösungen für Big Data Analytics, Business Intelligence, Data Governance, Demand Planning und Data Mining & Analytics
- Marketing: Lösungen für integriertes Marketing Management, Marketing Operations, Multi-Channel Kampagnenmanagement, Digital Messaging, Marketing Analytics & Customer Data Management

Big-Data-Technologien
Mit der neuen Teradata Unified Data Architecture™ können MapReduce-basierte Big Data Lösungen wie Aster, Hadoop und Teradata einfach zu einer einheitlichen und leistungsfähigen Analyseumgebung zusammengefasst werden. Durch den Einsatz von Teradata Intelligent Memory wird die Performance von In-Memory-Technologien kostenoptimiert in der Teradata Systemwelt verfügbar.

11.19 TU Berlin - DIMA

Kontaktinformationen
Prof. Dr. Volker Markl
Fachgebiet Datenbanksysteme und Informationsmanagement (DIMA)
Einsteinufer 17, Sekr. EN 7 | 10587 Berlin
Tel.: +49 (0) 30 314 23555 | sekr@dima.tu-berlin.de
ww.dima.tu-berlin.de

Big-Data-Lösungen
STRATOSPHERE als Flagschiffprojekt des Fachgebiets ist eine von der DFG geförderte Forschergruppe, in der fünf Fachgebiete an drei Universitäten in Berlin und Potsdam die skalierbare Analyse von großen Datenmengen in Echtzeit untersuchen. Das im Projekt entwickelte System Stratosphere ist eine open-source Plattform zur Analyse von großen Datenmengen mit geringer Latenz, welches weit über die Funktionalität und Performance von den derzeit üblichen Systemen für Big Data Analytics hinausgeht.

Einsatz-Szenarien
Stratosphere wird derzeit von mehreren Unternehmen, Universitäten und Forschungsinstitutionen im Kontext von Big Data Analytics evaluiert/eingesetzt, (z. B. Telekom, Internet Memory Research, INRIA, KTH Stockholm, Universität Trento, SZTAKI Budapest) und ist als skalierbare Datenanalyseinfrastruktur Bestandteil des Smart Data Innovation Lab.

Big-Data-Technologien
DIMA führt Forschungsarbeiten in den Gebieten Technologie von Informationssystemen, Textmining, Informationsmarktplätze, Business Intelligence, Informationsmodellierung und Datenbanktheorie durch. Dabei stehen im Bereich der Technologie »Modelle und Methoden der massiv-parallelen Informationsverarbeitung«, robuste Anfrageoptimierung sowie neue Rechnerarchitekturen für das Informationsmanagement im Fokus der aktuellen Forschung.

11.20 T-Systems

Kontaktinformationen
T-Systems International GmbH
Hahnstraße 43d | 60528 Frankfurt
Tel.: + 49 (0) 69 20060-0 | info@t-systems.com
www.t-systems.de/bigdata
www.t-systems.de/big-data

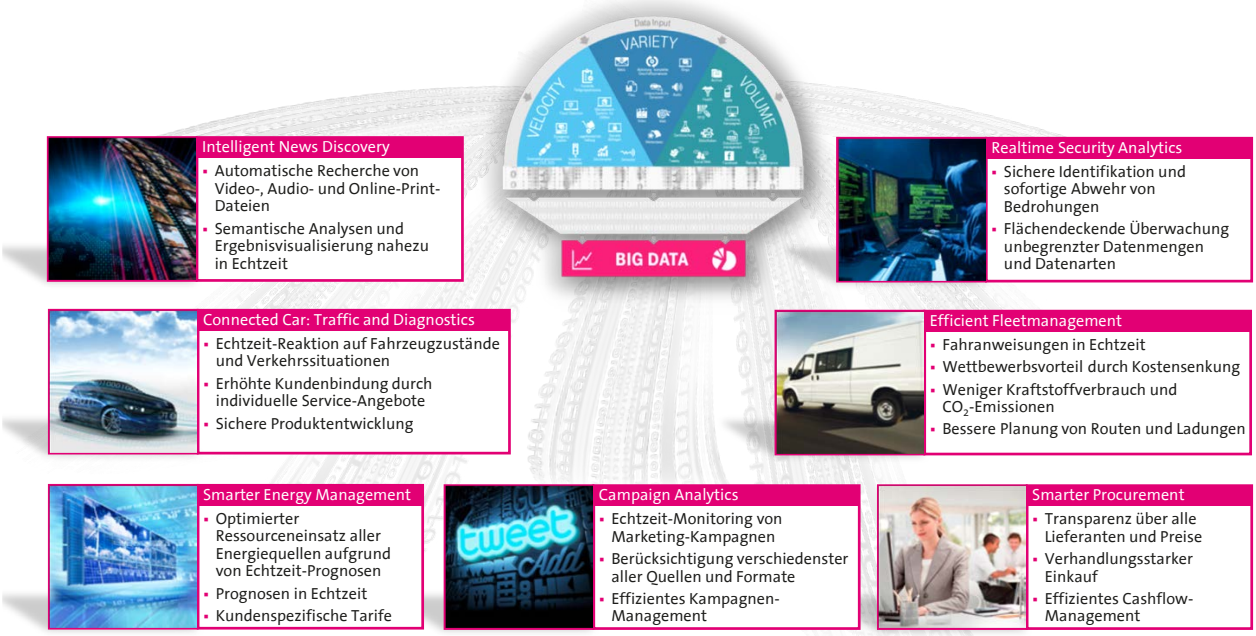
Big-Data-Lösungen
Die smarten Big Data Lösungen von T-Systems vereinen bewährte BI-Ansätze mit neuen Technologien. Die Schwerpunkte diesen smarten Daten-Managements sind:

- Enterprise Business Intelligence: Diese Lösungen sind Grundlage für operative und strategische Entscheidungen.
- High Performance Business Intelligence: Auf Basis dieser Lösungen können Unternehmen ihre Massendaten in Echtzeit verarbeiten.
- Emerging Technologies, Hadoop und NoSQL: Diese Technologien sind Grundlage für die Verarbeitung extrem hochvolumiger, unstrukturierter Daten.

Ganz gleich was Ihre Herausforderung bei der Analyse von Daten sind, smartes Data Management stellt sicher, dass die richtige Lösung zur Umsetzung Ihrer Anforderungen bereitgestellt wird.

Einsatz-Szenarien
T-Systems bietet Ihnen Systemlösungen, die je nach Anforderung jedes »V« optimal abdecken und die Sie beim Angang Ihrer Herausforderungen individuell unterstützen. Ob Sie durch Volume, Velocity oder Variety herausgefordert werden, T-Systems erschließt das entscheidende V – den Value.
Horizontal und alle Branchen.

Beispiele für innovative Einsatz-Szenarien von T-Systems:



Entsprechend Ihrer individuellen Herausforderungen entwickeln wir maßgeschneiderte Big-Data-Lösungen als Grundlage für nachhaltige Prozessoptimierungen.

Big-Data-Technologien

End-to-End: T-Systems bietet Ihnen sämtliche Big-Data-Services von Beratung, Potenzialanalyse, Strategieentwicklung über Realisierung bis zu Hosting und kontinuierlicher Optimierung Ihrer Big-Data-Lösungen aus einer Hand. Langjähriges Branchen-Know-how und große Implementierungs-kompetenz ermöglichen die sichere, individuelle Umsetzung auch komplexer Big-Data-Lösungen.

Best-of-Breed: Als produktunabhängiger Systeminteg-rator mit umfangreichem Know-how der Big-Data-Tools der führenden Anbieter können Sie sich bei T-Systems auf erstklassige Lösungen verlassen, die entsprechend Ihrer individuellen Anforderungen konzipiert werden.

Cloud: Jede Big Data Lösung bieten wir dynamisch aus der T-Systems Cloud an, so dass Sie Ihre Services ganz nach aktuellen Bedarf sicher und auf Basis klar definierter KPI's beziehen können.

11.21 PwC

Kontaktinformationen
PwC AG
Florian Buschbacher
Big Data & Data Analytics
Friedrichstr. 14 | 70174 Stuttgart
Tel.: +49 (0) 711 25034 3345
florian.buschbacher@de.pwc.com
www.pwc.de

Big-Data-Lösungen
Mit unserem Digital-Transformation-Ansatz bieten wir Ihnen einen integrierten Beratungsansatz, zugeschnitten auf Ihre Branche, der Sie dabei unterstützt, Wachstumspotentiale in der sich weiter entwickelnden digitalen Welt zu identifizieren und auszuschöpfen.

Auf Basis Ihrer digitalen Fitness helfen wir Ihnen dabei, die Dynamiken der digitalen Wirtschaft und Bedürfnisse Ihrer Kunden zu verstehen sowie die damit verbundenen Chancen, aber auch Risiken erfolgreich anzugehen.

Unsere Data Analytics Experten helfen Ihnen Big Data von der Unternehmensstrategie abgeleitete Ziele umzusetzen. Sie sind spezialisiert auf die Entwicklung, Implementierung und Prüfung hochkomplexer Algorithmen, der Analyse strukturierter, unstrukturierter und polystrukturierter Daten sowie Datenströme und beraten umfassend zu Themen des Data Managements.

Einsatz-Szenarien
Strategieentwicklung, Strategieumsetzung, digitale Transformation, Know Your Customer Analytics, Fraud Detection Analytics, Management Analytics, Business Analytics, Risk Analytics

Big-Data-Technologien
PwC ist als unabhängiger Berater in der Lage, zu unterschiedlichsten Technologieansätze mit tiefem Expertenwissen zu beraten.

Unser Verständnis zu Systemen, Kontrollen und Prozessen sowie regulatorischer Anforderungen erlaubt es uns, die richtigen Technologien und Tools auszuwählen und für Sie maßgeschneidert zu implementieren.

12 Glossar

Analytics Analyse
hier Gewinnung von Erkenntnissen durch komplexe Abfragen auf polsystrukturierte Daten, Datenbanken und Data-Warehouses mit spezifischen Abfragesprachen wie SQL oder Pig

Analytics Appliance
vorkonfigurierte oder paketierte Lösungen aus Hardware und Software für die Koordinierung von polystrukturierten Daten, die Ausführung von Analysen und die Präsentation der Erkenntnisse

Big Data
die Gewinnung neuer Informationen – die in kürzester Zeit sehr vielen Nutzern zur Verfügung stehen müssen – mittels enorm großer Datenbestände aus unterschiedlichsten Quellen, um dadurch schneller wettbewerbskritische Entscheidungen treffen zu können.

Broker
Makler/Buchmacher, hier: Rolle des Übermittlers von Daten zwischen Quelle und Anwender

Business Analytics
Ermittlung von Kennzahlen für Unternehmen, durch die Analyse größerer Datenmengen mit dem Ergebnis neuer Erkenntnisse aufgrund einer breiteren Datenbasis.

Business Intelligence
Gewinnung von Erkenntnissen über Zusammenhänge zwischen Informationen aus polystrukturierten Daten aus unterschiedlichsten Quellen

CAP-Theorem
Laut dem CAP-Theorem kann ein verteiltes System zwei der folgenden Eigenschaften erfüllen, jedoch nicht alle drei: C = Consistency = Konsistenz, A = Availability = Verfügbarkeit, P = Partition Tolerance = Partitionstoleranz (siehe Wikipedia)

CEPH
ein Dateisystem, das gleichzeitig Objekte, Dateien und Datenblöcke verwalten kann

Complex Event Processing
Complex Event Processing (CEP, Verarbeitung komplexer Ereignisse) ist ein Themenbereich der Informatik, der sich mit der Erkennung, Analyse, Gruppierung und Verarbeitung voneinander abhängiger Ereignisse beschäftigt. CEP ist somit ein Sammelbegriff für Methoden, Techniken und Werkzeuge, um Ereignisse zu verarbeiten, während sie passieren, also kontinuierlich und zeitnah. CEP leitet aus Ereignissen höheres, wertvolles Wissen in Form von sog. komplexen Ereignissen ab, d. h. Situationen, die sich nur als Kombination mehrerer Ereignisse erkennen lassen (vgl. Wikipedia).

Customer Analytics
Gewinnung von Erkenntnissen über das Kundenverhalten (überwiegend in Consumer-orientierten Unternehmen), beispielsweise mit dem Ziel der Entwicklung massenindividualisierter Produkte und Dienstleistungen

Data Management
Methoden und Verfahren zur Verwaltung von Daten, oft über Metadaten (Daten, die Daten beschreiben)

Data Mining
Anwendung statistischer Methoden auf sehr große Datenmengen, bspw. Im Gegensatz zur manuellen Auswertung über Funktionen eines Tabellenkalkulationsprogrammes

Data Science
Datenkunde, die Kenntnis bzw. Anwendung neuer Verfahren zur Arbeit mit Daten und Informationen wie zum Beispiel die Verwendung semantischer Verfahren oder die Erschließung von neuen Datenquellen (Sensordaten) und die Erarbeitung von Mustern oder statistischen Verfahren zur Auswertung solcher Daten

Eventual Consistency
Eine Schnittmenge des CAP-Modells hinsichtlich der ereignisbezogenen Konsistenz von Modellen .

Fraud Detection
Erkennung von Betrugsversuchen durch die Analyse von Transaktionen und Verhaltensmustern

Hadoop
Open-Source-Version des MapReduce-Verfahrens, in verschiedenen Distributionen erhältlich.

HANA
Ursprünglich: High-Performance Analytical Appliance; ein von SAP entwickeltes Produkt zum Betrieb von Datenbanken im (sehr großen) Hauptspeicher eines Computersystems

In-Memory
Bei In-Memory werden die Daten nicht physisch auf Datenträger gespeichert und wieder ausgelesen, sondern im Arbeitsspeicher gehalten und dadurch mit sehr hoher Geschwindigkeit verarbeitet.

Lambda-Architektur
Eine konstruktiv nutzbare Vorlage für den Entwurf einer Big-Data-Anwendung. Die in der Architektur vorgesehene Modularisierung spiegelt typische Anforderungen an Big-Data-Anwendungen wider und systematisiert sie.

Lustre
Linux-basierendes Betriebssystem für den Betrieb von Cluster-Architekturen

Machine Learning
Oberbegriff für die künstliche Generierung von Wissen aus Erfahrung: Ein künstliches System lernt aus Beispielen und kann nach Beendigung der Lernphase verallgemeinern. Das heißt, es lernt nicht einfach die Beispiele auswendig, sondern es »erkennt« Gesetzmäßigkeiten in den Lerndaten. So kann das System auch unbekannte Daten beurteilen. Zum Beispiel automatisierte Diagnoseverfahren, Erkennung von Kreditkartenbetrug, Aktienmarktanalysen, Klassifikation von DNA-Sequenzen, Sprach- und Schrifterkennung und autonome Systeme. (siehe Wikipedia)

Mahout
wörtlich: Elefantentreiber; hier: eine Apache-Komponente zum Aufbau von Bibliotheken für das Machine Learning

MapReduce
Verfahren zur Datenverwaltung und Indizierung

Metadaten
Daten zur Beschreibung von Daten, unter anderem, um Datenmodelle zu entwickeln.

Open Data
Konzept zum Zugang zu hoheitlichen Daten zu jedermann, beispielsweise Auskunft über die bei einer Meldestelle gespeicherten Daten über einen Bürger und die Einrichtungen, an die die Daten übermittelt worden sind.

Open Source
quelloffene Werke, zum Beispiel Software bei der man den Quellcode erhält

Predictive Analytics
das Treffen von Prognosen durch die Analyse von Daten Im Gegensatz zur Analyse historischer Zusammenhängen und Erkenntnissen; auch durch die Analyse von Daten, die möglicherweise urächlich nicht miteinander in Zusammenhang stehen

Predictive Maintenance
vorausschauende Wartung: durch die Auswertung vonMaschinendaten, beispielsweise akustischen Kennfeldern, lassen sich theoretische Vorhersagen über drohende Störungen treffen. Durch die vorausschauende Wartung werden Ersatzteile installiert, bevor das Originalteil ausfällt. Dadurch reduzieren sich Stillstandszeiten.

Python
Programmiersprache, oft innerhalb der Apache-hadoop-Softwarewelt verwendet

R
eine freie Programmiersprache für statistisches Rechnen und statistische Grafiken. Sie ist in Anlehnung an die Programmiersprache S entstanden und weitgehend mit dieser kompatibel. (siehe Wikipedia)

Sentiment-Analyse
Ableitung von Meinungsbildern aus den Mitteilungen, Nachrichten und Kommentaren von Verbrauchern in Social-Media

Streaming
im Gegensatz zur Verarbeitung von lokalen Daten ablaufende Verarbeitung kontinuierlicher Datenströme, beispielsweise laufende Übertragung und Wiedergabe von Video im Gegensatz zum Download und anschließendem Abspielen von Video

Transactional Data
Daten für/aus transaktionalen Systemen (zum Beispiel vollständige Abbildung einer Bargeldabhebung an einem Geldautomaten mit Bestandteilen wie Kartenerkennung, Authentifizierung, Autorisierung, Geldausgabe, Buchung, Prüfung, Abrechnung, etc.)

Variety
die Vielzahl an Datenquellen und Vielfalt an Datenarten in Big-Data-Szenarien

Velocity
die immer höheren Anforderungen an die IT-Systeme hinsichtlich der Bereitstellung von Berechnungsergebnissen, bspw. Lieferung von Kennzahlen innerhalb von Minuten statt Tagen

Volume
die immer größer werdenden zu handhabenden Datenmengen durch die Einbindung von immer größeren Daten(banken) in Big-Data-Szenarien

13 Sachwortregister

Der Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. vertritt mehr als 2.100 Unternehmen, davon gut 1.300 Direktmitglieder mit 140 Milliarden Euro Umsatz und 700.000 Beschäftigten. 900 Mittelständler, 200 Start-ups und nahezu alle Global Player werden durch BITKOM repräsentiert. Hierzu zählen Anbieter von Software & IT-Services, Telekommunikations- und Internetdiensten, Hersteller von Hardware und Consumer Electronics sowie Unternehmen der digitalen Medien und der Netzwirtschaft. Der BITKOM setzt sich insbesondere für eine Modernisierung des Bildungssystems, eine innovative Wirtschaftspolitik und eine zukunftsorientierte Netzpolitik ein.

BITKOM

Bundesverband Informationswirtschaft,
Telekommunikation und neue Medien e.V.

Albrechtstraße 10 A
10117 Berlin-Mitte
Tel.: 030.27576-0
Fax: 030.27576-400
bitkom@bitkom.org
www.bitkom.org