Management von Big-Data-Projekten

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:	Wulf Maier (Hewlett-Packard GmbH) Dr. Mathias Weber (BITKOM)
Copyright:	BITKOM 2013
Grafik/Layout:	Design Bureau kokliko/ Astrid Scheibe (BITKOM)
Titelbild:	Daniela Stanek (BITKOM) unter Verwendung von © envfx – Fotolia.com

Management von Big-Data-Projekten

Leitfaden

BITKOM

Liste der Abkürzungen

BDSG	Bundesdatenschutzgesetz
BI	Business Intelligence
BSI	Bundesamt für Sicherheit in der Informationstechnik
CEP	Complex Event Processing
CIC	Customer Interaction Center
CoE	Center of Excellence
CPU	Central Processing Unit
DW	Data Warehouse
EFM	Efficient Fleet Management
ETL	Extract-Transformation-Load
EU	Europäische Union
EVU	Energieversorgungsunternehmen
EWR	Europäischer Wirtschaftsraum
GeLiGas	Geschäftsprozesse Lieferantenwechsel Gas
GPKE	Geschäftsprozesse zur Kundenbelieferung mit Elektrizität
GPS	Global Positioning System
ILM	Information Lifecycle Management
IM	Information Management
IMDG	In Memory Data Grids
IP	Internet Protocol
IT	Informationstechnologie
MaBIS	Marktregeln für die Durchführung der Bilanzkreisabrechnung Strom
MDM	Meter Data Management
MDS	Managed Desktop Services
MLCM	Mobile Logistics Cost Management
MUS	Multi Utility Server
PAYD	Pay as you drive
PIA	Privacy Impact Assessment
PoC	Proof of Concept
RFID	Radio-Frequency Identification
RLM	Registrierende Lastgang Messung
RO*I*	Return on Information
ROI	Return on Investment[1]
SLP	Standardlastprofil
SSD	Solid State Disk
TK	Telekommunikation
TKG	Telekommunikationsgesetz
TMG	Telemediengesetz
VO	Verordnung
WLAN	Wireless Local Area Network

[1] Wenn mit RO*I* Return on Information gemeint ist, wird das I stets kursiv gesetzt.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

1 Geleitwort

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

Unternehmen, Wissenschaft oder auch der Staat sehen sich heute mit einem rapiden Anstieg des Datenvolumens konfrontiert, denn Smartphones, Social Media, Sensorik in allen möglichen Geräten, RFID-Chips oder neue E-Government-Anwendungen werden immer stärker genutzt. Die Digitalisierung von Infrastrukturen in den Bereichen Energie, Gesundheit, Verkehr, Bildung und öffentliche Verwaltung lässt die Datenmengen weiter steigen. Doch Big Data ist nicht einfach die Verarbeitung riesiger Datenmengen. Mit dem Begriff Big Data wird die Gewinnung und Nutzung entscheidungsrelevanter Erkenntnisse aus unterschiedlichen Datenquellen bezeichnet – aus unternehmenseigenen Datenbanken, vernetzten Produktionsmaschinen oder aus dem freien Internet. Diese Daten können zudem einem schnellen Wandel unterliegen und in bisher ungekanntem Umfang anfallen. In der Kombination dieser Merkmale besteht das Wesen von Big Data. Den Wert haben aber nicht die Daten an sich, sondern die Erkenntnisse, die sich mit neuen Verfahren gewinnen lassen – mit Business Analytics. In der digitalen Welt treten Daten als vierter Produktionsfaktor neben Kapital, Arbeitskraft und Rohstoffe.

Big-Data-Technologien spielen in Unternehmen überall dort ihre Stärken aus, wo qualitativ unterschiedliche Daten in hohen Volumina anfallen, unter anderem in Forschung und Entwicklung, Produktion, Distribution und Logistik, Finanz- und Risiko-Controlling sowie in Marketing und Vertrieb. Die Einsatzgebiete für Big-Data-Lösungen gehen aber weit über die Wirtschaft hinaus: Von der kontrollierten Nutzung der Daten digitalisierter Infrastrukturen etwa können nahezu alle Menschen profitieren. Damit schaffen Big-Data-Lösungen einen gesamtgesellschaftlichen Mehrwert.

Big Data ist eine neuartige Technologie, die vielfältige gesellschaftliche und politische Facetten aufweist. Die Nutzung großer Datenmengen verlangt auf der einen Seite Vertrauen und auf der anderen Seite eine hohe Schutz- und Sicherheitskompetenz – vor allem dann, wenn die Technologie auf personenbezogene Daten Anwendung finden soll.

Die politische Herausforderung besteht darin, rechtliche Rahmenbedingungen zu finden, die sowohl die Nutzung von Big Data zulassen als auch die Persönlichkeitsrechte in ausreichendem Maße schützen. Intelligente Ansätze

sind gefragt, um die Potenziale neuer Technologien in rechtlich und gesellschaftlich akzeptablen Grenzen umzusetzen. Dazu gehören beispielsweise komplexe Anonymisierungs- oder Pseudonymisierungsprozesse, die einen zulässigen Umgang auch mit personenbezogenen Daten ermöglichen.

Für die exportorientierte deutsche Wirtschaft führt an Big Data kein Weg vorbei. Auch aus gesellschaftlicher Perspektive ist der Einsatz von Big-Data-Technologien wünschenswert. Sie helfen uns, einige der wichtigsten Herausforderungen unserer Zeit zu lösen – wir brauchen smarte Energienetze für den Ausstieg aus der Kernkraft, intelligente Verkehrssysteme zur Beibehaltung hoher Umweltqualität sowie intelligente Gesundheitssysteme zur Sicherstellung flächendeckend hochwertiger medizinischer Betreuung. Diese Potenziale sollten wir verstehen und unter Abwägung der Risiken und Chancen ganz bewusst im notwendigen Umfang nutzen.

Prof. Dieter Kempf, Präsident, BITKOM e.V.

2 Management Summary

Big-Data-Publikationsserie

Der vorliegende Leitfaden ist die zweite Publikation des BITKOM-Arbeitskreises Big Data. Im ersten Leitfaden[2] wurde erläutert (vgl. Tabelle 1), was unter Big Data zu verstehen ist und worin der wirtschaftliche Nutzen beim Einsatz von Big Data besteht.

Nr.	Kernthese
1	In der digitalen Welt treten Daten als vierter Produktionsfaktor neben Kapital, Arbeitskraft und Rohstoffe.
2	Viele Unternehmen werden konventionelle und neue Technologien kombinieren, um Big-Data-Lösungen für sich nutzbar zu machen.
3	Der überwiegende Teil der in Unternehmen vorliegenden Daten ist unstrukturiert, kann aber in eine strukturierte Form überführt sowie quantitativen und qualitativen Analysen zugänglich gemacht werden.
4	Empirische Studien sowie zahlreiche Einsatzbeispiele belegen den wirtschaftlichen Nutzen von Big Data in vielen Einsatzgebieten.
5	Einige Funktionsbereiche von Unternehmen sind für den Big-Data-Einsatz prädestiniert. Dazu gehören Bereiche wie Marketing und Vertrieb, Forschung und Entwicklung, Produktion sowie Administration/ Organisation/Operations.
6	Der Einsatz von Big-Data-Methoden sollte bereits in der Konzeptionsphase aus rechtlicher Sicht geprüft werden.
7	Hohe zweistellige Wachstumsraten im Markt über einen längeren Zeitraum sind ein weiterer Beleg für die wirtschaftliche Bedeutung von Big-Data-Lösungen.
8	Es ist im volkswirtschaftlichen Interesse, Erfahrungen und Best Practices bei der Nutzung von Big Data effektiv zu kommunizieren.

Tabelle 1: Kernthesen des ersten BITKOM-Big-Data-Leitfadens

Big Data – eine Begriffsbestimmung

Im ersten BITKOM-Leitfaden wurde ausgeführt: Big Data unterstützt die wirtschaftlich sinnvolle Gewinnung und Nutzung entscheidungsrelevanter Erkenntnisse aus qualitativ vielfältigen und unterschiedlich strukturierten Informationen, die einem schnellen Wandel unterliegen und in bisher ungekanntem Umfang zu Verfügung stehen (vgl. Abbildung 1). Big Data spiegelt den technischen Fortschritt der letzten Jahre wider und umfasst dafür entwickelte strategische Ansätze sowie eingesetzte Technologien, IT-Architekturen, Methoden und Verfahren.

Mit Big Data erhalten Manager eine deutlich verbesserte Grundlage für die Vorbereitung zeitkritischer Entscheidungen mit besonderer Komplexität.

[2] Vgl. [BITKOM, 2012]

Aus Business-Perspektive verdeutlicht Big Data, wie auf lange Sicht die Daten zu einem Produkt werden. Big Data öffnet die Perspektive auf die »industrielle Revolution der Daten«, während gleichzeitig Cloud Computing den IT-Betrieb industrialisiert.

Aus IT-Perspektive markiert Big Data die aufkommenden Herausforderungen sowie neuen technologischen Möglichkeiten für Speicherung, Analyse und Processing schnell wachsender Datenmengen.

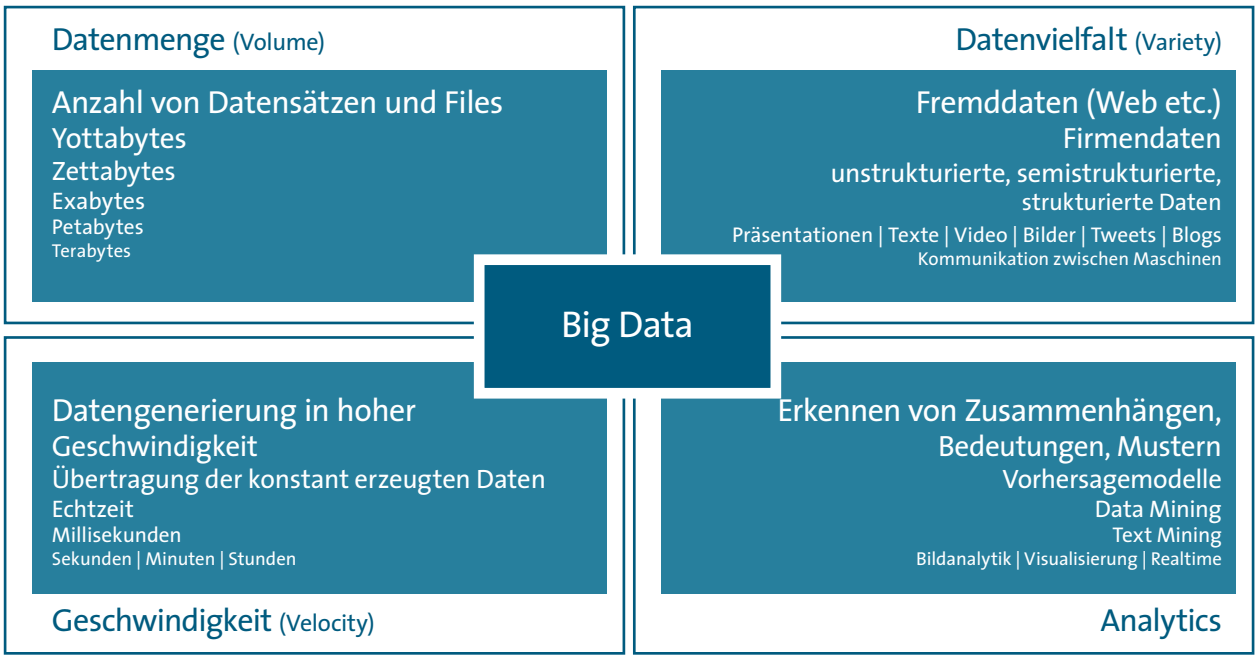


Abbildung 1: Merkmale von Big Data

Im vorliegenden zweiten BITKOM-Leitfaden über Big Data erfährt der Interessierte, wie man Big-Data-Projekte aufsetzt, wie Big-Data-Expertise im Unternehmen aufgebaut wird und welche Technologien[3] das neue Phänomen charakterisieren.

Wertschöpfungs- und Business-Modelle in der Daten-Wirtschaft

»Big Data – Die Daten werden zum Produkt« – diese Kernaussage umreißt präzise die Auswirkungen des Big-Data-Phänomens auf die Wertschöpfungs- und Geschäftsmodelle in den Industrie- und Dienstleistungsbranchen. So ist evident, dass durch die voranschreitende Digitalisierung von nahezu allen Geschäftsmodellen und -prozessen das Volumen der zur Verfügung stehenden Daten explodiert und deren Aktualität eine neue Stufe erreicht; es entstehen vollkommen neue Nutzungs- und Kommerzialisierungsmöglichkeiten für Daten. Je nach Branche ergeben sich somit vielfältige Optionen, um mit Big-Data-Lösungen bestehende Geschäftsprozesse zu optimieren und komplett neue Produkte und Dienstleistungen zu entwickeln.

[3] Der dritte BITKOM-Leitfaden über Big Data wird die Technologien im Einzelnen vorstellen.

Das Kapitel 3 liefert Entscheidern erste Ansätze und Beispiele, um die strategischen Handlungsmöglichkeiten für das eigene Unternehmen bzw. die eigene Branche zu bestimmen und zu evaluieren.

Datenschutz in Big-Data-Projekten

Bei der Entwicklung und bei der Umsetzung von Big-Data-Projekten spielt der Datenschutz insbesondere dann eine wichtige Rolle, wenn personenbezogene Daten verwendet werden sollen. Personenbezogen sind solche Angaben, die persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren Person betreffen.

Anknüpfend an die bereits im ersten Big-Data-Leitfaden[4] des BITKOM erläuterten Grundlagen des Datenschutzes werden im vorliegenden Dokument zwei Themen in den Mittelpunkt gerückt (vgl. Kapitel 4), die für Big-Data-Projekte eine besonders große Bedeutung haben:

- Datenschutz-Folgenabschätzung (Privacy Impact Assessments) und
- Vorgehen bei Anonymisierung.

Vorgehensmodell zur Umsetzung von Big-Data-Projekten

Wegen der hohen Komplexität von Big-Data-Projekten ist es empfehlenswert, sich bei ihrer Entwicklung und Umsetzung an einem Vorgehensmodell zu orientieren (vgl. Kapitel 5). Ein Vorgehensmodell unterstützt Unternehmen dabei, alle Schritte und Prozesse von Big-Data-Projekten transparent und nachvollziehbar zu gestalten. Im Sinne der Nachhaltigkeit sollten Big-Data-Projekte von der frühen Planung bis zur mittel- und langfristigen Optimierung durchgängig begleitet werden.

Das in diesem Leitfaden vorgeschlagene Vorgehensmodell umfasst in acht Phasen alle Aktivitäten von der Identifikation möglicher Big-Data-Potenziale, über die konkrete Planung, die Umsetzung einschließlich der Konsolidierung der IT-Infrastruktur und die Erschließung neuer Datenquellen bis hin zum Betrieb und zur Optimierung von Geschäftsprozessen.

Erfolgsfaktoren und Management-Aufgaben bei der Einsatzvorbereitung und Nutzung von Big Data

Big-Data-Lösungen können messbare Beiträge für die Wertschöpfung leisten, wenn die Entscheider eine Reihe von Erfolgsfaktoren beachten (vgl. Kapitel 6). Dazu zählen u.a. die Entwicklung einer wertorientierten Big-Data-Strategie und einer daraus abgeleiteten Roadmap, der Business-Fokus, die Etablierung eines umfassenden Big-Data-Innovationsprozesses, die regelmäßige Erfolgsüberprüfung sowie die Einrichtung einer passenden Organisation.

Kompetenzentwicklung der Mitarbeiter für Big-Data-Projekte

Wissen über die Einsatzmöglichkeiten von Big Data und die damit verbundenen Technologien ist zurzeit noch rar. Dieses Defizit muss durch Aus- und Weiterbildung der Spezialisten in der Software-Entwicklung, im IT-Betrieb, in den Fachabteilungen sowie im Management zügig abgebaut werden (vgl. Kapitel 7). Dazu eignen sich insbesondere Kollaborationsplattformen. Besonderer Wert muss dabei auf die Förderung von Kreativität gelegt werden.

Big-Data-Projekte sind einerseits normale IT-Projekte, für die etablierte Methoden und Verfahren des Projektmanagements zur Verfügung stehen. Andererseits stellt die Erschließung der Möglichkeiten von Big Data für jedes Unternehmen einen Innovationsprozess mit einer Vielzahl von Facetten und Implikationen dar. Wie tief der mit Big Data eingeläutete Wandel für ein Unternehmen ist, hängt

[4] vgl. [BITKOM 2012]

von seiner Strategie ab. Für einige Anwendungsfälle kann Big Data durchaus eine disruptive Technologie darstellen.

Architekturen und Basistechnologien für Big Data

Eine funktionale Referenzarchitektur eignet sich als Grundlage für die Big-Data-Umsetzung. Dazu stehen Open-Source-Werkzeuge sowie kommerzielle Hard- und Software zur Auswahl (Kapitel 8). Innovative Technologien unterstützen die Umsetzung der funktionalen Aspekte.

Deployment- und Betriebsmodelle für Big-Data-Anwendungen

Alle Organisationen, die Big-Data-Lösungen etablieren wollen, sehen sich beim Deployment und beim Betrieb von Big-Data-Anwendungen mit ähnlichen Herausforderungen konfrontiert.

Eine Reihe von Parametern (vgl. Abschnitt 9.1) bestimmen, in welchem Maß die Architektur skalierbar ist; die Skalierbarkeit bildet wiederum die Voraussetzung für das Deployment eines erfolgreichen Big-Data-Projekts. Für jeden Parameter werden die Auswirkungen auf das Deployment und das Betriebsmodell dargestellt. Auf dieser Basis werden Big-Data-relevante Grundsätze für skalierbare Architekturen und deren Notwendigkeit aufgezeigt. Wie eine geeignete Architektur betrieben werden kann, beschreibt das Kapitel 9.2.

3 Wertschöpfungs- und Business-Modelle in der Daten-Wirtschaft

»Big Data – Die Daten werden zum Produkt« – diese Kernaussage umreißt präzise die Auswirkungen des Big-Data-Phänomens auf die Wertschöpfungs- und Geschäftsmodelle in den Industrie- und Dienstleistungsbranchen. So ist evident, dass durch die voranschreitende Digitalisierung von nahezu allen Geschäftsmodellen und –prozessen das Volumen der zur Verfügung stehenden Daten explodiert und deren Aktualität eine neue Stufe erreicht; es entstehen vollkommen neue Nutzungs- und Kommerzialisierungsmöglichkeiten für Daten. Je nach Branche ergeben sich somit vielfältige Optionen, um mit Big-Data-Lösungen bestehende Geschäftsprozesse zu optimieren und komplett neue Produkte und Dienstleistungen zu entwickeln.
Das Kapitel 3 liefert Entscheidern erste Ansätze und Beispiele, um die strategischen Handlungsmöglichkeiten für das eigene Unternehmen bzw. die eigene Branche zu bestimmen und zu evaluieren.

3.1 Megatrends

Im Abschnitt 3.1 werden die Megatrends beschrieben, die die Grundlage für die Entwicklung neuer Big-Data-zentrierter Geschäfts- und Wertschöpfungsmodelle bilden.

3.1.1 Industrial Internet

In den letzten 10 Jahren war das Einsatzspektrum moderner Internettechnologien auf wenige Branchen beschränkt. So transformiert das Internet bislang hauptsächlich die Informations- und Medienbereitstellung. Durch neue mobile Endgeräte und leicht bedienbare Interface-Technologien werden zunehmend auch Dienstleistungen in Form von Apps und Subskriptions-Services angeboten.

In den industriell geprägten Branchen hat das Internet bislang nur wenige Auswirkungen auf die Produktions- und Vermarktungsprozesse gehabt. Dies beginnt sich mit dem Eintritt in das Big-Data-Zeitalter zu ändern. Die Kombination aus kostengünstigen Sensortechnologien, leistungsfähigen IT-Infrastrukturen und einer hohen Nachfrageelastizität aufgrund vorausschauender und hoch flexibler Analyse- und Planungssysteme schafft eine vollkommen neue Produktionswelt.

In der Welt des Industrial Internets werden die Material- und Produktionsflüsse weiter optimiert, da nahezu alle Eingangsressourcen einzeln lokalisiert und verfolgt werden können. So wandert das Feedback aus den Nachfragemärkten nahezu in Echtzeit durch die verschiedenen Stufen der Lieferanten- und Produktionskette und sorgt für eine optimale Steuerung des Produktionsumfangs sowie der verbrauchten Materialien und Energieträger.

Produktivitätsschub durch Big Data

Es ist zu erwarten, dass der Einsatz von Internettechnologien in Kombination mit Big-Data-Verfahren einen neuen Produktivitätsschub entfachen wird. Analysten rechnen mit einer Verdreifachung der Zahl industriell eingesetzter Sensoren bis 2015. Parallel dazu steigen das Datenvolumen und die Zahl der Datenquellen. Allein die Aufrüstung der Energie-Infrastruktur hin zu Smart Grids wird zu einer Datenexplosion in der Industrie führen. Hinzu kommen die weitere Digitalisierung von Prozess- und Logistikketten sowie die Generierung von Messdaten in Maschinen- und Anlagenparks.

Das Industrial Internet beschreibt die tiefgreifenden Veränderungen im produzierenden Gewerbe und den angrenzenden Dienstleistungssegmenten. Es schafft eine

stärkere Verzahnung der Produktions- und Nachfrageseite und erhält aus Big Data wesentliche Impulse.

3.1.2 Aufmerksamkeits-Wirtschaft

Aufmerksamkeit ist die neue Leitwährung. Da in der digitalisierten Medienwelt alle Regungen und Bewegungen des Nutzers nachvollziehbar und auswertbar sind, wird die Verweildauer und Interaktion der Nutzer zur Leitwährung im Internet. In der Welt der sozialen Medien, die die Teilhabe der Nutzer im Fokus haben, wertet man die Nutzeraktivitäten in Engagement-Metriken aus. Klicks und Verweildauer definieren den Erfolg der meisten Internetangebote – egal ob Suchmaschine, Portal oder Social Network. Auch die Generierung von Traffic für den Verkauf auf Online-Shops folgt dieser Regel. Big Data ist daher schon heute ein fester Bestandteil im Management-Portfolio der Internetunternehmen. So werten diese detailliert das Verhalten ihrer Nutzer aus, um Werbewirkung und Verkaufsraten zu erhöhen.

Verständnis der Kundenbedürfnisse erhöhen

Mit der zunehmenden Digitalisierung unseres Alltags, die uns Informationen aus unserer beruflichen wie privaten Sphäre überall und sofort zur Verfügung stellt, wird die Aufmerksamkeit allerdings zu einem knappen Gut. Zwar steigt derzeit der Medienkonsum noch an. Doch das Wachstum bei den neuen Medien (Internet, Smartphone) geht deutlich zu Lasten der alten Medien (Print, TV). Derzeit beträgt die tägliche private Mediennutzung 585 Minuten[5]. Analysten erwarten ab 2016 eine Sättigung der Mediennutzung. Die Effektivität von klassischer Online-Werbung wird deutlich abnehmen. Die Kosten steigen drastisch an, um einen neuen Nutzer auf das eigene Internetangebot zu locken.

Um wettbewerbsfähig zu bleiben und neue Geschäftsmodelle und Services erfolgreich lancieren zu können, sind Unternehmen im Internet zukünftig noch viel stärker darauf angewiesen, ihre Kunden wirklich zu verstehen.

Das bedeutet, dass Marketing-Entscheider in der Lage sein müssen, aus dem Online-Verhalten der Kunden die richtigen Schlüsse zu ziehen und Angebote passgenau und in Echtzeit zu präsentieren.

Hierzu ist eine technologische Basis in Form von kosteneffizienten Big-Data-Datenbanken und –Tools erforderlich. Außerdem werden bei der Entwicklung neuer Algorithmen und Marketingprozesse verhaltenswissenschaftliche Erkenntnisse eine sehr wichtige Rolle spielen.

Erfolgsformel in der Aufmerksamkeits-Wirtschaft

In der Verbindung von Big-Data-Technologien und von Erkenntnissen der »Behavioral Economics« werden neue Werbe- und Recommendation-Algorithmen entwickelt, die in der Lage sind, die Präferenzen des Nutzers und seine individuellen Aufmerksamkeits- und Zeitprofile zu berücksichtigen. Die Aufgaben der Data Scientists werden sich mittelfristig um psychologische und soziologische Fragestellungen erweitern. Für die Erfolgsformel in der Aufmerksamkeits-Wirtschaft (vgl. Formel 1) liefert Big Data die grundlegenden Instrumente und Infrastrukturen, um die schier explodierenden Mengen an Kunden- und Transaktionsdaten kreativ und effizient auszuwerten.

Statistik + IT + Psychologie = Markterfolg

Formel 1: Erfolgsformel in der Aufmerksamkeits-Wirtschaft

[5] Ohne berufliche Nutzung von Smartphones etc., aber inklusive Parallelnutzung von z. B. Radio und Internet. Vgl. [SOM, 2013]

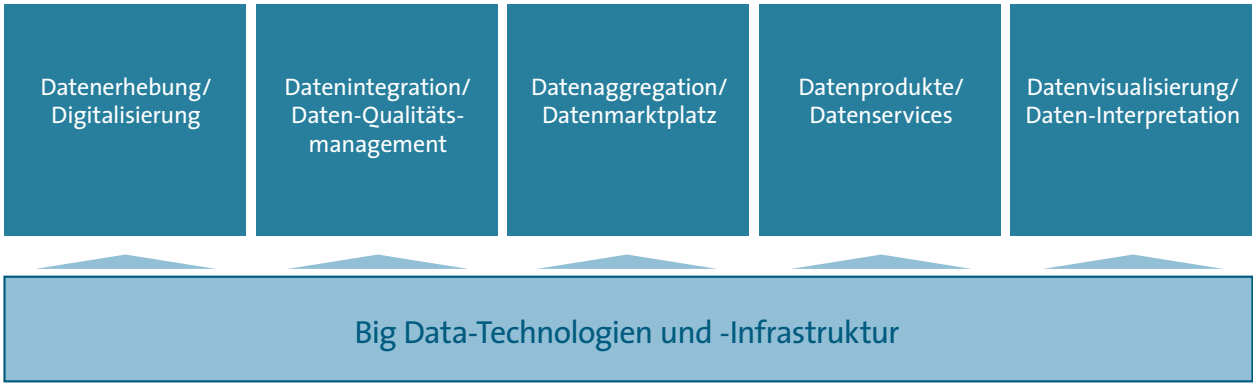


Abbildung 2: Wertschöpfungskette in der Daten-Wirtschaft

3.2 Wertschöpfungskette der Daten-Wirtschaft

Auf dem Weg zu einer »Daten-Wirtschaft« verändern sich auch die Wertschöpfungsketten. So kommen neue Wertschöpfungsstufen hinzu (z. B. Daten-Marktplätze), die jeweils unterschiedliche Skills und Automatisierungsgrade aufweisen. Das in Abbildung 2 dargestellte Modell skizziert die Wertschöpfungskette der Daten-Wirtschaft auf mittlere Sicht.

Data Technology Supply

Zu Beginn der Wertschöpfungskette stehen die Ausrüster. Denn ohne die entsprechenden Technologien, Infrastrukturen und Tools lassen sich keine neuen Services und Wertschöpfungsmodelle entwickeln und implementieren. Big Data erfordert bestimmte Investitionen: Nicht nur die Hardware, sondern auch die personellen Ressourcen werden auf absehbare Zeit kostenintensiv bleiben.

3.2.1 Datenerhebung – Digitalisierung

Eine fundamentale Rolle in der Wertschöpfungskette spielen die Akteure, die neue Daten erheben bzw. erschaffen – durch die Aufzeichnung und Sammlung von Daten existierender Geschäfts- und Produktionsprozesse[6] oder aber durch die Digitalisierung bislang analoger Daten. Beispiele hierfür sind die Digitalisierung von Büchern und Archiven, die digitale Kartographie oder die Videoüberwachung von Städten und öffentlichen Gebäuden.

3.2.2 Datenintegration – Daten-Qualitätsmanagement

Damit Daten zu Gold werden können, müssen Unternehmen eine zentrale Hürde meistern: Aufräumen, denn die Parametrisierung, Standardisierung und Integration der Datenbestände erfordert Zeit und Ressourcen. Die Prozesse des »Data Cleaning« können aber in der Praxis auf absehbare Zeit nicht vollautomatisiert werden. Zur möglichst effizienten Umwandlung großer Datenbestände in verwertbare Formate sind professionell aufgesetzte Prozesse und die Auswahl geeigneter Tools entscheidend. Für diese Aufgaben empfiehlt sich die Auslagerung an spezialisierte Dienstleister. Analysten erwarten, dass sich die Anzahl solcher Dienstleister in den kommenden 2-5 Jahren vervielfachen wird. Die Dienstleister werden sich auf bestimmte Branchen und Datentypen spezialisieren. Die Geschäftsmodelle sind meist Dienstleistungsverträge oder standardisierte Subskriptions-Services, die Data Cleaning in Abhängigkeit von der Datenmenge, der Anzahl der Quellen und der Diversität der Formate umfassen.

[6] z. B. Logging von im Produktionsprozess oder Sammlung von GPS-Daten über eine mobile App

3.2.3 Datenaggregation – Datenmarktplatz

Ein wesentlicher Teil der Big-Data-Wertschöpfungskette liegt sicherlich in der Aggregation von Daten und deren Vermakelung über Datenmarktplätze[7].

So sammelt und normiert beispielsweise das US-Startup Gnip.com Social-Media-Daten über eine Vielzahl von Communities, die dann in aggregierter Form als Datenservice an Unternehmen bzw. deren Marketingabteilungen und Social-Media-Agenturen zur Auswertung bereitgestellt werden. Die Dienstleistung bzw. Wertschöpfung liegt hier in der Durchführung bestimmter Aufgaben wie z. B. Format-Normierung, URL-Auflösung, Spracherkennung, Doublettenabgleich etc. Da diese Aufgaben nur einmal durchgeführt werden und sich die Kosten auf alle Kunden verteilen, ergibt sich ein attraktives Geschäftsmodell.

Ein weiteres Beispiel ist das Unternehmen Factual, das auf seiner Plattform orts- und produktbezogene Daten an App-Entwickler und Werbetreibende verkauft. App-Entwickler können über eine standardisierte API die Adressdaten und Bewertungen von über 60 Millionen lokalen Geschäften und Restaurants abrufen und in ihre Apps einbauen.

Datenaggregation über Datenmarktplätze

Eine weitere interessante Wertschöpfungsvariante der Datenaggregation stellen die Datenmarktplätze dar. Hier schaffen die Marktplatzbetreiber Plattformen und einheitliche Standards für den Verkauf und die Nutzung verschiedener Datensätze oder Daten-Streams.

Das isländische Unternehmen Datamarket beispielsweise bietet Datenlieferanten auf seiner Plattform ein Modell zum Vertrieb von Datenpaketen und Services an. So können beispielsweise Marktforschungs- und Beratungsunternehmen ihre Daten und Expertise auf der Plattform zu einheitlichen Konditionen vertreiben. Kunden erhalten Zugang über standardisierte APIs, so dass Abfragen, Visualisierungen oder auch die Programmierung eigener Analytik-Anwendungen möglich werden.

Nach Überzeugung von Analysten werden sich die Datenmarktplätze branchen- und anwendungsspezifisch entwickeln. Allein bis 2015 wird sich der Datenverkauf über Aggregatoren und Marktplätze nahezu verzehnfachen.

3.2.4 Datenprodukte – Datenservices – »Produktisierung«

Der Entwicklung neuer datenbasierter Apps und Services kommt das größte Marktpotenzial zu. Auch an der Wertschöpfung werden neuentwickelte Apps und Services einen großen Anteil haben, da Anwender – privat wie beruflich – die höchste Zahlungsbereitschaft für integrierte Dienste mit einem klar definierten Nutzen haben. Grundlage der neuen Data Apps und Services können sein:

- die Kombination unterschiedlicher Datenquellen[8],
- die Aufbereitung in Form von Dashboards und Visualisierungen[9],
- Analyse- und Prognosemodelle auf Basis spezieller Datenbestände[10].

Der unternehmerischen Vielfalt und Kreativität sind hier nahezu keine Grenzen gesetzt.

---

7 Vgl. z. B. das Projekt MIA, in dessen Rahmen einen Online-Marktplatz für Informationen und Analysen auf Basis der Daten des deutschen Webs entwickelt wird. http://www.trusted-cloud.de/de/795.php

8 Beispiel: Reiseplanung für Trendsportarten wie Windsurfen. Hier werden Wetterdaten mit Reisezielen kombiniert, um möglichst günstige Windbedingungen zu gewährleisten.

9 Beispiel: Darstellung von Verbrauchsdaten auf IaaS-Plattformen, die Unternehmen einen Überblick über ihre Cloud-Provider sowie deren Performance und Kosten.

10 Beispiel: Vorhersagen von Flugreisepreisen für bestimmte Airlines und Flugstrecken oder von Preisen für Autos bzw. Investitionsgüter.

3.2.5 Datenvisualisierung – Dateninterpretation

Einen wirklichen Mehrwert für Unternehmen können große Datenbestände nur erlangen, wenn sie richtig interpretiert werden. Der Beratung und visuellen Aufbereitung wird daher ein zunehmender Stellenwert zukommen. Analysten erwarten, dass sich das Beratungsangebot und der Einsatz von Visualisierungswerkzeugen spezifisch für einzelne Branchen entwickeln werden. Statistik-Know-how wird nur wertvoll sein, wenn es durch tiefes Branchenwissen ergänzt wird.

3.3 Geschäftsmodelle in der Daten-Wirtschaft

In der Daten-Wirtschaft lassen sich die Geschäftsmodelle und Business Cases in vier Kategorien einteilen (vgl. Abbildung 3).

3.3.1 Optimierung

Die Auswertung bereits existierender Datenbestände kann für die Optimierung bestehender Geschäftsprozesse und -modelle einen sehr großen Mehrwert liefern.

Es bietet sich daher als Einstieg in das Big-Data-Business für viele Unternehmen an, mit der »Optimierungs«-Strategie zu starten und zu Beginn die unternehmenseigenen Datenbestände besser zu nutzen. So lassen sich beispielsweise wertvolle Rückschlüsse aus den Wechselwirkungen des Kauf- und Online-Verhaltens der Kunden ziehen, die Auswirkungen auf die Personaleinsatzplanung in den Filialen vor Ort aufdecken können. Vorreiter auf diesem Gebiet sind sicherlich die Anbieter von Billig-Flügen, die ihre Gewinn-Management-Systeme mit einer Vielzahl weiterer Parameter, z. B. aus dem Online-Verhalten, kombiniert und optimiert haben.

3.3.2 Monetarisierung

Häufig gleichen Datenbestände einem noch nicht gehobenen Schatz. So lassen sich mit bereits existierenden Daten neue Geschäftsmodelle oder Produkte kreieren, sofern die Nutzung der Daten rechtlich zulässig ist. Beispiele hierfür sind:

- die anonymisierte Auswertung der Nutzer- und Standortdaten von Telefonnutzern zur Optimierung von lokalisierten Diensten und von ortsbezogener Werbung,

Geschäftsmodelle in der Data Economy

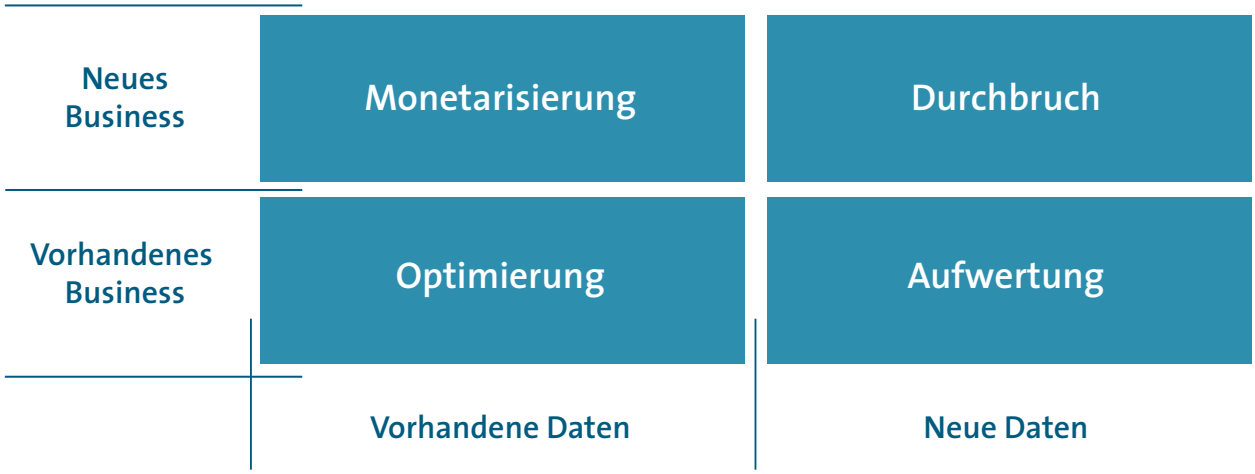


Abbildung 3: Geschäftsmodelle in der Daten-Wirtschaft

- der Weiterverkauf von aggregierten Transaktionsdaten und Nutzungsprofilen durch Kreditkartenfirmen.

- Internetunternehmen wie z. B. Google entwickeln auf Basis der Nutzungsdaten und des Suchverhaltens neue Analysedienste wie Google Trends.

- Der Navigationsanbieter TomTom vermarktet die GPS-Daten seiner Kunden über den kostenpflichtigen Dienst TomTomLive.

3.3.3 Aufwertung

Bestehende Geschäftsmodelle und Dienstleistungen lassen sich auch durch neue Daten aufwerten. Diesen Weg gehen beispielsweise Reiseunternehmen, die detaillierte Wetterprognosen integrieren, um ihre Marketingaktivitäten sowie die Auslastung ihrer Reiseziele zu optimieren. Ein weiteres Beispiel bildet das Verkehrsmanagement in Metropolregionen über Mautsysteme, die den Verkehrsfluss über Preisanpassungen steuern.

3.3.4 Durchbruch

Die »Durchbruchs«-Strategie stellt die Königsklasse bei der Entwicklung neuer Geschäftsmodelle für die Daten-Wirtschaft dar. Hier werden auf Basis der Sammlung und Digitalisierung neuer Datenbestände neue Produkte und Services erschaffen. Beispiele hierfür sind das Energiedaten-Startup Enercast, das ortsbezogene Leistungsprognosen für die Betreiber von Solar- und Windparks anbietet. Auch die digitale Kartographie von Städten á la Google Streetview schafft vollkommen neue Services z. B. für die Hotellerie und Immobilienwirtschaft. Ebenso lassen sich die Geschäftsmodelle der Social Media Monitoring Provider zu den »Durchbruchs«-Modellen zählen, die Feedback und Einstellungen der Nutzer zu verschiedenen Themen, Produkten und Markten zu aussagekräftigen Online-Analysen verdichten.

3.4 Beispiele für neue und optimierte Business-Modelle

3.4.1 Vernetzte Fahrzeuge – Verkehr und Diagnose

Branche	Automobilwirtschaft
Grundlegende Technologien	Wetter- und Datensensoren, mobile On-Board Unit oder Mobile Apps
Geschäfts-modell	Monetarisierung und Aufwertung (vgl. Unterabschnitte 3.3.2 sowie 3.3.3)

Herausforderung

In der wettbewerbsintensiven Automobilindustrie bieten ausgezeichnete Produkte und starke Marken keine Garantie mehr für den langfristigen Fortbestand eines Unternehmens. Die Hersteller versuchen, die Kunden mit neuartigen Services zu überzeugen, die mit Big-Data-Technologien entwickelt werden (vgl. Abbildung 4). Die Grundlage dafür entsteht mit der Digitalisierung und Vernetzung moderner Fahrzeuge – sie sorgen für eine sprunghafte Zunahme auswertbarer Daten: Die mehr als 40 Millionen Fahrzeuge auf Deutschlands Straßen können eine Vielzahl von Bewegungs-, Zustands-, Verschleiß- und Umgebungsdaten melden, die für Fahrer, Reparaturwerkstätten, Hersteller und zahlreiche weitere Organisationen von großem Interesse sind. Aus der Datenflut ergeben sich neue Geschäftsmodelle, die durch den Einsatz von Big Data verwirklicht werden können.

Big Data für die Produktoptimierung

Kunden äußern sich über Social-Media-Kanäle über ihr Fahrzeug. Die gezielte Auswertung solcher Informationen liefert Hinweise darauf, wie die Kundenerwartungen besser zu treffen sind, um so die Kundenzufriedenheit zu erhöhen.

Die Produktoptimierung durch Einsatz von Big-Data-Analysen kann bereits in der Produktentwicklung ansetzen. Ein Beispiel ist die Auswertung von Versuchsdaten, die mit Bestandsdaten aus vergangenen Erprobungen in Beziehung gesetzt werden. So lassen sich die Ursachen von Situationen analysieren, die in irgendeiner Hinsicht auffällig sind. Im Versuch werden Messdaten sowie Anmerkungen der Prüfingenieure automatisiert ausgewertet und unternehmensweit mit ähnlichen Erfahrungsbildern abgeglichen. Im Ergebnis werden Fehler vermindert und Rückrufaktionen vermieden.

Big Data in der Kundenbetreuung

Im Bereich Kundenbetreuung (Aftersales) und bei der Garantieabwicklung kann Big Data dabei helfen, die Wartungsintervalle zu optimieren. Dafür bietet sich die detaillierte anonymisierte Analyse des Nutzungsverhaltens der Kunden an. Werkstattbesuche können so bedarfsgerecht festgelegt und Kunden individuell werblich angesprochen werden. Big Data eignet sich ebenfalls für die Garantieabwicklung und die Überwachung von Anomalien. Bei Abweichungen von einer Norm kann auf empfehlenswerte Serviceaktionen hingewiesen werden.

Insgesamt unterstützt der professionelle Umgang mit Nutzungs- und Servicedaten das Management der After-Sales-Prozesse. Kundenzufriedenheit, Werkstattauslastung und die Wahrscheinlichkeit von Wiederholungskäufen nehmen zu.

Weitere Anwendungsfelder

Für vernetzte Fahrzeuge sind viele weitere Services denkbar. Dazu gehören

- Carsharing,
- nutzungsbasierte Gestaltung und Abrechnung von Versicherungsleistungen nach dem PAYD-Grundsatz[11],
- ortsbasierte Dienste wie die Bereitstellung lokaler Wetterinformationen,
- Warnungen vor Streckengefahren (Unfall, Glatteis,...),
- Hilfestellungen bei Parkplatzsuche und –reservierung,
- Intermodale Navigation unter Einbeziehung anderer Verkehrsmittel.

Vorstellbar ist auch die Verknüpfung von Daten über Verkehrssituationen mit ortsbasierten Serviceangeboten.

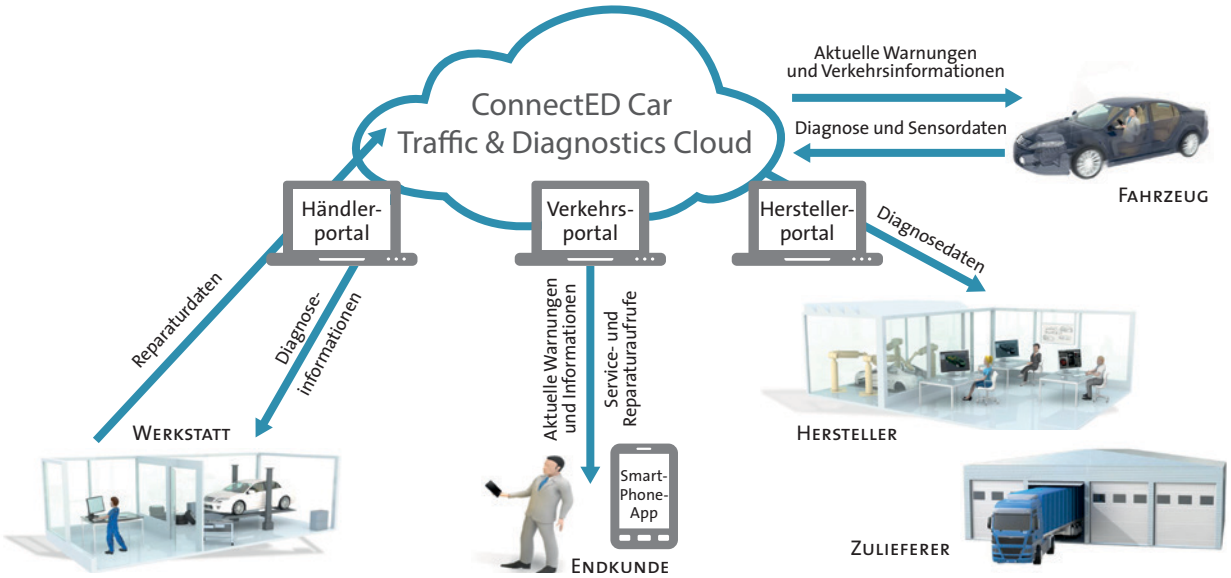


Abbildung 4: Technische Lösung für den Big-Data-Einsatz in der Automobilbranche

[11] Pay as you drive (PAYD) – Typ einer Kfz-Haftpflichtversicherung, bei der die Prämienhöhe von der Intensität der Fahrzeugnutzung und der Fahrweise abhängig ist.

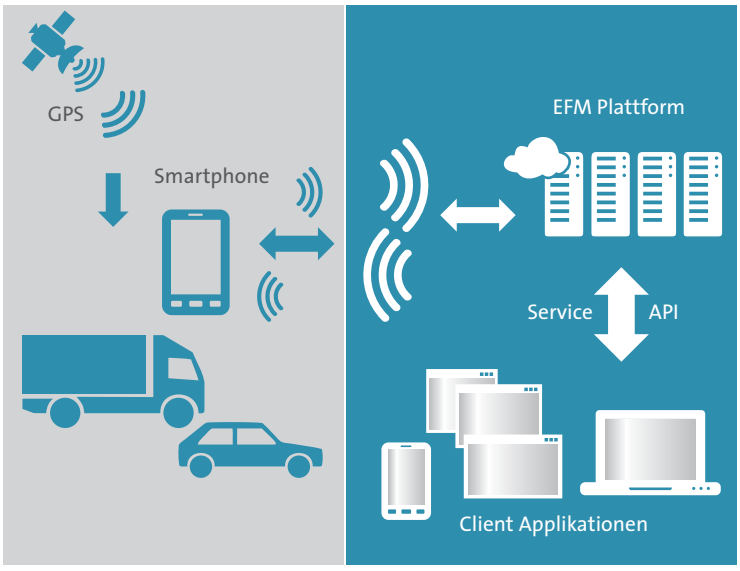
3.4.2 Effizientes Fahrzeugflotten-Management

Branche	Logistik sowie Betreiber großer Fahrzeugflotten
Grundlegende Technologien	Mobiles Endgerät (GPS) mit Windows-Betriebssystem, Plattform über Cloud, OEM-neutrale Hardware
Geschäfts-modell	Monetarisierung (vgl. Unterabschnitt 3.3.2)

Herausforderung

Mehr als 50 Millionen Kraftfahrzeuge rollten 2011 über Deutschlands Straßen. Sie verursachten laut ADAC 189.000 Staus, die sich auf rund 450.000 Staukilometer summierten. Für Transport- und Logistikunternehmen sind die Staus mit hohem Zeit- und Geldverlust verbunden – und mit verärgerten Kunden, wenn deren Pakete nicht pünktlich beim Adressaten eintreffen.

DB Schenker zum Beispiel ist mit 95.000 Mitarbeitern an rund 2.000 Standorten in 130 Ländern vertreten und transportiert allein im europäischen Landverkehr mehr als 95 Millionen Sendungen pro Jahr. Die steigende Arbeitsbelastung und der Stress sowie der wachsende Preisdruck und steigende Energiekosten entwickeln sich nicht nur für die Fahrer selbst und DB Schenker zum Problem. Zusätzlich verlangt der Gesetzgeber mit Emissionsschutzauflagen bis zum Jahre 2020 eine Reduktion der CO_2-Emissionen um 30 Prozent. Um in diesem Umfeld wettbewerbsfähig und erfolgreich zu sein, sind Transparenz über Verbräuche, Kosteneffizienz und Umweltverträglichkeit pro Fahrzeug und Fahrt sowie die sofortige Beeinflussung des Fahrverhaltens in Echtzeit notwendig. Aktuelle Verbrauchsinformationen und Emissionsdaten stehen allerdings nicht zur Verfügung, so dass sich das Fahrverhalten nicht unmittelbar beeinflussen lässt und der Verbrauch nicht sinkt.



- Mobiles Endgerät übermittelt Logistikdaten in Echtzeit (Mobilfunk, GPS)
- Zentrale EFM-Plattform analysiert Daten, errechnet kritische Parameter wie Zuladung/Fracht, aktuelle/kumulierte CO2-Emissionen, momentaner/kumulierter Kraftstoffverbrauch, Position, Höhe, Steigung und Gefälle
- Mandantenfähiges System
- Zentrale Profilverwaltung pro Fahrzeug mit Normwerten
- Dashboard für Abweichungen
- Visualisierung mittels Google-Maps
- OEM-neutrale Hardware

Abbildung 5: Big-Data-Lösung zum Management großer Fahrzeugflotten

Big Data für die Produktoptimierung

Die in Abbildung 5 dargestellte skalierbare Transport- und Logistiklösung MLCM (Mobile Logistics Cost Management) könnte die Probleme der Branche lösen. In einem Pilotversuch der DB Schenker in China fahren in 50 Trucks Smartphones mit, die mit dem mobilen Windows-Betriebssystem laufen. Auf dem Smartphone läuft eine App, die via GPS Ort, Geschwindigkeit, Beschleunigung und Fahrzeiten einschließlich Stillstand ermittelt. Unter Berücksichtigung der Ladung und des Fahrzeuggewichts kann anhand dieser Daten mit Hilfe von Referenzprofilen direkt der Kraftstoffverbrauch berechnet und dem Streckenprofil jeder Route zugeordnet werden. Sekundenschnell gelangen die Daten über einen Server parallel an das Logistikunternehmen und den jeweiligen Fahrer. Disponent und Fahrer erhalten so ein detailliertes Bild über Streckenabschnitte mit erhöhtem Verbrauch. Erste Ergebnisse des Pilotversuchs zeigen, dass sich mit dem Echtzeit-Monitoring der Gesamtflotte der jährliche Kraftstoffverbrauch um rund vier Prozent pro Fahrzeug senken lässt – bei einer Fahrzeugflotte mit mehreren 1000 Transportern ein nicht unerheblicher Betrag.

Nutzen von Big Data im Bereich Logistik

Logistik-Dienstleister und Flottenbetreiber sehen sich einem wachsenden Druck ausgesetzt, den fahrzeugbezogenen Schadstoffausstoß zu minimieren und transparent darzustellen. Die Reduktion von Leerfahrten und Umwegen sowie eine kraftstoffsparende Fahrweise reduzieren die Betriebskosten und stärken die Wettbewerbsfähigkeit. Die Lösung basiert auf Echtzeit-Übermittlung von Bewegungsdaten, ihrer schnellen Auswertung und auf Realtime-Vorschlägen zur Optimierung von Strecke und Fahrweise. Damit können Fahrzeuge ökonomisch und ökologisch sinnvoll Termine einhalten.

3.4.3 Big Data für Energieversorgungsunternehmen

Branche	Energie
Grundlegende Technologien	Smart Metering Plattform (IP-Netz, MDS-System), MDM-Metering-Portal (EVU Branchen-template SAP)
Geschäfts-modell	Monetarisierung und Aufwertung (vgl. Unterabschnitte 3.3.2 sowie 3.3.3)

Herausforderung

Die Energiebranche steht mitten im Wandel. Die Umstellung auf erneuerbare Energien ist eines der größten Infrastrukturprojekte seit der industriellen Revolution. Deutschland hat sich zum Ziel gesetzt, im Jahr 2020 rund 35 Prozent der Stromerzeugung mit Hilfe von Wind-, Wasser- und Solarkraft sowie Biomasse zu gewinnen. Gleichzeitig fordert eine EU-Richtlinie, dass Deutschland bis 2022 rund 44 Millionen Smart Meter installieren muss. Schon seit Januar 2010 müssen in Gebäuden mit neuem Anschluss an das Energieversorgungsnetz oder bei einer größeren Renovierung Zähler verwendet werden, die den tatsächlichen Energieverbrauch und die tatsächliche Nutzungszeit widerspiegeln. Diese Smart Meter sind auch für Verbraucher mit einem Jahresverbrauch von mehr als 6000 kW/h relevant.

Auch die Beziehung zu den Kunden wird sich für die Energieversorger massiv verändern. Kunden sollen über ihren Energieverbrauch häufiger als bisher informiert werden und die Abrechnung soll mehrmals pro Jahr erfolgen. Weiterhin sollen Verbraucher die Möglichkeit erhalten, individuelle Tarife nutzen zu können, damit sie ihren Verbrauch auf unterschiedliche Preise anpassen können.

Aus diesen Veränderungen ergeben sich für die Energiebranche ganz neue Herausforderungen an die IT-Infrastruktur. Bisher waren die IT-Prozesse auf die zentrale Erzeugung und Versorgung der Kunden abgestimmt.

Der Strom floss nur in eine Richtung. Jetzt werden Verbraucher zu Erzeugern und die Energie wird zunehmend dezentral in das Stromnetz eingespeist. Damit benötigen die Stromnetzbetreiber aktuelle und deutlich mehr Daten, um das Netz stabil halten zu können. Mit jedem Smart Meter fließen zudem Daten zu den Energieversorgern, die diese für Verbrauch und Abrechnung in Echtzeit verarbeiten müssen.

Auf die Verarbeitung dieser Datenvolumina sind die IT-Systeme der Versorger noch nicht eingestellt. Dabei ist ein effizientes, integriertes Datenmanagement für die Erfüllung aller Anforderungen unentbehrlich. Ein Energieversorger mit 200.000 Zählern muss gegebenenfalls bis zu 200.000 Telekommunikationsverträge bestellen, einrichten und verwalten. Ein wichtiger Ansatz besteht darin, die Daten bedarfsgerecht zur Verfügung zu stellen.

Dies lässt sich nur über Big-Data-Lösungen erreichen, da schon bei einem übertragenen Tageswert pro Zähler täglich 4,8 Millionen Datensätze anfallen, jährlich über 1,75 Milliarden. Die IT-Abteilung des Energieversorgers speichert und verwaltet hierfür jährlich ca. 1.000 Terabyte an Datenvolumen.

Big Data für die Produktoptimierung

Big Data für Energieerzeuger kann solche Datenvolumina speichern und verarbeiten. Als Bindeglied zwischen den Systemen zur Zählerdatenerfassung und -übermittlung sowie den nachgelagerten betriebswirtschaftlichen Systemen der verschiedenen Akteure ermöglicht Meter Data Management die sichere und schnelle Speicherung und Verarbeitung von Massendaten.

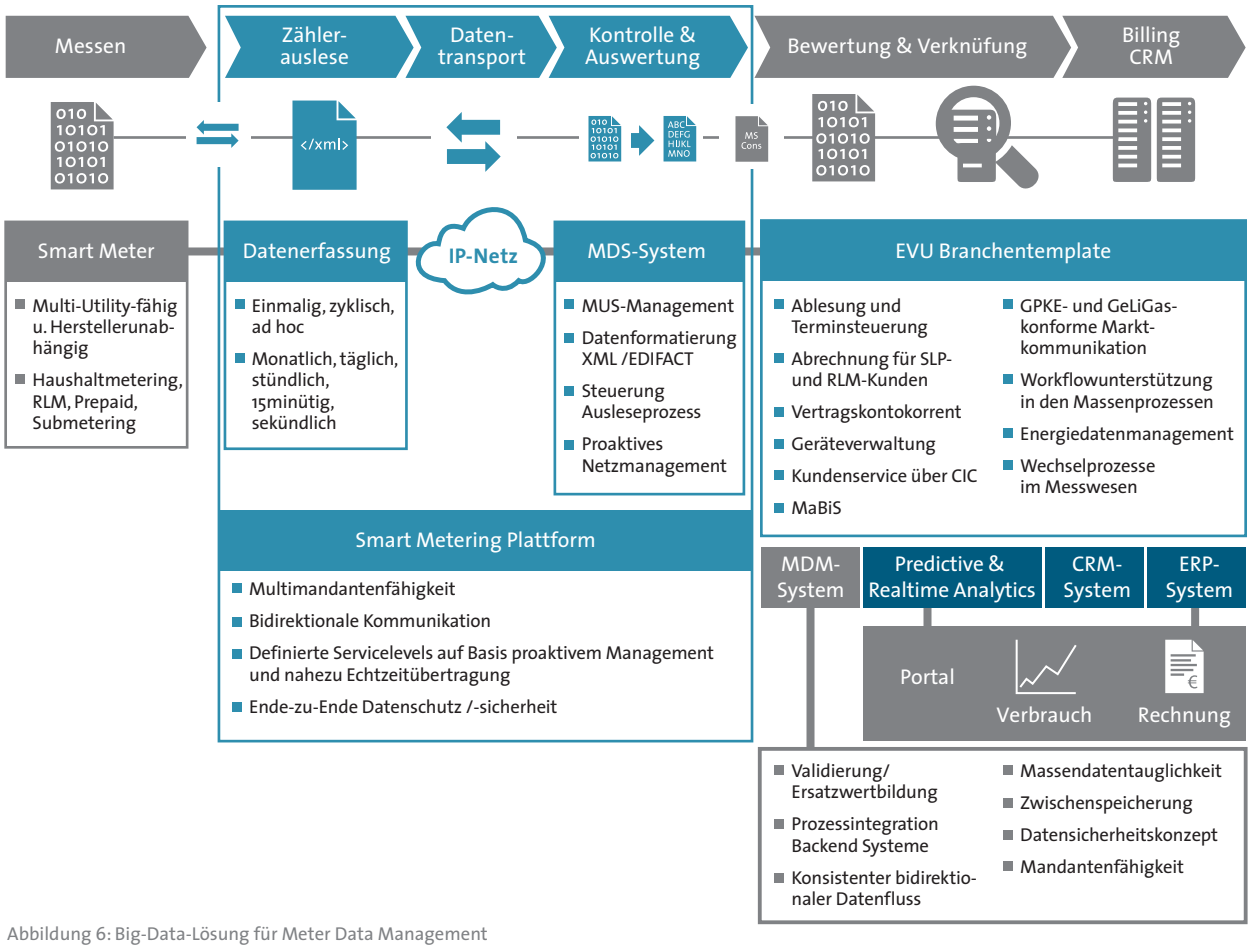


Abbildung 6: Big-Data-Lösung für Meter Data Management

MDM bietet mehrere Vorteile:

- Es ist als leistungsfähiges, zentrales IT-System für Datenvolumina im Petabyte-Bereich (Big-Data-Management) ausgelegt.
- Es ermöglicht einen hoch performanten und sicheren Datenaustausch zwischen Erfassungsort und Rechenzentren über mehrfach gesicherte Datennetze – Mobil- oder Festnetz.
- Es unterstützt alle relevanten, technischen Standards und lässt sich daher einfach in bestehende IT-Systeme integrieren, etwa in SAP-Systeme.
- Es ist offen für zukünftige Systementwicklungen, zum Beispiel im Rahmen von intelligenten Stromnetzen – Smart Grids.
- Es berücksichtigt aller rechtlichen Rahmenbedingungen und Besonderheiten des Datenschutzes.
- Es visualisiert als modular aufgebaute Komplettlösung die Messwerte verschiedener Zähler- und Medientypen, indem es Zählerdaten geschützt und formatunabhängig zur Verfügung stellt.

Nutzen von Big Data im Bereich Energiewirtschaft

Durch die automatisierte Übertragung und Analyse der Massendaten fließen die Daten direkt in das Billing-System der Unternehmen ein. Damit können die Energieversorgungsunternehmen die rechtlichen Vorgaben erfüllen. Sie verpflichten Stromerzeuger dazu, individuelle Tarife anzubieten, was von Kunden gewünscht wird. Zudem ermöglicht die Echtzeitverarbeitung der Massendaten den Blick auf den aktuellen Verbrauch, was Einsparungspotenziale aufzeigt.

Mit den Echzeitanalysen des Stromverbrauchs bekommen die Energieversorger auch ein Instrument an die Hand, die Stromerzeugung zeitnah zu steuern. Durch den hohen Anteil dezentral erzeugter Energie kommt es in den Stromnetzen vermehrt zu starken Spannungsschwankungen. Da die Stromerzeugung heute auf Erfahrungswerten basiert, stehen mit den Big-Data-Analysen genauere Daten zur Verfügung, um daraus bedarfsgerecht Energie bereitzustellen, Prognosen in Echtzeit und M2M-Kommunikation zu ermöglichen.

Eine medienbruchfreie Kommunikation bietet zudem erhebliche Einsparungspotenziale bei den Prozesskosten, so dass nach Angaben von stromanbieter.net (2012) eine Reduktion der Stromkosten von bis zu 23 Prozent möglich wäre.

4 Datenschutz in Big-Data-Projekten

Bei der Entwicklung und bei der Umsetzung von Big-Data-Projekten spielt der Datenschutz insbesondere dann eine wichtige Rolle, wenn personenbezogene Daten verwendet werden sollen. Personenbezogen sind solche Angaben, die persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren Person betreffen.
Die Grundlagen des Datenschutzes sind bereits im ersten Big-Data-Leitfaden des BITKOM[12] beschrieben worden. Im vorliegenden Leitfaden werden zwei Themen in den Mittelpunkt gerückt, die für Big-Data-Projekte eine besonders große Bedeutung haben:

- Privacy Impact Assessments und
- Vorgehen bei Anonymisierung.

4.1 Privacy Impact Assessment

Wenn die Verarbeitung von Daten besondere Risiken für die betroffenen Personen mit sich bringt, will der EU-Gesetzgeber künftig eine Pflicht zur Datenschutz-Folgenabschätzung (Privacy Impact Assessment) schaffen. Im Moment befindet sich der Entwurf der EU-Datenschutzverordnung noch in der Diskussionsphase, und die Regelungen zur Datenschutz-Folgenabschätzung[13] werden in dem einen oder anderen Punkt sicher noch geändert. Man kann aber davon ausgehen, dass die Datenschutz-Folgenabschätzung zur Pflicht wird. Gerade im Bereich von Big-Data-Anwendungen werden solche Privacy Impact Assessments große Bedeutung erlangen.

4.1.1 Anwendungsfälle für ein Privacy Impact Assessment

Die Datenschutz-Folgenabschätzung dient dem präventiven Schutz personenbezogener Daten, also z. B. Name, E-Mailadresse oder Bankdaten einer natürlichen Person. Unternehmen sollen dann verpflichtet sein, eine Folgenabschätzung vorzunehmen, wenn die Verarbeitung der Daten auf Grund ihres Wesens, ihres Umfangs und ihrer Zwecke konkrete Risiken für Rechte und Freiheiten betroffener Personen birgt.

Der Entwurf der Datenschutzverordnung nennt auch konkrete Anwendungsfälle: Die systematische und umfassende Auswertung persönlicher Aspekte, beispielsweise die Analyse der wirtschaftlichen Lage, des Aufenthaltsorts, des Gesundheitszustands, der persönlichen Vorlieben, der Zuverlässigkeit und des Verhaltens einer Person, wenn dies als Grundlage für Maßnahmen mit Rechtswirkung dient oder sonst erhebliche Auswirkungen haben kann. Als zweiten Anwendungsfall sieht der Entwurf die Verarbeitung von Gesundheits- oder anderen sensitiven Daten in großem Umfang, wenn damit Maßnahmen mit Bezug auf einzelne Personen vorbereitet werden. Gleiches soll gelten, wenn umfangreiche Dateien mit Daten über Kinder, genetischen oder biometrischen Daten verarbeitet werden. Außerdem will der Gesetzgeber den Datenschutzbehörden gestatten, selbst eine Liste von Verarbeitungsvorgängen aufzustellen, bei denen eine Datenschutz-Folgenabschätzung stattfinden soll.

4.1.2 Vorgehen bei einem Privacy Impact Assessment

Was den Inhalt betrifft, so enthält der Verordnungsentwurf nur allgemeine Vorgaben für die Datenschutz-Folgenabschätzung. Sie muss eine allgemeine Beschreibung der geplanten Verarbeitungsvorgänge enthalten

[12] Vgl. [BITKOM, 2012], S. 43 ff.
[13] Art. 33 und 34 des Entwurfs

sowie eine Bewertung der Risiken für Rechte und Freiheiten der betroffenen Personen und der geplanten Abhilfemaßnahmen und der Vorkehrungen, um den Datenschutz sicherzustellen und die Nachweise dafür zu erbringen. Die Checkliste im Abschnitt 10.1 gibt einen Überblick.

Die Verordnung sieht außerdem vor, dass das verantwortliche Unternehmen die Zustimmung der betroffenen Personen zu der beabsichtigten Verarbeitung einholt. Gerade bei Big-Data-Verfahren, die eine Vielzahl von Daten verarbeiten, dürfte dies in der Praxis schwierig sein, ganz abgesehen davon, dass zunächst festzulegen sein wird, wie eine solche Meinung einzuholen ist und welchen qualitativen und quantitativen Umfang diese haben muss.

4.1.3 Rechtsfolgen

Die Datenschutz-Folgenabschätzung hat unterschiedliche Rechtsfolgen:

- Konsultationspflicht,
- Vorlagepflicht und
- Genehmigungspflicht.

Konsultationspflicht

Die wichtigste Rechtsfolge ist die Pflicht des Unternehmens, vor Beginn der Datenverarbeitung die zuständige Aufsichtsbehörde zu Rate zu ziehen, damit die rechtlichen Vorgaben eingehalten und bestehende Risiken gemindert werden. Diese Konsultationspflicht besteht dann, wenn sich aus Folgenabschätzung hohe konkrete Risiken für die Rechte der Betroffenen ergeben oder wenn die Datenschutzbehörde in der obengenannten Liste der Verarbeitungsvorgänge eine solche Konsultation vorsieht.

Im Gesetzgebungsverfahren ist an dieser Pflicht zur Konsultation erhebliche Kritik geübt worden; der Innenausschuss des Europaparlaments hat vorgeschlagen, dass Unternehmen auch ihren betrieblichen Datenschutzbeauftragten konsultieren können, entsprechend der bisherigen Regelung im deutschen Recht in § 4 d Abs. 5 BDSG.

Vorlagepflicht

Während die Konsultationspflicht nur besteht, wenn hohe Risiken für die Betroffenen drohen, sollen die Unternehmen in allen Fällen verpflichtet sein, die Datenschutz-Folgenabschätzung der zuständigen Aufsichtsbehörde vorzulegen und ihr Informationen zu erteilen.

Genehmigungspflicht

In bestimmten Fällen im Zusammenhang mit dem Export von Daten aus dem Europäischen Wirtschaftsraum[14] heraus sieht der Entwurf der Verordnung vor, dass zunächst die Genehmigung der Datenschutzbehörde eingeholt wird.

Folgen von Verstößen

Bei Verstößen gegen die Pflicht, eine Folgenabschätzung durchzuführen oder fehlender Konsultation der Datenschutzbehörde drohen Bußgelder bis zur Höhe von 1 Mio. EUR oder 2 % des Unternehmensumsatzes[15].

4.1.4 Kritikpunkte

Insgesamt ist einzuschätzen, dass der europäische Gesetzgeber ein wichtiges Ziel verfolgt, wenn er personenbezogene Daten präventiv schützen will. Es bestehen jedoch zumindest drei Kritikpunkte:

- Der bürokratische Aufwand durch dieses Verfahren wird sich erheblich erhöhen, und das gilt gerade bei der Nutzung von Big Data.

- Es ist zu wünschen, dass der Gesetzgeber die Möglichkeit schafft, die Folgenabschätzung mit dem

[14] EU-Staaten sowie Island, Liechtenstein und Norwegen
[15] Art. 79 Nr. 6 lit.

betrieblichen Datenschutzbeauftragten durchzuführen, da dieser mit den Abläufen im Unternehmen besser vertraut ist.

- Außerdem ist zu bemängeln, dass die EU-Kommission die Fälle nur unpräzise beschreibt, in denen eine Folgenabschätzung stattfinden muss, und den Datenschutzbehörden auch noch die Befugnis geben will, diese Fälle auszudehnen. Das macht es Unternehmen schwer, die Pflichten zu erkennen und einzuhalten.

4.1.5 Privacy Impact Assessment – Checkliste

Eine PIA-Checkliste ist im Anhang zu diesem Leitfaden enthalten (vgl. 10.1).

4.2 Anonymisierung und Pseudonymisierung

Nach dem Bundesdatenschutzgesetz (BDSG) gilt in Bezug auf die Verarbeitung von personenbezogenen Daten ein sogenanntes Verbot mit Erlaubnisvorbehalt, das heißt, personenbezogene Daten dürfen nur dann erhoben, verarbeitet oder gespeichert werden, wenn dies gesetzlich ausdrücklich erlaubt ist oder der Betroffene darin eingewilligt hat. Personenbezogen sind nach der Gesetzesformulierung solche Angaben, die persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren Person betreffen, also beispielsweise der Name, die E-Mail-Adresse, Kreditkartendaten oder das Geburtsdatum einer Person.

Im Rahmen von Big-Data-Projekten ist daher zunächst zu prüfen, ob in Bezug auf die personenbezogenen Daten, die genutzt werden sollen, entsprechende Einwilligungen der betroffenen Personen vorliegen. Eine solche Einwilligung kann z. B. in Kundenerklärungen zu Werbeeinwilligungen liegen. Eine gesetzliche Erlaubnis für die Datennutzung kann sich aus §§ 28 ff BDSG ergeben. Hier ist es ratsam, im Rahmen des Privacy Impact Assessments eine sorgfältige Prüfung vorzunehmen.

Eine Datennutzung ist darüber hinaus auch dann möglich, wenn anonymisierte Daten genutzt werden.

4.2.1 Anonymisierung

Daten gelten gemäß § 3 Abs. 6 BDSG als anonymisiert, wenn die Einzelangaben nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer Person zugeordnet werden können. Es ist also nicht erforderlich, dass die Rückgängigmachung der Anonymisierung für jedermann unmöglich oder unverhältnismäßig ist.[16]

Ausreichend ist vielmehr, dass die verantwortliche Stelle keine realistische Möglichkeit zur De-Anonymisierung hat. Auch wenn einzelne Aufsichtsbehörden dies anders sehen, so liegt dem BDSG doch der Begriff der relativen Anonymität zugrunde: Wenn also ein Dritter mit entscheidenden Zusatzwissen bestimmte Daten einer bestimmten Person zuordnen kann, folgt daraus nicht zwangsläufig die Personenbezogenheit dieser Daten.

Ursprünglich personenbezogene Daten können dadurch anonymisiert werden, dass Identifikationsmerkmale gelöscht oder bestimmte Merkmale aggregiert werden. Aggregation von Merkmalen heißt, dass exakte Angaben durch allgemeinere ersetzt und die Daten dann zusammengefasst werden: Beispielsweise eine Gruppenbildung anhand des Geburtsjahres anstelle des genauen Geburtsdatums oder anhand einer weiträumigen Gebietsangabe anstelle der Adressangabe.

Eine Anonymisierung kann auch dadurch vorgenommen werden, dass aus einem Bestand personenbezogener Daten einzelne Angaben ohne Personenbezug herausgefiltert werden. Sie können dann isoliert für statistische oder planerische Zwecke verwendet oder weitergegeben werden, ohne dass auf die Vorgaben des BDSG Rücksicht

[16] Vgl. [RoScho, 2000]

genommen werden muss. Dies gilt allerdings nur, wenn nicht zu befürchten ist, dass die Daten später wieder zusammengeführt werden, was im Einzelfall auch vertragliche Regelungen erforderlich machen kann.

4.2.2 Pseudonymisierung

Die Anwendung des BDSG kann auch durch eine Pseudonymisierung ausgeschlossen werden. Hierbei wird der Name und etwaige andere Identifikationsmerkmale durch ein Kennzeichen ersetzt, um die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren[17]. Mit anderen Worten: Bei der Anonymisierung werden Identifikationsmerkmale gelöscht, bei der Pseudonymisierung nur ersetzt.

Dort wo das Gesetz ausdrücklich eine Anonymisierung verlangt[18], genügt eine Pseudonymisierung nicht. Die Pseudonymisierung ist immer dort anzuwenden, wo – im Sinne der Datensparsamkeit nach § 3 a – die Kenntnis der Identität des Betroffenen nicht notwendig ist.

Durch die Pseudonymisierung wird die unmittelbare Kenntnis der Identität des Betroffenen für die Vorgänge ausgeschlossen, bei denen der Personenbezug nicht zwingend erforderlich ist, vor allem in Wissenschaft und Forschung.

Bei der Pseudonymisierung gibt es grundsätzlich zwei Verfahrensarten:

- Erstens die Erzeugung von Zufallswerten und deren Zuordnung zum Betroffenen mittels einer Referenzliste. Für den Inhaber dieser Liste bleiben die Daten dann personenbezogen.

- Zweitens die Erstellung von Pseudonymen durch Hash-Verfahren mit geheimen Parametern. Wenn die Rückrechnung der ursprünglichen Daten mit sehr hohem Aufwand verbunden ist, spricht man auch von einer Einweg-Pseudonymisierung.

4.2.3 Anonymisierung und Auswertung von Daten

Für ein wirksames Anonymisierungsverfahren ist eine sinnvolle Kombination aus technischen und organisatorischen Maßnahmen, die auch im anonymisierten Zustand Bezugsmöglichkeiten zwischen den einzelnen Datensätzen zulässt, von entscheidender Bedeutung.

Im Fall von Standortdaten können

- ortsbasierte Filter sowie
- ein regelmäßiger Wechsel des Anonymisierungsschlüssels

direkte bzw. indirekte Rückschlüsse auf einzelne Individuen anhand der Datenbasis effektiv verhindern.

Trotz dieser datenschutzrechtlich notwendigen Einschränkungen können auf Grundlage und unter Hinzuziehung von Wahrscheinlichkeitsberechnungen dennoch wertvolle Langzeitaussagen getroffen werden. Das ändert nichts an der wirksamen Anonymisierung und verbessert gleichzeitig die Möglichkeiten zur Nutzung der Daten. Das Praxisbeispiel eines wirksamen Anonymisierungsverfahrens im Anhang[19] zeigt: Es ist möglich, unter Einbehaltung aller datenschutzrechtlichen Bestimmungen eine genügend umfassende Datenbasis für vielfältige (potenzielle) Big-Data-Anwendungen zu realisieren.

4.2.4 Anonymisierung und TK-Recht

Die Reichweite des Begriffs Anonymisierung ist über das BDSG hinaus relevant, etwa für den Umgang mit Standort[20] – und Verkehrsdaten durch TK-Unternehmen.

---

17 §3 Abs. 6a BDSG
18 §§ 4d Abs. 4 Nr. 2, 4 f Abs. 1, Satz 5, 30, 40
19 Vgl. Abschnitt 10.4
20 Standortdaten geben den Standort eines Mobiltelefons oder Smartphones an.

Die angemessene Behandlung dieser Daten wirft neuerdings Fragen auf. Speziell entwickelte Anwendungen können die Ortungsdaten der Nutzer erfassen und anonymisieren. So anonymisierte Daten können dann Hinweise auf Kundendatenströme geben. Mögliche Interessenten solcher Daten sind beispielsweise Einzelhandelsunternehmen. Auch das gezielte Versenden von Werbenachrichten ist mithilfe solcher Daten und entsprechenden Programmen prinzipiell möglich.

Mit Verkehrs- oder Standortdaten verhält es sich aber wie mit personenbezogenen Daten im Allgemeinen: TK-Dienstleister dürfen nach dem TKG Standort- und Verkehrsdaten nur dann verwenden oder verarbeiten, wenn eine besondere Rechtfertigung vorliegt.

Auch hier dürfte bei Big Data häufig als einzige realistische Rechtfertigungsmöglichkeit die Anonymisierung der Daten verbleiben. Es kommt also wiederum darauf an, ob der Bezug zwischen Daten und Person ohne unverhältnismäßig großen Aufwand hergestellt werden kann.

Anonymisierung bei Weitergabe der Daten

Bei der Weitergabe der Daten an interessierte Dritte wird teilweise die Auffassung vertreten, dass hier eine Anonymisierung gar nicht möglich sei, da die Drittunternehmen eventuell über eigene Datenbestände verfügten, die es dem Drittunternehmen erlaubten, die betroffene Person zu identifizieren. Diesbezüglich ist anzumerken, dass die Unverhältnismäßigkeit des Aufwands zur De-Anonymisierung nur anhand der Umstände des Einzelfalls und nicht pauschal beurteilt werden kann. Wenn Fallkonstellationen denkbar sind, in denen die Zusammenführung von Datenbeständen zur Wiederherstellung eines Personenbezugs ausreicht, bedeutet dies nicht, dass die Anonymisierung[21] generell unmöglich wäre.

Die beteiligten Unternehmen müssen sich dieser Möglichkeit bewusst sein und Maßnahmen ergreifen, die eine De-Anonymisierung verhindern. Es liegt daher immer in der Verantwortung der beteiligten Unternehmen, die Einhaltung des Datenschutzrechts zu gewährleisten. Ausreichend flankiert ist diese Pflicht von den bußgeld- und strafrechtlichen Vorschriften des BDSG beziehungsweise TKG.

[21] Das Bundesamt für Sicherheit in der Informationstechnik (BSI) hat in einer Informationsschrift die verschiedenen Grade der Anonymität dargestellt: [BSI, o.J.]

5 Vorgehensmodell zur Umsetzung von Big-Data-Projekten

Wegen der hohen Komplexität von Big-Data-Projekten ist es empfehlenswert, sich bei ihrer Entwicklung und Umsetzung an einem Vorgehensmodell zu orientieren. Es unterstützt Unternehmen dabei, alle Schritte und Prozesse von Big-Data-Projekten transparent und nachvollziehbar zu gestalten. Im Sinne der Nachhaltigkeit ist es wichtig, Big-Data-Projekte von der frühen Planung bis zur mittel- und langfristigen Optimierung durchgängig zu begleiten.
Das in diesem Leitfaden vorgeschlagene Vorgehensmodell umfasst in acht Phasen alle Aktivitäten von der Identifikation möglicher Big-Data-Potenziale, über die konkrete Planung, die Umsetzung einschließlich der Konsolidierung der IT-Infrastruktur und die Erschließung neuer Datenquellen bis hin zum Betrieb und zur Optimierung von Geschäftsprozessen.

5.1 Bedeutung eines Vorgehensmodells

Wegen der hohen Komplexität von Big-Data-Projekten ist es empfehlenswert, sich bei ihrer Umsetzung an einem Vorgehensmodell zu orientieren. Ein Vorgehensmodell unterstützt Unternehmen dabei, alle Schritte und Prozesse bei der Entwicklung und Umsetzung von Big-Data-Projekten transparent und nachvollziehbar zu gestalten.

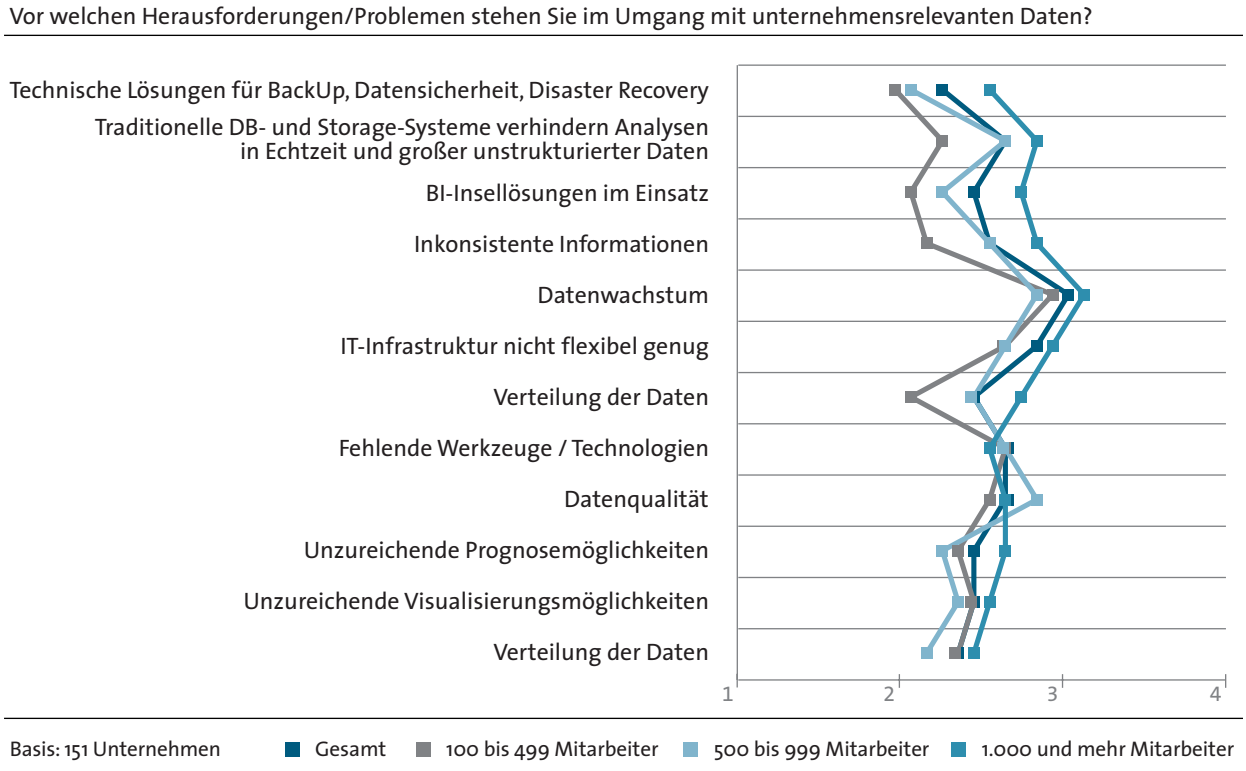


Abbildung 7: Probleme im Umgang mit unternehmensrelevanten Daten – Zielstellungen für Big-Data-Projekte

Ein Vorgehensmodell muss sich dabei an dem konkreten Bedarf des Anwenderunternehmens ausrichten. In Abbildung 7[22] ist dargestellt, welche Probleme beim Umgang mit unternehmensrelevanten Daten gesehen werden. Aus den erkannten Problemen leiten sich Ziele von Big-Data-Projekten ab, die wiederum mit einzelnen Schwerpunkten in den Phasen eines Vorgehensmodells verknüpft werden können.

Das Big-Data-Vorgehensmodell in Abbildung 8 umfasst acht Phasen für die Planung, Umsetzung und Optimierung von Big-Data-Projekten:

- Assessment
- Readiness
- Implementierung und Integration
- Konsolidierung und Migration
- Nutzung der neuen Daten
- Reporting und Predictive Analytics
- End-to-End-Prozesse
- Optimierung.

In Abhängigkeit von der Projektspezifik werden die Phasen unterschiedlich gewichtet sein.

Die einzelnen Phasen des Vorgehensmodells werden im Abschnitt 5.2 kurz beschrieben. Die in den Kapiteln 6 bis 9 dargestellten Empfehlungen, Konzepte und Technologien zur Planung und Umsetzung von Big-Data-Initiativen weisen zumeist einen deutlichen Bezug zu den einzelnen Phasen des Vorgehensmodells auf und konkretisieren es weiter.

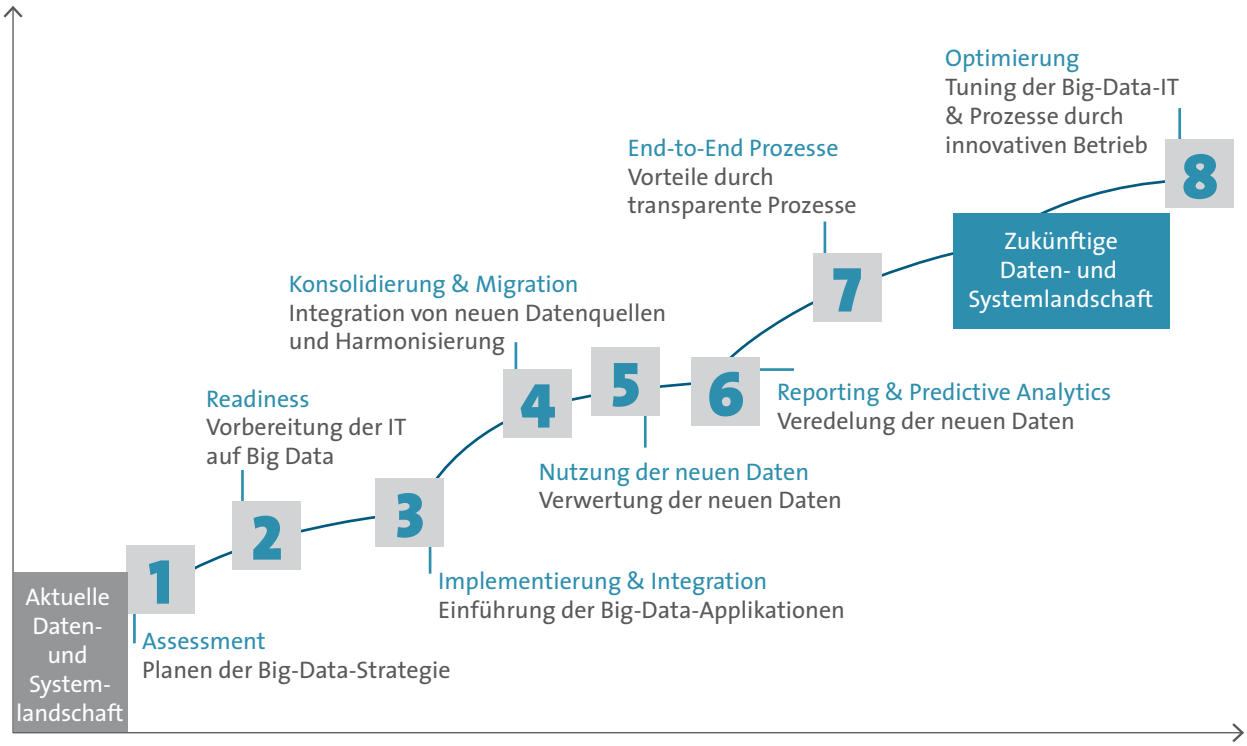


Abbildung 8: Big-Data-Vorgehensmodell

[22] Vgl. [EGAG 2012]

5.2 Phasen des Vorgehensmodells

5.2.1 Assessment

Dem Big-Data-Assessment kommt für die Durchführung der Big-Data-Projekte eine entscheidende Bedeutung zu. Potenziale durch den Einsatz von Big Data werden identifiziert und Herausforderungen zu ihrer Freisetzung ermittelt.

In Abhängigkeit von der Dimension des Projektes und der Situation im Unternehmen kann ein Assessment auch mehrere Wochen dauern. Drei Assessmentstufen werden empfohlen:

- Big-Data-Discovery-Workshop:
  Die wichtigsten Erfahrungen und Fähigkeiten des Unternehmens im Umfeld von Big Data werden aufgenommen, erste Ziele und Möglichkeiten von Big Data erörtert.

- Big-Data-Readiness-Assessment:
  Ziel dieses Ansatzes ist es, für einen bestimmten Unternehmensbereich die Chancen durch Big Data sowie die zu bewältigenden Herausforderungen zu identifizieren. Bestandteil dieses Ansatzes ist die Evaluierung der Big-Data-Maturity, die Aufschluss über den Big-Data-Status und damit wichtige Rückschlüsse über mögliche Big-Data-Ansatzpunkte und Projekte liefert.

- Big-Data-Evaluierungs- und Strategieansatz:
  Aufbauend auf dem Big-Data-Readiness-Assessment werden die identifizierten Potenziale entsprechend ihrer Bedeutung für das Unternehmen priorisiert. Vielversprechende Big-Data-Szenarien werden genauer beleuchtet und frühzeitig auf Machbarkeit geprüft, um dann ein erstes Lösungsdesign, ein Betriebs- und Wartungskonzept sowie Kosten-/Nutzenbetrachtungen zu entwickeln und in der Big-Data-Strategie und Roadmap zu verankern. Ein weiterer wichtiger Aspekt ist das Aufsetzen der Big-Data-Governance. Die zur Umsetzung ausgewählten Szenarien werden abschließend in der Big-Data-Strategie verankert, und die zeitliche Umsetzung wird im Rahmen einer Roadmap detailliert (vgl. Kapitel 6).

Big-Data-Maturity-Modell

Der Stand der Big-Data-Kompetenz (vgl. Kapitel 7), die aktuelle Infrastruktur sowie die Management-Unterstützung (vgl. Kapitel 6) sind maßgebliche Faktoren, ob und wie Big-Data-Applikationen eingeführt bzw. erweitert werden.

Eine erste Einschätzung, welchen Status Big-Data-Konzepte zu einem bestimmten Zeitpunkt in einem Unternehmen aufweisen und worin die Zielsetzung für zukünftige Big-Data-Initiativen bestehen kann, muss in einer Gap-Analyse konkretisiert werden, die während des Readiness-Assessments durchgeführt werden kann.

Grundlage für die Gap-Analyse ist das Big-Data-Maturity-Modell. Es beschreibt den Big-Data-Reifegrad eines Unternehmens. Wesentliche Erfahrungen, Fähigkeiten, Prozesse, Initiativen und Projekte im Umfeld von Big Data werden umfassend durchleuchtet. Auf Basis dieser Ergebnisse werden die Stärken und Schwächen, aber auch die Potenziale des Unternehmens in diesem Umfeld detailliert aufgezeigt.

Realistische Projekte zur Einführung einer Big-Data-Infrastruktur, zur Implementierung von Analytics-Ansätzen, zur Optimierung von Geschäftsprozessen sowie eine Big-Data-Roadmap können mithilfe dieser Ergebnisse vorgeschlagen werden.

Das Big-Data-Maturity-Modell ermöglicht eine erste Bewertung anhand der folgenden Kriterien:

- Big-Data-Kompetenzen
- Big-Data-Infrastruktur (Prozesse, Tools)
- Erfahrungen und erfolgreiche Projekte im Big-Data-Umfeld
- Strategische Verankerung innerhalb des Unternehmens (Big-Data-Competence-Center sowie transparente Big-Data-Governance)

- Gewichtung von Big Data zur Umsetzung der Geschäftsstrategie
- Big-Data-Geschäftsprozess-Architektur und Big-Data-IT-Referenzarchitektur
- Konsolidierung IT-Landschaft und Anbindung an Big-Data-Cloud
- Aktuelle Big-Data-Initiativen
- Big-Data-Roadmap
- Kosten-/Nutzenanalyse auf Basis von Big Data
- Big-Data-Leitprinzipien.

Die Bewertungsergebnisse vermitteln einen Eindruck vom Big-Data-Status innerhalb des Unternehmens oder eines Unternehmensbereichs und werden in einem sechsstufigen Big-Data-Maturity-Modell abgebildet (vgl. Abbildung 9 sowie Anhang 10.2):

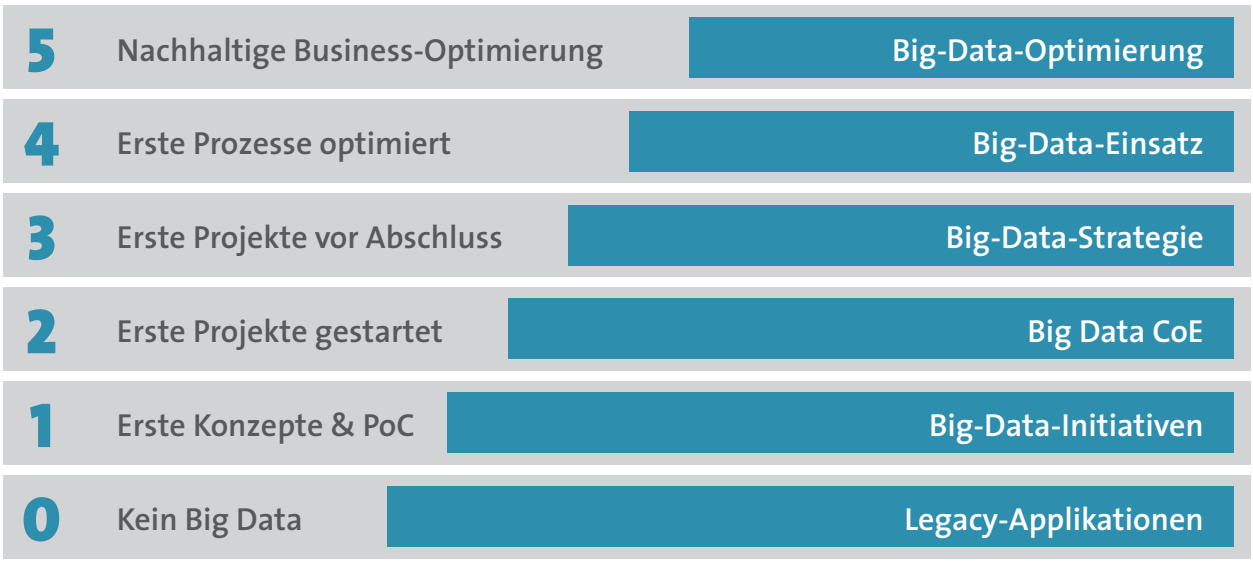


Abbildung 9: Big-Data-Maturity-Modell

Big-Data-Governance

Big-Data-Governance bildet die Voraussetzung, um die Strukturen und Prozesse im Information Management eines Unternehmens transparent und effizient steuern zu können.

Mit der anhaltenden Datenexplosion wird Big-Data-Governance zu einer notwendigen Grundlage für Big-Data-Projekte (vgl. Kapitel 6).

Aus zwei Gründen stellt Big Data eine besondere Herausforderung für eine Governance dar:

- Erstens kommt ein Großteil der neuen und potenziell wertvollen Daten aus externen Quellen.
- Zweitens sind diese Daten überwiegend unstrukturiert. Sie werden im Vergleich zu den internen transaktionalen Daten schlechter erfasst und verstanden.

Da diese Daten zunehmend als Asset verstanden werden, gewinnt Big Data an Bedeutung und damit auch das Verständnis für eine nachhaltige Big-Data-Governance[23]. Big Data ist somit auch eine Chance für eine nachhaltige Information Management Governance.

Eine Möglichkeit zum Aufbau transparenter Strukturen in diesem Bereich besteht in der Etablierung eines Big Data Center of Excellence (CoE)[24].

In der Vergangenheit waren die Aufgaben von Datenqualität, sicherheit und schutz nicht selten unterschiedlichen organisatorischen Einheiten zugeordnet. In Big-Data-Projekten müssen diese Aufgaben zusammengeführt werden.

- Die Datenqualität wird unter anderem durch die Anzahl der Quellen, durch Abstraktions- und Verschiebe-Mechanismen, durch Umspeichern und durch Interpretationen beeinflusst. Für das Qualitätsmanagement sind besondere Tools unerlässlich. Die Mitarbeiter in diesem Bereich müssen kontinuierlich weitergebildet werden, da völlig neue Kompetenzprofile gefordert sind (vgl. Kapitel 7).
- Die Datensicherheit umfasst den Schutz gegen Verlust oder die Verfälschung von Informationen.
- Im Datenschutz (vgl. Kapitel 4) kommt dem Management eine besondere Verantwortung zu: Fehler in diesem Bereich können einen enormen Imageschaden verursachen.

5.2.2 Readiness

Eine Grundlage für die erfolgreiche Umsetzung von Big-Data-Projekten bildet auch der Aufbau der erforderlichen Hardware- und Software-Infrastruktur sowie der benötigten Kompetenzen. Auf der Grundlage der Ziele aus der Assessment-Phase müssen nun die Anforderungen

[23] Folgende Fragen stehen an (Auswahl):
Wie können Big-Data-Daten mit Metadaten beschrieben werden?
- Wie sehen Restart- und Recovery-Verfahren aus?
- Wie geht man mit Datenredundanzen und auslaufender Datenaktualität um?
- Wann müssen Daten geografisch verteilt werden?
- Wie kann man sicherstellen, dass Ereignisströme aus Gründen des Nachweisbarkeit aufbewahrt werden?
- Wer hat Zugriff auf die Daten und wo werden diese Zugriffsrechte geregelt?

Für den Einstieg in Big-Data-Projekten empfiehlt sich eine pragmatische, schrittweise Entwicklung dieser neuen Architektur-Paradigmen, denn die Erfahrungen können die Paradigmen beeinflussen.
In Big-Data-Projekten ist der Daten- und Informations-Lebenszyklus unter Governance-Gesichtspunkten zu bewerten:
- Wer ist für die Daten zuständig (Owner)?
- Wie lange müssen Daten aufbewahrt werden?
- Wie erfolgt der Übergang von Daten in das Archivierungssystem?
- Wann dürfen oder müssen Daten gelöscht werden?
- Welche Datenschutzrichtlinien und -gesetze sind im Aufbau von Big-Data-Lösungsarchitekturen zu berücksichtigen?

[24] Aspekte, die es beim Aufbau eines Big Data CoE zu beachten gilt, sind in der Anlage 10.3 dargelegt.

an die zukünftige Big-Data-Infrastruktur mit der vorhandenen IT abgeglichen werden. Es ergeben sich folgende Fragen:

- In welchen Bereichen weist die IT Lücken auf?
- Welche Technologien eignen sich zur Umsetzung der benötigten Infrastruktur (vgl. Kapitel 6)
- Ist die Netzwerkinfrastruktur und Serverkapazität den neuen und schnell zur Verfügung stehenden Daten gewachsen?
- Sind die Anforderungen an die Datensicherheit erfüllt?
- Sind die aktuellen Software-Releases für die neuen Lösungen und Konzepte ausgelegt?
- Sind die notwendigen Kompetenzen und Ressourcen für die Implementierung und den Betrieb der neuen Big-Data-Landschaft vorhanden (vgl. Kapitel 9)?
- Welche Komponenten der Big-Data-Infrastruktur können aus der Cloud bezogen werden?
- Können mögliche neue Anforderungen an den Datenschutz (vgl. Kapitel 4) erfüllt werden?

Aufgedeckte Defizite müssen umgehend beseitigt werden – spätestens jedoch vor der Einführung neuer Big-Data-Applikationen und -Prozesse. Zusätzlich müssen Grundlagen in den Bereichen Organisation, Prozesse, Compliance und Governance[25] geschaffen werden.

5.2.3 Implementierung und Integration

In dieser Phase geht es um das Design und die Implementierung der Big-Data-Lösung, die auch zukünftigen Anforderungen gewachsen sein sollte.[26] Die Implementierung schließt auch die Integration in die bestehende IT–Landschaft ein. Hier spielt vor allem die Anbindung an Cloud-Lösungen eine wichtige Rolle, da Big-Data-Services zukünftig vor allem aus der Cloud bezogen werden.

Die meisten bisherigen Big-Data-Anwendungen in Unternehmen sind in erster Linie durch

- die Ausweitung von traditionellen Business-Intelligence- und Business-Analytics-Lösungen auf größere Datenmengen oder
- die Speicherung großer Datenmengen und deren schnelle Verarbeitung

charakterisiert. Häufig werden Mischarchitekturen eingesetzt, die klassische Data-Warehouse-Architekturen einbeziehen. Mit neueren Technologien werden riesige Datenvolumina einer Vorfilterung unterzogen, Massendaten analysiert und Daten in Nah-Echtzeit verarbeitet.

Diese Technologien sind Bausteine für die Umsetzung fachbereichs- und themenübergreifender komplexer IT-Szenarien, wie sie beispielsweise

- im Gesundheitswesen (E-Health),
- im Straßenverkehr (interaktive, sekundenaktuelle Verkehrs- und Routenplanung) und
- im Management von Unternehmen (Verknüpfung von BI mit MES, ERP und Social-Media-Informationen)

zu finden sind.

[25] Es ist zu klären, wer für die neuen Daten sowie die Informationsflüsse verantwortlich ist.
[26] Im Abschnitt 8.2.3 wird vertiefend auf die Wege zur Transformation eingegangen.

Die Phasen 3 bis 6 des Vorgehensmodells sind für die Einbindung und Nutzung von analytischen Plattformen (vgl. Kapitel 8) essenziell.

5.2.4 Konsolidierung und Migration

Nach Aufbau und Integration der Big-Data-Lösung gilt es nun, ihre Möglichkeiten voll auszuschöpfen. Entscheidend ist der Fokus, der mit der Big-Data-Lösung verfolgt wird. Geht es darum, die schon vorhandene Infrastruktur zu optimieren oder sollen neue Datenquellen erschlossen werden?

Je nach Anforderungen können Harmonisierungsstrategien sowie Datensimplifikation oder exakte Prozessbeschreibungen zur Einbindung neuer Datenquellen von hoher Bedeutung sein.

Klare Verantwortlichkeiten für die Daten und deren Nutzung sowie Konzepte zur Gewährleistung sowohl der Transparenz in der Datenvielfalt als auch einer adäquaten Datenqualität sind notwendig (vgl. Abschnitt 9.1). Dem Master Data Management kommt eine entscheidende Bedeutung zu.

5.2.5 Nutzung der neuen Daten

Fokus in dieser Phase ist es, die neuen Datenquellen gewinnbringend für das Unternehmen zu nutzen. Die neuen Daten müssen je nach Format zur Verfügung gestellt und aufbereitet werden, um neue Erkenntnisse für die Entscheidungsvorbereitung zu gewinnen. Die Analyse der neuen Daten wird mit den Legacydaten des Unternehmens verknüpft. Die Nutzung der Daten wird unternehmensspezifisch sein; daher sind oft individuelle Analysealgorithmen notwendig. Die Einbindung von analytischen Plattformen (vgl. Abschnitt 8.1) spielt hier eine wesentliche Rolle.

5.2.6 Reporting und Predictive Analytics

Die neuen Daten und die daraus abgeleiteten Erkenntnisse bilden nun die Grundlage zur Optimierung vorhandener Reporting-Prozesse.

Zusätzlich bilden die neu erschlossen Daten – kombiniert mit Legacy-Daten und aufbereitet durch anforderungsspezifische Algorithmen – die Möglichkeit für Prognosen zukünftiger Trends.

Ein wichtiger Aspekt für die Zukunft ist Information Assurance. Aufgrund der enormen Mengen an Daten und ihrem steigenden Wert für Unternehmen muss sichergestellt werden, dass Verluste, Verfälschungen oder Veränderungen der Daten erkannt werden und diesen Gefahren entgegengewirkt werden kann. Weiterhin muss die Validität der Daten sichergestellt sein. Ansätze zum Monitoring der Informationsflüsse sind zu erarbeiten.

5.2.7 End-to-End Prozesse

Big Data ermöglicht es, dass sämtliche Daten zur Abbildung komplexer Geschäftsprozesse in hoher Granularität erfasst und durch neue Daten angereichert werden. Damit kann ein komplettes (End-to-End) Monitoring von Geschäftsprozessen Realität werden, was wiederum Wege zur Optimierung vorhandener und zur Entwicklung vollkommen neuer Geschäftsprozesse öffnet.

5.2.8 Optimierung

Nach dem Aufbau einer Big-Data-Landschaft gelangt ihr zuverlässiger Betrieb und ihre weitere Optimierung in den Fokus. Das Application Management gewährleistet die reibungslose Funktion der Big-Data-Lösungen und trägt zur kontinuierlichen Verbesserung der Betriebskonzepte bei (vgl. Kapitel 9). Weiterhin sind Konzepte notwendig, um die implementierten Lösungen und Prozesse kontinuierlich zu optimieren.

6 Big-Data-Projekte in Unternehmen – Erfolgsfaktoren und Management-Aufgaben bei der Einsatzvorbereitung und Nutzung

Big-Data-Lösungen können dann messbare Beiträge für die Wertschöpfung leisten, wenn die Entscheider eine Reihe von Erfolgsfaktoren beachten. Dazu zählen u.a. die Entwicklung einer wertorientierten Big-Data-Strategie und einer daraus abgeleiteten Roadmap, der Business-Fokus, die Etablierung eines umfassenden Big-Data-Innovationsprozesses, die Auswahl geeigneter Technologien, die regelmäßige Erfolgsüberprüfung sowie die Einrichtung einer passenden Organisation.

Einsatzvorbereitung und Nutzung von Big Data profitieren von Kreativität und neuen Denkweisen, denn bisherige Ansätze zur Berechnung eines Return on Investment stoßen an Grenzen. Wichtig sind Investitionen in Aus- und Weiterbildung sowie die Entwicklung von Keimzellen, in denen Big-Data-Ideen zu Big-Data-Szenarien wachsen können.

Big-Data-Lösungen können messbare Beiträge für die Wertschöpfung leisten. Um einen Erfolg bei ihrer Entwicklung und Nutzung zu erzielen, sollten Unternehmen eine Reihe wichtiger Faktoren beachten (vgl. Tabelle 2). Auf weitere wichtige Erfolgsfaktoren wie den Einsatz agiler Projektmethoden und Vorgehensweisen sowie die Entwicklung von Expertise wird im Kapitel 7 eingegangen.

Tabelle 2: Erfolgsfaktoren für Big-Data-Projekte

Existenz einer wertorientierten Big-Data-Strategie
Business-Fokus
Unterstützung von Seiten des Senior-Managements aus den Fachbereichen
Bewertung des Return on Information
Orientierung an bewährten Grundsätzen und Praktiken des Information Managements
Definition einer wertorientierten Roadmap
Etablierung eines umfassenden Big-Data-Innovationsprozesses
Regelmäßige Erfolgsüberprüfung und Ausrichtung der geplanten Maßnahmen
Auswahl geeigneter Technologien
Einrichtung einer passenden Organisation

Existenz einer wertorientierten Big-Data-Strategie

Für den Erfolg ist ausschlaggebend, dass Organisationen ihre Big-Data-Aktivitäten aus der Unternehmensstrategie ableiten. Für Big Data sollte eine das gesamte Unternehmen umfassende, wertorientierte Strategie entwickelt werden, die insbesondere die Fachbereiche und die IT-Abteilung einbindet. Eine solche Strategie beschreibt:

- die Vision:
  Was soll erreicht werden? Wie sehen die Ziele aus?
- das Vorgehen:
  Wie kann das Ziel erreicht werden? Wie sieht der Ausgangspunkt aus?
- die Roadmap:
  Welche Wege führen zum Ziel?
- die ausführbaren Schritte und Initiativen:
  Mit welchem Umsetzungsplan können die Ziele erreicht werden?

Die Teilschritte zur Entwicklung einer Big-Data-Strategie sind in der Tabelle 3 aufgeführt.

Tabelle 3: Teilschritte zur Entwicklung einer Big-Data-Strategie

Bewertung der bestehenden Prozesse, der Daten, der Organisation sowie der Daten- und technischen Architekturen
Identifikation der für das Unternehmen wertvollen Daten. Welche Bedeutung haben sie? Wie werden sie am besten für Unternehmensziele genutzt?
Ausrichtung an strategischen Unternehmenszielen und Prioritäten
Erstellung einer Landkarte analytischer Fähigkeiten, abgebildet auf benötigte Daten- bzw. technische Architekturen
Identifikation geeigneter Tools und Plattformen
Formale Anforderungsbeschreibung, der erwarteten Ergebnisse und des quantitativen Nutzens
Empfehlungen zu Struktur- und Prozessverbesserungen
Entwicklung von Zeit-, Kosten- und Ressourcen-Schätzungen

Die Erarbeitung der Big-Data-Strategie setzt eine effektive Kommunikation zwischen den Fachbereichen eines Unternehmens und den Abgleich der Ziele voraus. Das Management muss Klammern zwischen der IT- und den Fachabteilungen bilden, die organisatorischen Voraussetzungen für eine neue Form der Zusammenarbeit schaffen und dabei immer das Gesamtunternehmen im Blick behalten. Temporäre, interdisziplinäre Teams aus IT- und Fachabteilungen bieten sich dazu an.

Business-Fokus

Big-Data-Projekte dienen keinem Selbstzweck; vielmehr muss ein zu erzielender Business-Mehrwert Treiber der Entwicklung sein. Je besser dieser zu erzielende Business-Mehrwert formuliert und quantifiziert wird, desto einfacher fällt es, einseitig von der Technologie getriebene Initiativen zu vermeiden. Die Klarheit darüber, welche analytischen Prozesse bzw. welche informationsgestützten Entscheidungen ein erfolgreicheres Geschäft ermöglichen, wird auch den Einsatz einer Big-Data-Lösung in den Fachbereichen positiv beeinflussen.

Senior-Management-Unterstützung aus den Fachbereichen

Business-Mehrwerte können nur in den Fachbereichen eines Unternehmens entstehen. Deshalb ist es unerlässlich, dass die Fachbereichsleitung die Führung bei der Ausrichtung von Big-Data-Initiativen und der Festlegung ihrer Ziele und Aufgaben übernimmt.

Erfolgreiche BI- und Data-Warehouse-Projekte zeigen: Sponsoren aus dem Top-Management der Fachbereiche erleichtern das Zusammenwirken von Vertretern unterschiedlicher Disziplinen deutlich, die in den Projekten benötigt werden. Besonders tiefgreifende Veränderungen können ohne Unterstützung von Fachbereichs- oder Geschäftsleitung nicht durchgesetzt werden.

Bewertung des Return on Information

Bei der Monetarisierung von Daten werden neue Berechnungen durchgeführt, die als »Return on Information« (ROI) bezeichnet werden. Organisationen, die einen hohen ROI erzielen, stellen einen Zusammenhang zu differenzierten und nachhaltigen Wettbewerbsvorteilen und zu besseren Produkten und Dienstleistungen her. Der Wert, der aus Informationen abgeleitet werden kann, ist von folgenden Einflussgrößen abhängig:

- Volumen der Detaildaten, die analysiert werden können,
- Anzahl der Benutzer, die Zugriff auf diese Daten haben,
- Tiefe der Analyse, die mit den Daten ermöglicht wird,
- Zeiteinheit zwischen Erzeugung der Daten und der Zugriffs- und Analysemöglichkeit (Latenz).

Diese Werte müssen zu den Kosten in Relation gesetzt werden, die bei Anschaffung und Betrieb der Big-Data-Lösung entstehen (vgl. Formel 2)[27].

$$\frac{\frac{\text{Daten Volumen} \times \text{Analyse Tiefe} \times \#\text{Benutzer}}{\text{Latenz}} = \text{Daten Wert}}{\text{Hardware} + \text{Lizenz} + \text{Support} = \text{Kosten}} = \textbf{ROI}$$

Return on Information

Formel 2: Return on Information (ROI)

Die Entwicklungen auf dem Anbietermarkt bestätigen die Anwendbarkeit der Formel 2: Es werden Lösungen angeboten, die für immer größere Datenvolumina ausgelegt sind sowie eine zunehmende Analyse-Tiefe, steigende Benutzerzahlen und Analysen in Echtzeit ermöglichen. Gleichzeitig werden Betriebskosten durch Appliances gesenkt sowie Hardware und Lizenzmodelle an wachsende Volumina angepasst.

Orientierung an bewährten Grundsätzen und Praktiken des Information Managements

Die bewährten Praktiken und Grundlagen traditioneller Informations-Management-Initiativen wie Enterprise-Daten-Architektur, Stammdaten-Management, Datenintegration, Datenqualität und Metadaten-Management behalten auch im Kontext von Big Data ihrer Gültigkeit. Eine Abkehr von diesen bewährten Praktiken und Grundlagen lässt die Gefahr entstehen, dass eine Big-Data-Lösung nur isoliert und nicht für einen unternehmensweiten Wertbeitrag genutzt werden kann. Zugleich müssen einige neue Ansätze entwickelt werden, um erfolgreich zu sein.

Definition einer wertorientierten Roadmap

Unternehmen sollten einen wertorientierten Plan (Roadmap) entwickeln, der die Initiativen einer langfristigen Strategie beschreibt, aber gleichzeitig kurzfristige Ergebnisse gewährleistet.
Spezifische Informationen, die im direkten Zusammenhang mit dem Kerngeschäft entstehen, können einen Wettbewerbsvorteil darstellen. So differenziert, wie sich das Unternehmen am Markt mit seiner Unternehmensstrategie positioniert, kann die Wertschöpfung aus Informationen sein. Die Investitionen in die Informations-Exploration bzw. -Innovation werden daher davon abhängig sein, welchen besonderen Mehrwert – bis zu einer möglichen Alleinstellung – eine Information für das Unternehmen aufweist. Die Priorisierung von Initiativen wird daher auf dem geschäftlichen Nutzen basieren.

Dabei bedeuten solche Initiativen nicht unbedingt eine Neueinführung von Big-Data-Technologien. Häufig werden Unternehmen an existierenden Organisationen, Verfahren oder Querschnitts-Funktionen ansetzen und diese optimieren.

[27] Vgl. [HP / Vertica 2013]

Etablierung eines umfassenden Big-Data-Innovationsprozesses

Mit der Bewältigung riesiger Datenvolumina unterstützt Big Data ein tieferes Verständnis für das Business. Die Innovation Big Data umfasst das gesamte Unternehmen und sollte daher in eine innovationsorientierte Unternehmenskultur eingebettet werden, in der das Lernen aus Erfahrung gefördert wird.

Jedem Innovationsprozess ist eine gewisse Unbestimmtheit wesenseigen. Folglich können nicht alle Anforderungen an ein Big-Data-Projekt und sämtliche Nutzungskomponenten vorab vollständig beschrieben werden.

Die Technik von Big Data erlaubt die Einbeziehung neuer Datenarten, die Durchführung zusätzlicher Auswertungen sowie die Erweiterung der Ressourcen im laufenden Betrieb und ist damit für den Einsatz agiler Entwicklungsmethoden[28] besonders geeignet.

Diese Technologie kann auch dazu eingesetzt werden, den Nutzern Verfahren in die Hand zu geben, die ihnen Freiheit bei der Formulierung neuer Analysen lässt. Professionell im Unternehmen aufgesetzt, schafft Big Data eine breite Basis für neue Einsichten und Innovation.

Die Einsatzmöglichkeiten von Big Data sowie die innovativen Wirkungen im Unternehmen sollten zu einem frühen Zeitpunkt durch Machbarkeitsstudien und Pilotlösungen überprüft werden. Der Aufbau von Big-Data-Testumgebungen eröffnet den Fachbereichen und der IT Wege zum Ausprobieren der neuen Technologien und zum Sammeln eigener Erfahrungen. So kann der Gefahr entgegen gewirkt werden, dass Fehler im Lösungs-Design erst am Ende eines Projektes erkannt werden.

Tabelle 4: Schritte im Wandel der Unternehmenskultur

Schritt	Kultureller Wandel
1	Umdenken, denn Big Data lässt sich nicht als Return on Investment über x Jahre darstellen.
2	Ausbildung und Weiterbildung sowie Investitionen in neue Berufe und in den BI-Bereich, um die Informationen in großen Datenmengen zu »explorieren«.
3	Keimzellen schaffen, in denen Big-Data-Ideen zu Big-Data-Szenarien werden. Den Mitarbeitern Freiräume schaffen und lassen.
4	Leitungsentscheidung darüber, ob die Fachabteilung oder die IT-Abteilung für die Exploration von Daten und den Aufbau von Big-Data-Szenarien zuständig ist bzw. ob hier eine Stabsfunktion für das Unternehmen hilfreich ist.
5	Neue Algorithmen finden, statt eines Scale-outs vorhandener Lösungen.

Big-Data-Projekte und können recht komplex sein. Die Speicherung und Auswertung von einigen hundert Terabytes bildet oft den Einstieg in komplexere Szenarien. Mit der Bewältigung der Komplexität sind nicht selten die eigentlichen Wettbewerbsvorteile für die Anwender verknüpft. Die Komplexität der Big-Data-Projekte setzt einige Management-Entscheidungen voraus, die auch einen schrittweisen Wandel der Unternehmenskultur in Richtung Innovationsförderung bedeuten (vgl. Tabelle 4).

[28] Für Innovationsprojekte eignen sich agile Entwicklungsmethodiken. Sie führen schnell zu ersten Resultaten und erlauben die Berücksichtigung von Erfahrungen, die in den ersten Projektschritten gesammelt wurden.
Erfahrungsgestützte Lernfortschritte können in Big-Data-Projekten als agiler, selbstlernender maschineller Prozess implementiert werden. Dafür wurde der Begriff Machine Learning geprägt. Machine Learning setzt in Software implementierte Algorithmen ein, die mit jedem Zyklus aus den Daten hinzulernen, so dass über die Zeit der stetig verbesserte Algorithmus immer bessere Ergebnisse erzeugt.
Der Einsatz von Machine Learning bedeutet, einen nicht-deterministischen, agilen Lernprozess zu konzipieren und zu überwachen. Der Entwurf der Lernszenarien bildet eine neue Aufgabenstellung, die bei den Mitarbeitern ein hohes Abstraktionsvermögen voraussetzt. Die dafür erforderlichen Fähigkeiten reichen weit über bestehende Datenbank- oder Entwicklungs-Skills hinaus. Da entsprechende Ausbildungsangebote noch fehlen, sollte unter mathematisch-algorithmisch Begabten nach Talenten für diese Art der Datenverarbeitung gesucht werden.

Eine Vorstellung detaillierter Prozess- oder Phasen-Modelle würde den vorliegenden Leitfaden sprengen. Es empfiehlt sich der Einsatz z. B. von

- Business Use Cases zur Bewertung des Nutzenpotenzials von Big Data,
- Proof-of-Concepts,
- Früh-Erkennungs-Methoden in der konkreten Datenanalyse[29],
- Ergebnisoffenen Meilenstein-Bewertungen und Reviews.

Diese Ansätze sind mit robusten Implementierungs-Methoden zu verbinden, z. B. mit agilen Entwicklungsmethoden wie »Agile BI«[30].

Regelmäßige Erfolgsüberprüfung und Ausrichtung der geplanten Initiativen

Bei der Ausgestaltung des Big-Data-Innovationsprozesses liegt es im Ermessen des Managements, die regelmäßige Erfolgsüberprüfung als Teil dieses Prozesses zu betrachten oder auf einer höheren Ebene vorzunehmen.

Aus zwei Gründen ist es in jedem Fall wichtig, dass die Big-Data-Initiativen und ihr Zusammenspiel regelmäßig mit Blick auf ihre grundsätzliche Ausrichtung überprüft werden:

- Erstens ändert sich das wirtschaftliche Umfeld sehr schnell.

- Zweitens ist nicht auszuschließen, dass erhoffte Ergebnisse nicht oder nicht vorhergesehene positive Nutzeffekte eintreten.

Wenn die Big-Data-Strategie mit der Unternehmensstrategie abgeglichen ist, kann über Anpassungen leichter im Gesamtkontext entschieden werden.

Als Metrik zur Abschätzung von Umsetzungsaufwand und -risiko eignet sich der Implementierungs-Reifegrad einer gewünschten Funktionalität. Zu seiner Bestimmung wird das zu lösende Problem in Teilprobleme bzw. -aufgaben zerlegt; anschließend werden klar definierte Ziele, Leistungsmerkmale, Dateninhalte und Datenquellen zugeordnet. Das dient der Bewertung,

- ob mit Blick auf die Ziele Zugriff auf die notwendigen Daten vorhanden ist,
- wie hoch der Aufwand zur Zielerreichung sein wird.

Im Kontext von Big Data empfiehlt sich zusätzlich die Identifikation strategischer (Meta-)Daten bzw. Daten-Muster.

Technologie-Auswahl für Big Data

Big Data erhält Impulse aus zahlreichen Technologiebereichen (vgl. Kapitel 8).

Die Möglichkeiten zur Verarbeitung von Daten unterscheiden sich bei Big-Data-Lösungen deutlich von herkömmlicher transaktionalen Datenbank-Anwendungen oder von Data-Warehouse-Lösungen.
Neue Möglichkeiten ergeben sich durch

- die vollständige Verarbeitung von Daten im Hauptspeicher (In-Memory),
- die Nutzung neuer Daten-Speicherungsstrukturen,
- die Analyse riesiger Mengen von Daten über verteilte Dateisysteme,
- die Implementierung unterschiedlicher Transaktionsmodelle oder
- den Zugriff auf noch unqualifizierte neue Datenquellen.

Frei von tradierten Denkmustern[31] muss untersucht werden, welches Design für eine Big-Data-Lösung sich am besten für bestimmte Klassen von Anforderungen eignet.

---

[29] z. B. auf Basis eines geringeren Datenumfanges bzw. geringeren Graden der Einbettung in bestehende Prozesse

[30] Vgl. [TDWI, 2013]

[31] Solche Denkmuster sind z. B.: »Ein Data Warehouse ist das beste System für Datenanalysen« oder »Relationale Datenbanken bilden das Mittel der Wahl, wenn es um Transaktionen geht«.

Zwei Entscheidungen sind von besonderer Tragweite:

- die Auswahl der Hardware:
  Die Betriebsqualität und die Performanz von Big-Data-Lösungen wird stark von der eingesetzten Hardware beeinflusst. Neue Fragen sind zu beantworten (Auswahl):
  - Welche möglichen Vorteile ergeben sich beim Einsatz von SSD-Plattenspeichern?
  - Wie skalieren Big-Data-Software-Lösungen bei der Erweiterung von CPU- und Hauptspeicher-Kapazitäten bzw., bei einer größeren Anzahl von eingesetzten Rechnerknoten?
  - Sind Appliances besser geeignet als Standard-Server?
  - Welchen Einfluss haben parallele Architekturen auf die Algorithmen, die Anwendungsarchitektur und das Laufzeitverhalten der Lösung?
- die Entscheidung für Open-Source-Software oder für kommerzielle Lösungen:
  Big-Data-Technologien wie z. B. Hadoop sind mittlerweile als Open Source verfügbar. Es gilt zu evaluieren, wie sich Open-Source-Produkte und Services von denen kommerzieller Anbieter unterscheiden. Die Erfahrungen von Open Source im Bereich der Anwendungsentwicklung sind nützlich und sollten auf Big-Data-Evaluierungen übertragen werden:
  - Wo ist Support erhältlich?
  - Sind Upgrades der Software und die Kontinuität der Weiteentwicklung sichergestellt?
  - Welche Standards werden unterstützt?
  - Welchen Umfang weisen Schulungsangebote aus?
  - Wie aktiv sind Communities?
  - Wie werden neue Anforderungen berücksichtigt?
  - Welche Kosten entstehen?
  - In welchem Maß werden die Anforderungen an die Funktionalität erfüllt?
  - Auf welchen Plattformen sind die Lösungen verfügbar?

Einrichtung einer geeigneten Organisation

Zu den wesentlichen Erfolgsfaktoren gehört auch der Aufbau einer passenden Organisation (vgl. auch Abschnitt 10.3). Dabei ist u.a. die Frage zu klären, ob die formulierten Ziele aus organisatorischer Sicht überhaupt erreichbar sind und ob die notwendigen Kenntnisse und Erfahrungen vorliegen. Ebenso müssen Rollen und Verantwortung geregelt werden. Es ist z. B. zu klären, wer Daten kreiert, nutzt, prüft, pflegt und verbessert, wer welche Entscheidung trifft und wer die Big-Data-Lösung entwickelt und betreibt (vgl. Tabelle 5).

Für die Entwicklung einer Big-Data-Organisation sollte geprüft werden, ob erprobte Organisationsformen aus den Bereichen BI und DW nützlich sein können.[32]

Tabelle 5: Ausgewählte Aufgaben bei der Organisationsentwicklung für Big-Data-Lösungen

Aufgabe	Erläuterung
Leitungs- (Governance-) Modell	Wer verantwortet und entscheidet Ziele und Initiativen?
Big-Data-Competence-Center	Ausgehend von den mit BI gesammelten Erfahrungen könnte die Entwicklung eines Big-Data-Competence-Center s sinnvoll sein. Es dient der Bündelung bzw. Spezialisierung auf der Fach- und der IT-Seite für Big-Data-Fragestellungen.
Data Governance	Verabschiedung von Standards und Richtlinien für die Daten sowie Festlegung der Data Owner.
Data Stewardship	Umsetzung bzw. Erfüllung der Data-Governance-Standards und Richtlinien
Skills und Ressourcen	Wie werden Teams mit spezifischer Big-Data-Kompetenz gebildet? Muss ggf. auf externe Expertise zurückgegriffen werden?)

[32] Vgl. z. B. [TDWI, 2012].

Zu Beginn eines Big-Data-Programms ist mit Blick auf die Größe der Herausforderung zu entscheiden, ob am Anfang eine projektorientierte Organisationsform sinnvoll ist und wie schnell eine permanente Organisation aufgebaut werden sollte.

Big-Data-Checkliste

In einer Big-Data-Strategie sollten Unternehmen einen Rahmen für ihre Big-Data-Initiativen festlegen. So kann sichergestellt werden, dass die neuen Daten oder Analysen auch effizient und zielgerichtet eingesetzt werden. Eine Checkliste hilft dabei, die wesentlichen strategischen Fragen im Auge zu behalten:

- Welche Herausforderungen soll die Datennutzung lösen?
- Warum sollen diese Herausforderungen gelöst werden? Wie sieht der Business Case aus?
- Welche Daten benötigt das Unternehmen dafür?
- Welche Daten liegen heute in welchen Systemen vor? Ist der Detaillierungsgrad ausreichend?
- Welche der erforderlichen Daten werden heute noch nicht systematisch erfasst?
- Können die fehlenden Daten als Nebenprodukt bestehender Prozesse erzeugt werden? Oder sind neue Erfassungswege dafür erforderlich?

Entsprechend lassen sich dann die wichtigsten Eckpunkte einer Infrastrukturstrategie definieren. Diese umfasst mehrere Aspekte:

- Daten-Infrastruktur/-Architektur:
- Das Unternehmen muss festlegen, welche Systeme für die jeweiligen Datensätze in Zukunft führend sein werden.
- Software-Infrastruktur:
  Unternehmen müssen die Mittel für die Datenanalysen festlegen. Normalerweise geht es dabei um etablierte BI-Werkzeuge, die Standard-Reports aus den vorhandenen Daten erstellen können. In einem Big-Data-Ansatz besteht diese Software-Infrastruktur aus einer Big-Data-Plattform wie Hadoop, Konnektoren zu den relevanten Datenquellen in der Daten-Architektur sowie Analyse-Tools wie Hive für Data Warehousing, Mahout für Machine Learning oder Pig als interaktive Shell.
- Technische Infrastruktur:
  Hier geht es um die technische Infrastruktur für die Umsetzung des Big-Data-Ansatzes. Für das Unternehmen bedeutet das eine klassische Make or Buy-Entscheidung. Wenn Analysen nur einmalig erfolgen bzw. große Schwankungen im Datenvolumen oder in der Analysenachfrage bestehen, dann lohnt sich eher, auf Cloud-basierte Infrastrukturen zurückzugreifen, als in eine eigene Hardware zu investieren. Aufschluss darüber liefert der in der Data Due Diligence entwickelte Business Case.

Unternehmen sollten daher zuerst eine umfassende Bestandsaufnahme auf Basis einer Checkliste erstellen. Dabei können sie wirtschaftlich sinnvolle Ansätze identifizieren, Technologiefragen klären und erste Schritte für eine Pilotumsetzung in die Wege leiten. Zu den wesentlichen Erfolgsfaktoren gehört auch der Aufbau einer passenden Organisation.

Eine Checkliste ist somit ein wichtiges Werkzeug, um die Ziele, aber auch die Reife und die Potenziale, die sich durch Big Data ergeben, zu erfassen und somit direkt an den Anfang von Big-Data-Initiativen zu stellen (vgl. Kapitel 5).

7 Kompetenzentwicklung der Mitarbeiter für Big-Data-Projekte

Wissen über die Einsatzmöglichkeiten von Big Data und die damit verbundenen Technologien ist zurzeit noch rar. Dieses Defizit muss durch Aus- und Weiterbildung der Spezialisten in der Software-Entwicklung, im IT-Betrieb, in den Fachabteilungen sowie im Management zügig abgebaut werden. Dazu eignen sich insbesondere Kollaborationsplattformen. Besonderer Wert muss dabei auf die Förderung von Kreativität gelegt werden.

Big-Data-Projekte sind einerseits normale IT-Projekte, für die etablierte Methoden und Verfahren des Projektmanagements zur Verfügung stehen. Andererseits stellt die Erschließung der Möglichkeiten von Big Data für jedes Unternehmen einen Innovationsprozess mit einer Vielzahl von Facetten und Implikationen dar. Wie tief der mit Big Data eingeläutete Wandel für ein Unternehmen ist, hängt von seiner Strategie ab. Für einige Anwendungsfälle kann Big Data durchaus eine disruptive Technologie darstellen.

7.1 Wege zur Kompetenzentwicklung

Big Data stellt Unternehmen vielfältige neue Möglichkeiten zur Verfügung, Daten zu speichern, zu bearbeiten und mit dem Ziel zu analysieren, mit verbesserten Entscheidungsprozessen wichtige Impulse für das Business zu geben[33]. Big Data unterscheidet sich wesentlich von der bisherigen Datenverarbeitung. Ohne ausreichende Big-Data-Expertise wird kein Big-Data-Projekt zum Erfolg führen.

Ein grundlegendes Verständnis von den neuen Möglichkeiten – aber auch von den Grenzen – muss daher in der gesamten Organisation entwickelt werden. Das gilt ganz besonders für das Management aller betroffenen Fachbereiche. Big Data erfordert neue Kenntnisse und Fähigkeiten bei allen involvierten Mitarbeitern aus den Bereichen Unternehmens- und IT-Architektur, Analyse, Softwareentwicklung, Betrieb und Wartung von IT-Lösungen.

Der Kompetenzaufbau wird erheblich aufwändiger und langwieriger sein als die Installation der Technik. Es ist im Einzelfall zu prüfen, welche Wege[34] der Kompetenzentwicklung zielführend sind:

- Weiterbildung der Mitarbeiter,
- Neueinstellung von Experten,
- Gewinnung externer Berater für bestimmte Aufgaben,
- Zusammenarbeit mit Berufsakademien, Organisationen der Forschung und Hochschulausbildung oder anderen Unternehmen.

In vielen etablierten Aufgabenbereichen werden sich durch neue technische Lösungen Veränderungen ergeben. Hierzu gehören u.a.

- Sicherstellung der Datenintegrität,
- Governance,
- IT-Sicherheit,
- Datenschutz.

[33] Vgl. [BITKOM, 2012]
[34] Mitunter ist auch das Outsourcing von Teilaufgaben an Service-Provider eine Option.

Das Big-Data-Team muss vor dem Projektstart gerüstet sein und eine interdisziplinäre offene Kultur der Zusammenarbeit entwickeln. Deshalb ist es notwendig die benötigten Kompetenzen frühzeitig aufzubauen (vgl. Kapitel 5, Phase 2). Für Unternehmen bedeutet das: Die Kompetenzentwicklung des Big-Data-Team sollte als strategische Investition betrachtet werden, denn Big Data wird zukünftig an Bedeutung gewinnen und verstärkt Business-Nutzen hervorbringen.

7.2 Neue Berufsbilder und Mitarbeiterprofile

Big Data erfordert eine Reihe von Expertiseprofilen, die sich in Zukunft möglicherweise zu vier Berufsbildern oder Mitarbeiterprofile verdichten werden:

- Data Innovator,
- Data Scientists,
- Big Data Developer/Programmierer und
- Data Artists.

In Deutschland gibt es dafür zurzeit noch keine Ausbildungsgänge. Der Bedarf an solchen Spezialisten wird in Hochschulen und Forschungsinstituten anerkannt. Einzelne Unternehmen haben wahlweise Big-Data-Analysts/-Scientist/-Artists in ihrem Beraterstab, andere wirken an der Aus- und Weiterbildung mit. Neue Software-Tools und Software-Entwicklungs-Umgebungen unterstützen die Arbeit der »Erforscher großer Datenmengen«.

Ein agiler Start in Big-Data-Projekte sollte genutzt werden, die benötigte Expertise schrittweise aufzubauen.

Data Innovator / Scout

Der Data Innovator/Scout wird seinen Fokus darauf setzen, innovative Geschäftsmodelle auf Basis Big Data zu identifizieren, zu evaluieren und zu erarbeiten. Es geht darum, die Wettbewerbsfähigkeit des Unternehmens zu erhöhen und die Potenziale von Big-Data-Technologien innovativ zu nutzen. Der Scout ist zuständig für die Pilotierung von Anwendungsideen und die Big-Data-Beratung für Fachabteilungen und Unternehmensleitung.

Data Scientist

Der Data Scientist wird sich hauptsächlich mit den Daten beschäftigen und mit ihnen experimentieren, um neue Verknüpfungen zu finden, deren Auswertung zusätzliche Erkenntnisse verspricht.

Data Scientists sollten kreativ, frei und möglichst bereichsübergreifend mit den Daten arbeiten können. Das setzt in vielen Unternehmen ein Umdenken über Hierarchien und »Hoheitsgebiete« voraus.

Die Berufsausbildung eines Data Scientists sollte Kenntnissen in Mathematik, Wirtschaftsstatistik, IT[35], Medien, Unternehmensführung, Psychologie sowie Branchenwissen vermitteln. Der Data Scientist ist also Fachinformatiker mit zusätzlichen Kenntnissen in weiteren Disziplinen.

Zudem wird es Wissenschaftler mit einer neuen Spezialisierung geben; sie werden in Zusammenarbeit mit Data Scientists neue Muster in Daten erkennen und Auswertungs-Algorithmen entwickeln.

Big-Data-Developer/Programmierer

Big-Data-Developer/Programmierer bauen auf den Erkenntnissen der Data Innovators/Scouts und Data Scientists auf und entwickeln neue Algorithmen und Applikationen. Das könnten beispielsweise Verfahren zur integrierten Verarbeitung von Daten aus unterschiedlichen Quellen oder neue mathematische Verfahren zur Datenanalyse sein. Big-Data-Developer sind erfahrene Programmierer mit tiefen Kenntnissen in Statistik, Mathematik und Semiotik, aber auch Philosophie und Wirtschaft.

[35] Grundlagen der Programmierung und Programmiersprachen, Datenbanken und Datenbanksprachen, Grundlagen der Informationstechnik, Kenntnisse über Netzwerke.

Der Bedarf an Big-Data-Developern wird in dem Maße steigen, wie die Anwenderunternehmen komplexere Big-Data-Szenarien in Angriff nehmen.

Data Artist

Der Data Artist ist für die Visualisierung der Daten verantwortlich. Er wird die oft komplizierten Zusammenhänge so aufbereiten und darstellen, dass die Partner in den Fachabteilungen oder im Management sie schnell erfassen und durchdringen können.

Zur Ausbildung zum Data Artist gehören u.a. Grafikdesign, Psychologie, Mathematik, IT und Kommunikation. Dabei kann auf dem Berufsbild des Mediengestalters aufgebaut werden. Anstelle der werblichen Ausrichtung rücken dann die informationstechnischen Inhalte.

Ein wichtiges Arbeitsgebiet der Data Artists ist die Datenqualität.

7.3 Anpassung bestehender Mitarbeiterprofile

Big-Data-Lösungen werden häufig durch Weiterentwicklung vorhandener Systeme oder durch Integration neuer Komponenten in solche Systeme geschaffen. Es ist daher naheliegend, dass Unternehmen die Möglichkeiten untersuchen, ihre Mitarbeiter zu Big-Data-Spezialisten weiterzubilden. Das bietet sich insbesondere bei fünf Berufsgruppen an:

- Anwendungsentwickler
  Java-Entwickler und andere Anwendungsprogrammierer müssen sich Wissen über Big-Data-Technologien und die damit verbundenen neuen Anwendungsarchitekturen aneignen. Die Programmierung paralleler Systeme sowie Kenntnisse über verteilte Dateisysteme[36] und In-Memory-Technologien gewinnen in Big-Data-Projekten an Bedeutung.

- Datenarchitekten
  Ein heutiger Data Architect wird u.a. in den Bereichen Datenintegration und Sicherstellung einer übergreifenden Data Governance hinzulernen müssen. Big-Data-Technologien wie z. B. In-Memory-Technologien, NoSQL-Datenbanken, Hadoop sowie Datenströme (Data in Motion) stellen neue Anforderungen an das Metadaten-Management. Regeln (Policies), Prozesse und Verantwortlichkeiten müssen neu definiert werden.

- Business Intelligence Analyst
  BI-Analysten entwerfen Datenmodelle für Data-Warehouse-Systeme und stellen den Fachabteilungen Reports und Analysen zur Verfügung. Ihre Kenntnisse und Fähigkeiten müssen auf neue Big-Data-Bereiche erweitert werden. Hier geht es z. B. um die Abbildung von Datenmodellen auf Basis von Key-Value, Graphen oder Dokumentenspeichern in NoSQL- oder In-Memory-Systemen oder um den Zugriff auf Systeme (z. B. SQL, REST, Web Services) zur Integration in bestehende Report- und Visualisierungsumgebungen.

- Datenbankadministrator (DBA)
  DBA sind zurzeit noch stark auf die Administration von relationalen Datenbanken und SQL orientiert. In Big-Data-Systemen entstehen neue Anforderungen. Beispielsweise geht es aufgrund der Verteilung von Daten auf verteilte Datei-Systeme oder neue Big-Data-Speicherungssysteme[37] um die Neu-Ausrichtung von Betriebsprozessen wie Disaster Recovery, Restart/Recovery, Backup/Archivierung.

- Systemadministrator
  Die Bereitstellung von stark skalierbaren und verteilten Big-Data-Systemen, die im Rechenzentrum des Unternehmens oder auf Cloud-Infrastrukturen betrieben werden können, erfordert ein breiteres Wissen über Systemkonfigurationen. Weitere Aspekte wie die Kapazitätsplanung für neue Big-Data-Systeme oder deren Überwachung kommen hinzu.

[36] z. B. Hadoop/HDFS – vgl. Kapitel 8.
[37] z. B. Hadoop, In-Memory, NoSQL

8 Architekturen und Basistechnologien für Big Data

Im Kapitel 8 wird eine funktionale Referenzarchitektur für ein Big-Data-System vorgestellt. Die Referenzarchitektur eignet sich als Grundlage für die Umsetzung. Dabei können Open-Source-Werkzeuge oder kommerzielle Hard- und Software eingesetzt werden. Weiterhin werden kursorisch relevante Technologien beschrieben, die die Grundlage für eine Umsetzung der funktionalen Aspekte bilden.

8.1 Analytische Plattform und Infrastruktur

Optimale Analyseergebnisse erfordern die umfassende, integrierte Auswertung aller strukturierten wie polystrukturierten Daten im Unternehmen. Für die effiziente Bewältigung von Geschäftsproblemen und die Erschließung neuer Geschäftschancen ist der Einsatz eines Bündels von Big-Data-Technologien (vgl. Abschnitt 8.3) unter Einbeziehung von traditionellen Data-Warehouse-, Business-Intelligence- und Business-Analytics-Lösungen anzuraten.

Eine analytische Plattform wird in der Regel aus verschiedenen Standardtechnologien bestehen – der entscheidende Mehrwert entsteht erst, wenn versierte Mitarbeiter Technologien und Daten kreativ kombinieren.

Die analytische Plattform erstreckt sich über alle Schichten der Informationsarchitektur:

- Frontend,
- Backend und
- Algorithmen

Frontend

Das Management und die weniger Technik-affinen Fachbereiche werden die Ergebnisse der analytischen Plattform dann intensiv nutzen, wenn sie verständlich dargestellt werden. Die Herausforderung besteht hier in der Verbindung von etablierten BI-Werkzeugen mit neuartigen Werkzeugen für komplexe Visualisierungen.

Backend

Das Backend besteht aus Bestands- und Neusystemen:

- Bestandssysteme sind Datenbanken, Business-Intelligence-Applikationen oder ein Data Warehouse. Entscheidend ist die erfolgreiche Integration in die Gesamtarchitektur, die mit Konnektoren vereinfacht werden kann.

- Neusysteme entstammen meist den Bereichen Hadoop, Stream Processing oder In-Memory Computing. Werden Daten aus Neusystemen wieder in Bestandssystemen gespeichert, müssen die Ergebnisse an die existierenden Schemata angepasst werden. Dies setzt einen möglichst agilen Ansatz auch in den klassischen BI-Systemen voraus, um mit den Big-Data-Merkmalen Velocity und Variety (vgl. Abbildung 1) effizient innerhalb der Gesamtarchitektur umgehen zu können.

Algorithmen

Bei den Algorithmen zur Verarbeitung der Daten werden die klassischen und die vorhersagenden Algorithmen unterschieden:

- Klassische Algorithmen sind retrospektiv und werden genutzt, um Muster in bestehenden Datenmengen zu identifizieren. Darüber hinaus sind Unternehmen mehr und mehr daran interessiert, einen verlässlichen Blick in die Zukunft zu erhalten.

- Vorhersagende Algorithmen nutzen eine größere Datenbasis und komplexere Modelle. Damit lassen sich verborgene Muster erkennen und genauere Vorhersagen treffen. Die vorhersagenden Algorithmen umfassen die Bereiche

  - Machine Learning und
  - Data Mining.

Beide gehören zum Bereich der Künstlichen Intelligenz. Der Fokus von Machine Learning liegt auf der (Wieder-) Entdeckung von bekannten Zusammenhängen in Daten. Im Unterschied dazu ist Data Mining ein explorativer Ansatz, mit dem versucht wird, unbekannte Zusammenhänge zu entdecken.

Einbindung analytischer Plattformen

Der erste Schritt zur Einbindung von analytischen Plattformen umfasst die Analyse der Geschäftsprobleme und -Potenziale auf der einen Seite. Auf der anderen Seite sind vorhandene und erschließbare Datenquellen zu identifizieren. Dieses Wissen und der Überblick über die Big-Data-Technologien sind Grundlage für den folgenden analytisch-kreativen Schritt – die Antwort auf die Frage: Wie lassen sich die Daten mit den neuen Möglichkeiten der Big-Data-Technologien kombinieren und welche Algorithmen sowie wissenschaftlichen Konzepte bringen neue wertvolle Informationen?

Das verantwortliche Team sollte daher interdisziplinär und fachbereichsübergreifend zusammengestellt sein[38].

Die Einbindung der analytischen Plattform beginnt sinnvollerweise bei den unternehmenseigenen Daten. Ausgehend von der bestehenden Datenbasis sind vorhandene Daten[39] in die Plattform zu integrieren und neue Datenquellen[40] zu erschließen. Die Integration erfolgt am besten über Konnektoren, die den Integrationsprozess vereinfachen und die Dauer, die Unsicherheit und die Kosten der Integration vermindern. Für die Erschließung neuer Daten durch Big-Data-Technologie bietet sich die im Abschnitt 8.2.1 vorgestellte Referenzarchitektur an.

Generell ist die Einbindung analytischer Lösungen eng mit dem Vorgehensmodell verknüpft – die Analyse der Geschäftspotenziale und herausforderungen ist integraler Bestandteil seiner 1. Phase (vgl. Unterabschnitt 5.2.1).

[38] vgl. dazu insbesondere Abschnitt 7
[39] z. B. relationale CRM-Daten
[40] z. B. Textanalyse der Kundenkorrespondenz

8.2 Architektur

8.2.1 Funktionale Architektur

Die funktionale Architektur eines Big-Data-Systems ist im Abschnitt 8.2.1 dargestellt. Nachfolgend werden die einzelnen funktionalen Anforderungsdimensionen detailliert beleuchtet. Da derzeit am Markt keine Technologie verfügbar ist, die allen Anforderungen entspricht, werden im Praxiseinsatz Technologien kombiniert.

Private-/Public-Cloud-Infrastruktur

Die Basis für eine Big-Data-Architektur bildet eine Schicht, welche die Speicherung von Informationen aus unterschiedlichsten Quellen – unstrukturiert und strukturiert – effizient und über einen längeren Zeitraum hinweg kostengünstig ermöglicht. Diese Infrastruktur kann in einer Cloud realisiert werden, wobei sich ein Unternehmen in Abhängigkeit von seinen konkreten Performance- und Compliance-Anforderungen für eine Public oder eine Private Cloud entscheiden wird. Neben der Speicherung von Daten kann in dieser Schicht auch eine erste Vorverarbeitung von Daten erfolgen, deren Ergebnis dann die Grundlage für eigentliche Big-Data-Analysen bildet (Refinement[41]).

Echtzeitverarbeitung / Streaming

Viele neue Datenquellen[42] erfordern die Verarbeitung von Informationen in nahezu Echtzeit, um schnellstmöglich auf neue (Ausnahme-)Situationen regieren zu können bzw. das Eintreten von gewissen Ereignissen[43] in einem Unternehmen zu melden. Spezialisierte Systeme nehmen dazu Daten in Echtzeit entgegen, bewerten diese nach zuvor hinterlegten Regeln und verwerfen die eingegangenen Daten danach wieder.

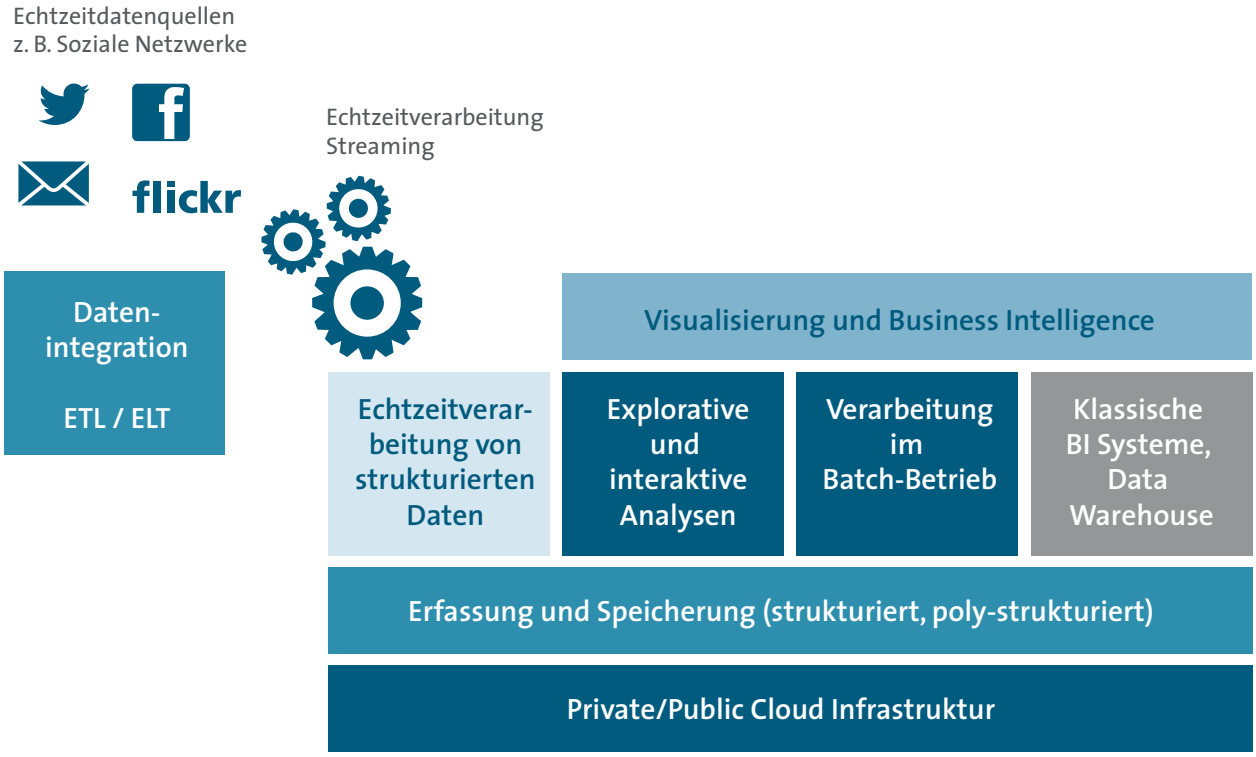


Abbildung 10: Referenzarchitektur eines Big-Data-Systems

41 Ein solches »Refinement« kann zum Beispiel die Transkription einer Audio-Datei in ein Text-Dokument sein.
42 z. B. die sozialen Netzwerke
43 »Managing by Exception«

Datenintegration – Extract-Transformation-Load

In der Praxis wird es oftmals sinnvoll sein, die Komplexität der Datenerfassung zu vermindern. Dazu bietet sich die Verwendung von Datenintegrations-Werkzeugen an, die entsprechende Konnektoren zu Datenquellen bieten. Auf diese Weise können Datenströme in einer zentralen Software gesteuert und überwacht werden. Solche Werkzeuge finden bereits im klassischen BI-Umfeld Verwendung, und so verfügen Unternehmen über Expertise in diesem Bereich.

Echtzeitverarbeitung von strukturierten Informationen

In Unternehmen entsteht oft der Bedarf, in nahezu Echtzeit auf – meist strukturierte – Informationen zuzugreifen.[44] Die Antwortzeit von entsprechenden Modulen in einer Big-Data-Lösung muss garantiert sein und mit der Zahl der Abfragen skalieren können. Hier bieten sich Lösungskomponenten im Umfeld von In-Memory Computing oder spezialisierte NoSQL-Datenbanksysteme geradezu an.

Explorative, interaktive Analysen

Daneben gibt es in immer mehr Unternehmen die neue Rolle des »Data Scientist« (vgl. Abschnitt 7.2). Diese Experten interessieren sich primär für neue Arten der Betrachtung von Daten und benötigen dazu effiziente Werkzeuge, um neue Aspekte in kurzer Zeit zunächst prototypisch umsetzen zu können. Sie gehen mit Daten explorativ und in hohem Maße interaktiv um. Ihre Kreativität wird von Werkzeugen unterstützt, mit denen sich auf einfache Art und Weise komplexe Analysen formulieren lassen.

Verarbeitung im Batch-Betrieb

Neue Konzepte und Ideen werden in der Regel nicht sofort auf dem Gesamtdatenbestand entwickelt und getestet. Hier kommen Werkzeuge zum Einsatz, die die interaktive Datenanalyse unterstützen. Bei vielversprechenden Prototypen erfolgt je nach verwendetem Werkzeug in einem weiteren Schritt die Umsetzung auf der Basis ergänzender Technologien. Hier kann z. B. ein zuvor für eine Teildatenmenge entwickelter Algorithmus auf einen größeren Gesamtdatenbestand angewendet werden. Wie bereits mit dem Begriff »Refinement« (vgl. S. 51) angedeutet, können solche Batch-Prozesse aus Rohdaten Informationen erzeugen, die dann nachfolgend exploriert werden.

Visualisierung und gesicherte Informationsverbreitung

Eine wichtige Anforderung an Big-Data-Lösungen bildet die Darstellung komplexer Zusammenhänge in Daten, um neue Muster zu erkennen bzw. um die Bereiche in den Daten zu identifizieren, die einer genaueren Betrachtung zu unterziehen sind. Dies setzt eine hohe Interaktivität von Visualisierungswerkzeugen voraus, die weit über das »klassische BI-Reporting« hinausgeht.

Neben der Visualisierung zur explorativen Analyse sollte die Big-Data-Architektur die gesicherte Informationsverbreitung im Unternehmen über die Kreise der Data Scientists und Analysten hinaus unterstützen. Erkenntnisse aus Big-Data-Analysen, z. B. eine weitere Dimension mit Bezug zu Kundengruppen, stellt einen Wettbewerbsvorteil dar. Damit ist es wichtig sicherzustellen, dass Zugriffe auf diese Informationen gesichert erfolgen, also nur einem bestimmten Personenkreis zugänglich sind oder aber protokolliert werden.

Die Nutzung der bestehenden BI-Infrastruktur[45] bietet sich als Weg an, aus Big-Data-Analysen gewonnene Einsichten unter Berücksichtigung der bereits Bestehenden Verfahren zum Schutz von Daten im Unternehmen in Geschäftsprozesse einzubetten.

---

[44] z. B. um den Inhalt einer Webseite dem aktuellen Nutzerverhalten anzupassen (»Next Best Action«)
[45] z. B. Enterprise Data Warehouse

Klassische BI-Systeme, Data Warehouse

Big-Data-Lösungsarchitekturen sollten sich darüber hinaus effektiv an bestehende, meist für strukturierte Daten[46] geschaffene Infrastrukturen anbinden lassen, um bereits getätigte Investitionen zu schützen und die Qualität in den Big-Data-Analysen zu gewährleisten. Viele relational strukturiert vorliegende und integrierte Unternehmensdaten eignen sich hervorragend zur Verbesserung von Big-Data-Analysen vorwiegend polystrukturierter Datenquellen.[47] Zusätzlich lassen sich auf diesem Wege Ergebnisse aus Big-Data-Analysen mit bereits vorhandenen Unternehmensdaten validieren.

Die Datenintegration sowie der effiziente Austausch von Informationen über die Grenzen von Teilsystemen hinweg spielen bei Big-Data-Architekturen eine entscheidende Rolle – auch für den Erfolg von Big Data insgesamt.

8.2.2 Umsetzung der Big-Data-Referenzarchitektur

Eine detaillierte Beschreibung, wie die in Abbildung 10 vorgestellte Big-Data- Referenzarchitektur in Hard- und Software umgesetzt werden kann, würde den vorliegenden Leitfaden sprengen. Sowohl die Open Source Community als auch kommerzielle Anbieter bieten technologisch vielfältige Lösungen in allen vorgestellten Bereichen an. Die aktuellen und vor allem die zukünftigen Anforderungen der Nutzer beeinflussen die Umsetzung der Big-Data-Architektur entscheidend. Viele Unternehmen werden eine bestehende BI-Architektur erweitern.

Beim Aufbau von Big-Data-Lösungen ist der Ausbildungsstand der Mitarbeiter ein wichtiger Faktor: Flache Lernkurven werden die Akzeptanz von Big-Data-Lösungen im Unternehmen erleichtern und damit das Potenzial von neuen Anwendungen erweitern. Je mehr Mitarbeiter in einem Unternehmen auf einfachem Wege Zugriff auf die Big-Data-Infrastruktur erhalten, desto eher können Mehrwerte aus Daten generiert werden.

Der Aufwand für den Betrieb einer Big-Data-Infrastruktur ist ebenfalls zu beachten. Eine hohe Komplexität kann sich schnell zu einem enormen Kostentreiber auswachsen. Aus diesem Grund empfiehlt sich die Ausnutzung von Synergieeffekten zwischen klassischer BI und Big Data. Ein Verzicht auf die Nutzung von Synergien zwischen bestehenden Systemen und einer Big-Data-Analyseplattform kann zu komplexen Betriebsprozessen und Datenbewegungen führen, die mittelfristig die Akzeptanz einer Big-Data-Analyseplattform nachhaltig vermindern können.

Einbindung von bestehenden Business-Intelligence- und Analytics-Lösungen

Eine besondere Herausforderung stellt die Einbindung von neuartigen Big-Data-Applikationen, -Verfahren und -Technologien in bestehende Business-Intelligence- und Analytics-Architekturen dar. Der Grund liegt in der hohen Variabilität im Kontext von Big Data.

Das Ziel sollte es sein, eine Infrastruktur bereitzustellen, die umfassend alte und neue Datenquellen nutzt und durch eine Kombination verschiedener Software- und Hardware-Komponenten sowie moderner Algorithmen Geschäftsprobleme effizient löst.

Der Zugang zu dieser Infrastruktur sollte möglichst einfach gestaltet werden, damit sich viele Mitarbeiter daran beteiligen können, innovative Lösungsansätze im Unternehmen in neue Geschäftsmodelle zu transformieren (vgl. Kapitel 3). Dabei ist der Fokus auf den Fail-Fast-Ansatz[48] zu legen, um schnell und fundiert entscheiden zu können, ob ein neuer Ansatz den gewünschten Mehrwert erbringen kann.

---

[46] z. B. ein Enterprise Data Warehouse

[47] Ein Beispiel ist die Verknüpfung von Informationen über Werbekampagnen mit den Informationen über das Nutzerverhalten auf einer Webseite.

[48] Fail-Fast bezeichnet die Fähigkeit eines Systems zur Früherkennung von Fehlern. Ein Fail-Fast System kann an seinen Schnittstellen Fehler oder Zustände, die zu Fehlern führen, erkennen und aufzeigen.

Die Einbindung neuer analytischer Lösungen in die vorhandene Infrastruktur ist eng mit der Phase 3 des Vorgehensmodells (vgl. Unterabschnitt 5.2.3) verknüpft. Die Nutzung vorhandener sowie die Erschließung neuer Datenquellen wird durch die Phasen 4 (vgl. 5.2.4) und 5 (vgl. 5.2.5) adressiert, der Aufbau und die Optimierung der Analysen durch Phase 6 (vgl. 5.2.6).

8.2.3 Ansätze zur Integration von Big-Data-Lösungen

Im Abschnitt 8.2.1 wurde eine funktionale Big-Data-Architektur vorgestellt. An dieser Stelle sollen nun Transformationsansätze beschrieben werden, wie diese Architektur Realität werden kann.

- Ansatz 1 – Einführung neuer Prozesse in ein vorhandenes System:
  Ein Big-Data-Konzept kann unter Beachtung der Leistungsgrenzen der bestehenden Systeme erstellt werden. Dieser Ansatz läuft darauf hinaus, innovative Lösungen mit der bestehenden Hard- und Software umzusetzen, indem z. B. neue Prozesse eingeführt werden. Bei diesem Ansatz besteht die Gefahr, dass die Auswahl möglicher Lösungen von vornherein einschränkt und das volle Potential von Big Data nicht ausgeschöpft wird.

- Ansatz 2 – Migration auf neue Systemkomponenten:
  Eine andere Möglichkeit besteht im Austausch und der Migration bestehender Lösungen auf eine alternative Software und ggf. auch Hardware. Ein solcher Wechsel ist mit hohen Kosten und eventuell auch Widerständen in der eigenen Organisation verbunden. Der notwendige Aufbau von Wissen im Unternehmen darf nicht unterschätzt werden und sollte in eine ganzheitliche Bewertung mit eingehen.
  Die empfehlenswerte Alternative ist die Kombination und Integration von Big-Data-fähiger Hard- und Software in die bereits bestehende Infrastruktur oder die gezielte Veränderung von Lösungen hin zu einer Shared-Nothing- oder In-Memory-fähigen Architektur.

- Ansatz 3 – Einführung neuer Technologien:
  Die Integration neuer Technologien für neue Lösungen hat den Vorteil, dass bestehende Standardsoftware und eine Vielzahl an Features genutzt werden können. Auch können diese Lösungen auf bereits integrierte Daten in einem Data Warehouse zugreifen, um neue Big-Data-Lösungen zu validieren oder mit weiteren Unternehmensdaten anzureichern.
  Gegen diesen Ansatz spricht, dass eventuell existierende Probleme der bestehenden Systeme mit in den Big-Data-Bereich übergehen und dort zu Limitierungen führen können. Datentransfers zwischen den Systemen sollten effizient möglich sein, um die Komplexität der Architektur nicht wachsen zu lassen. Der Austausch von Metadaten ist hier ein Schlüssel zum Erfolg.
  Nicht zu vernachlässigen ist bei diesem Ansatz der Druck auf die bestehenden Systeme, sich den neuen Gegebenheiten von Big Data anzupassen.

8.3 Basistechnologien

Im Abschnitt 8.3 werden die technologischen Grundprinzipien von Big-Data-Lösungen kursorisch erläutert, wobei von den praxisrelevanten Anforderungen ausgegangen wird.

Lineare Skalierbarkeit

Um dem ständig wachsenden Datenvolumen begegnen zu können, ist die lineare Skalierbarkeit eines Big-Data-Systems eine wichtige Voraussetzung. Hierunter wird die Fähigkeit einer Big-Data-Lösung verstanden, z. B. bei einer Verdopplung des Datenvolumens die gleiche Antwortzeit zu bieten, wenn die Kapazität des Systems ebenfalls verdoppelt wurde.

In der Praxis zeigt sich, dass eine gute Big-Data-Lösungsarchitektur in kurzer Zeit neue Nutzerkreise in einem Unternehmen anspricht, die ein schnelles Wachstums der Nutzerzahlen und des Workloads der Big-Data-Lösung zur Folge hat. Daher sollte die Architektur linear horizontal skalierbar angelegt sein. Darunter versteht man die

Erweiterung der Lösung um weitere Rechenknoten, um den Big-Data-Merkmalen Velocity und Variety zu entsprechen (vgl. Abbildung 1).

Massive Parallel Processing und Shared-Nothing-Architekturen

Mit Blick auf lineare Skalierbarkeit und massiv-parallele Verarbeitung haben sich Shared-Nothing-Architekturen am Markt etabliert, bei denen die Gesamtdatenmenge in möglichst gleich große Teile unterteilt wird, die dann auf unterschiedlichen Rechenknoten abgelegt werden. Die Verarbeitung der Daten erfolgt zunächst lokal durch jeden Rechenknoten im System auf dem jeweiligen Teildatenbestand. Die Ergebnisse der lokalen Berechnungen werden dann in einem weiteren Schritt zu dem einen gewünschten Ergebnis zusammengefasst. Nicht nur das MapReduce-Paradigma baut darauf auf, auch relationale Datenbanksysteme nutzen diese Architektur seit Jahrzehnten.

Diese Architektur eignet sich am besten für Problemstellungen, die die lokale Berechnung von Teilergebnisse erlauben, wenn also kein Datenaustausch zwischen den einzelnen Rechenknoten im System zur Berechnung der Teilergebnisse notwendig ist.

Die Verteilung von Daten über viele Rechenknoten hinweg bietet auch eine effiziente und einfache Möglichkeit, die Ausfallsicherheit eines Systems zu erhöhen. Dazu werden die Daten mehrfach in Form von Kopien auf unterschiedlichen Rechenknoten gespeichert. Je nach der konfigurierten Zahl der sogenannten. »Replica« können multiple Rechenknoten ausfallen, ohne dass es zu Datenverlusten kommt. Dieser Ansatz kann so weit gehen, dass auf ein Backup der Daten verzichtet werden kann.

MapReduce Verarbeitung

Das aus einer Google-Veröffentlichung von 2004 bekannte MapReduce-Paradigma[49] stammt konzeptionell aus der funktionalen Programmierung. Am Markt sind Implementierungen verfügbar, die auf unterschiedlichen Programmiersprachen basieren.

Die Daten werden in zwei Schritten verarbeitet:

- In dem »Map«-Schritt werden die Eingabedaten verarbeitet.
- Die dabei entstehenden Zwischenergebnisse werden an den zweiten Prozessschritt »Reduce« weitergereicht, um diese wieder zu verdichten.

Diese Art der Verarbeitung wird in der Regel zu Prozessketten verbunden, um das gewünschte Ergebnis zu berechnen (vgl. Abbildung 11).

Diese Art der Datenverarbeitung eignet sich hervorragend für die Berechnung in massiv-parallelen Systemen, da jeder Rechenknoten den »Map«-Schritt auf dem ihm zugeteilten Teildatenbestand ausrechnen kann. Erst bei der Übergabe der Zwischenergebnisse an den »Reduce«-Schritt müssen die Zwischenergebnisse zwischen den Rechenknoten ausgetauscht werden. In der Regel ist das im »Map«-Schritt erzeugte Datenvolumen deutlich geringer als das Volumen der Eingabedaten[50].

Lässt sich das zu lösende Problem auf der Datenseite in beliebig kleine Eingabemengen unterteilen, so kann durch die Bereitstellung von weiterer Hardware die Ausführungsgeschwindigkeit gesteigert werden.

Hadoop ist die wohl bekannteste Implementierung dieses Konzepts, wobei weitere Umsetzungen in kommerziellen Produkten angeboten werden.[51]

---

[49] Vgl. [DeGhe, 2004]

[50] z. B. Konvertierung von Audio zu Text

[51] Ein Beispiel sind relationale, um das MapReduce-Paradigma erweiterte Datenbanksysteme.

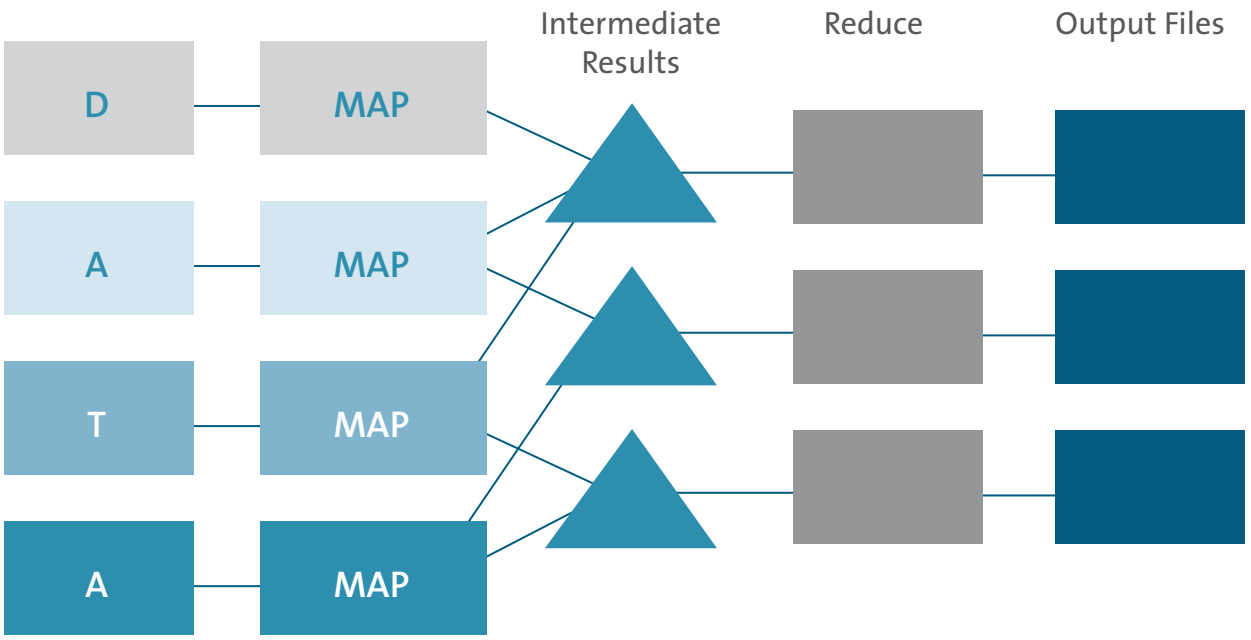


Abbildung 11: MapReduce-Verarbeitung[52]

In-Memory-Computing

Für extrem zeitkritische Anwendungen hat sich die In-Memory-Technologie etabliert. Bei dieser Variante der Datenhaltung wird auf Plattenspeicher als Speichermedium verzichtet. Mit den im Speicher des Systems gehaltenen Daten lässt sich die Verarbeitungsgeschwindigkeit enorm erhöhen. Auch bei sehr komplexen Fragestellungen werden kurze Antwortzeiten ermöglicht. Neue Anwendungsszenarien können erschlossen werden.

In diese Kategorie fallen sogenannte In Memory Data Grids (IMDG), die bei der Verarbeitung komplexer Ereignisse[53] häufig Anwendung finden. Die Nicht-Persistenz der Daten erfordert eine entsprechende Architektur der Anwendung oder parallel geschaltete Persistenz zur Sicherung gegen Ausfall einer Komponente. Vor diesem Hintergrund sind hybride Ansätze entwickelt worden, bei denen Systeme über eine Speicherhierarchie verfügen und automatisch entscheiden, welche Daten in welcher Speichertechnologie vorgehalten werden.

NoSQL

NoSQL lässt zwei Interpretationen zu.

- Zum einen werden damit Systeme bezeichnet, bei denen der Zugriff auf die gespeicherte Information nicht über die auf Mengenverarbeitung ausgerichtete Structured Query Language (SQL) erfolgt.

- Zum anderen wird der Begriff als »Not Only SQL« verstanden. Bei dieser Sichtweise geht es um die Einbettung der etablierten SQL-Datenbanken in die Big-Data-Architektur sowie um den Rückgriff auf andere Methoden und Abfragesprachen in Anwendungsfällen, für die sich SQL nicht eignet.

NoSQL Datenbanken implementieren häufig KeyValue-Stores, wobei der »Value« hier auch ein ganzes Dokument sein kann. Oder sie sind spezialisiert auf besondere Datenstrukturen wie z. B. Graphen, wie sie in der Analyse von sozialen Netzen eingesetzt werden.

[52] Quelle: http://commons.wikimedia.org/wiki/File:Mapreduce_(Ville_Tuulos).png
[53] Complex Event Processing (CEP)

Virtualisierung

Ein Schlüssel zur optimalen Ausnutzung der Hardware stellt die Virtualisierung dar. Hierbei wird von der eigentlichen verfügbaren Hardware durch Software abstrahiert, so dass unterschiedliche Anwendungen nebeneinander auf der gleichen Hardware ausgeführt werden können. Auf diese Weise erreicht man eine deutlich bessere Auslastung der Hardware, da nicht für jede Anwendung dedizierte Hardware bereitgehalten werden muss. Die Bereitstellung von virtuellen Ressourcen ist aus betriebswirtschaftlicher Sicht deutlich günstiger als die Anschaffung und Inbetriebnahme weiterer Hardware. Über diese Technologie bietet das Cloud Computing die Möglichkeit, bei Bedarf virtuelle Ressourcen kurzfristig verfügbar zu machen.

8.4 Semantische Analysen

Neben Datenvolumen und Geschwindigkeit, mit der Daten erfasst und verarbeitet werden müssen, ist die Variabilität der Daten ein charakteristisches Merkmal von Big Data (vgl. Abbildung 1). Mit anderen Worten: Big Data operiert überwiegend auf unstrukturierten Daten wie Text und Sprache und nur zu einem verschwindend geringen Teil auf strukturierten Daten, die klassischen BI-Systemen zugeführt werden können.

Um aus heterogenen Datenströmen Erkenntnisse und Aktivitäten ableiten zu können, ist in einem ersten Schritt eine Strukturierung zwingend erforderlich. Dafür können semantische Technologien zum Einsatz kommen, die wiederum oft Linked Open Data nutzen – Daten aus verknüpften, offen zugänglichen, semantischen Datenbanken.

Semantische Technologien erlauben es, auf der Basis von Domänen- und Hintergrund-Informationen sowie durch den Einsatz von Sprachtechnologien eine tiefere Analyse der Daten vorzunehmen, Informationen zu extrahieren und verborgene Zusammenhänge explizit zu machen. Diese Strukturierung ist jedoch nicht so zu verstehen, dass alle Daten zwingend in ein vordefiniertes relationales Schema gepresst werden. Vielmehr bleibt der unstrukturierte, heterogene Charakter der Daten erhalten – sie werden aber durch Zusatzinformationen und Annotationen ergänzt. In gewisser Weise wird aus Big Data also Smart Data, indem auf die Rohdaten eine semantische Struktur projiziert wird – meist in Form von Graph-orientierten semantischen Modellen.

Semantische Verfahren werden primär für die Tabelle 6 angeführten Verarbeitungsschritte genutzt:

Tabelle 6: Verarbeitungsschritte – Einsatz semantischer Verfahren

Filtern	Automatische Sichtung des Datenstromes und Vorselektion der potenziell relevanten Inhalte, um das »Rauschen« zu eliminieren
Klassifizieren	Einordnung der zuvor selektierten Inhalte in inhaltliche Kategorien, die dann zum Beispiel die weitere Verarbeitung steuern
Annotieren	Anreicherung der Daten um semantische Entitäten sowie Disambiguierung von Begrifflichkeiten gemäß Kontext
Extrahieren	Ableitung und Explizit-Machen von Zusammenhängen und Fakten, die dann zum Beispiel in Form von Relationen BI-Systemen zugeführt werden können

Semantische Verfahren profitieren enorm von den im Abschnitt 8.3 erläuterten Basistechnologien und skalierbaren Architekturen und können deshalb auch bei großen Datenmengen eingesetzt werden, während frühe Techniken der Künstlichen Intelligenz häufig noch schnell an Skalierungsgrenzen stießen.

Szenario: Wettbewerbsbeobachtung

Ein Beispiel soll den Nutzen semantischer Technologien transparent machen.

Für viele Unternehmen wird eine intensive Beobachtung des Marktes, der Wettbewerber und relevanter Technologien immer wichtiger[54]. Zusätzlich zu den frei im Internet verfügbaren Informationen kommen hier auch Quellen des Deep Web zum Tragen, etwa in Form von Patent-Datenbanken oder Publikations-Archiven, die unter Einbeziehung von Linked Open Data semantisch analysiert werden. Neben dem Datenvolumen besteht eine Herausforderung in den unterschiedlichen Begrifflichkeiten und Terminologien in heterogenen Quellen. Verfahren zur Sicherung der Datenqualität und Datenintegrität sind deshalb für Big-Data-Szenarien essenziell.

Mit Big-Data-Technologien in Kombination mit Semantik lassen sich die geschilderten Problemstellungen bewältigen. Nachdem zunächst mit Big-Data-Techniken große Teile des Internets durchforstet und die anfallenden enormen Datenmengen erfasst und gespeichert werden, kommen dann semantische Verfahren zum Einsatz, die insbesondere auf fachspezifische Ontologien zurückgreifen können (vgl. Tabelle 7):

Eine Besonderheit dieses Anwendungsfalles liegt darin, dass absolute Aussagen zu einem einzelnen Zeitpunkt kaum bewertet werden können, sondern vielmehr Trend-Analysen und Zeitreihen deutlicher auf Neuentwicklungen und Veränderungen in den Prioritäten hinweisen.

Dieses realisierte Szenario verdeutlicht, wie semantische Technologien im Zusammenspiel mit Linked Data eine Kombination, Integration und Korrelation von unternehmensinternen und öffentlich verfügbaren Datenquellen unterstützen –unabhängig davon, welchem Schema diese Datenquellen entsprechen. Variabilitäts- und Komplexitätsaspekte von Big Data (vgl. Abbildung 1) werden also durch Semantik besonders gut adressiert.

Tabelle 7: Einsatz semantischer Verfahren zur Wettbewerbsbeobachtung

Filtern	Mittels relativ einfacher und robuster Verfahren, wie etwa auf Basis von Stichwort-Listen, wird zunächst eine grobe Vorauswahl getroffen, um einen Großteil der offensichtlich irrelevanten Daten zu eliminieren und die weiteren Verarbeitungsschritte auf potenziell relevante Dokumente zu beschränken.
Klassifizieren	Klassifikationsverfahren, die auf Methoden aus dem Information Retrieval und der Künstlichen Intelligenz basieren, werden anschließend genutzt, um die relevanten Inhalte vor zu sortieren, in Kategorien einzuteilen und zum Beispiel spezifischen Wettbewerbs- oder Technologiefeldern zuzuordnen.
Annotieren	Die so ermittelten Inhalte werden im Detail analysiert, wobei die tiefe Analyse von Texten mittels linguistischer Methoden eine wesentliche Rolle spielt. Zur Unterstützung dieser Sprachtechnologien kommen fachspezifische Ontologien zum Einsatz, mittels derer Produkt- und Technologie-Bezeichnungen erkannt und zueinander in Relation gesetzt werden; verstärkt werden auch Quellen aus dem Internet via Linked Data eingebunden. Je nach Situation kommen weitere Techniken zum Einsatz, etwa aus der Bilderkennung, um Grafiken interpretieren oder chemische Strukturformeln erschließen zu können. Als Ergebnis ist der originäre Inhalt umfangreich semantisch annotiert, so dass eine Erschließung des Inhaltes möglich wird, zum Beispiel mittels semantischer Suchverfahren.
Extrahieren	Auf Basis der zuvor vorgenommenen Annotationen werden Fakten extrahiert und somit Zusammenhänge zwischen den annotierten Elementen explizit gemacht. Falls gewünscht, können diese Fakten in Drittsysteme übertragen und zum Beispiel klassischen BI-Systemen zugänglich gemacht werden.

[54] Dafür hat sich international der Begriff Competitive Intelligence durchgesetzt.

9 Deployment- und Betriebsmodelle für Big-Data-Anwendungen

Alle Organisationen, die Big-Data-Lösungen etablieren wollen, sehen sich beim Deployment und beim Betrieb von Big-Data-Anwendungen mit ähnlichen Herausforderungen konfrontiert.

Eine Reihe von Parametern (vgl. Abschnitt 9.1) bestimmen, in welchem Maß die Architektur skalierbar ist; die Skalierbarkeit bildet wiederum die Voraussetzung für das Deployment eines erfolgreichen Big-Data-Projekts. Für jeden Parameter werden die Auswirkungen auf das Deployment und das Betriebsmodell dargestellt. Auf dieser Basis werden Big-Data-relevante Grundsätze für skalierbare Architekturen und deren Notwendigkeit aufgezeigt. Wie eine geeignete Architektur betrieben werden kann, beschreibt der Abschnitt 9.2.

9.1 Dimension des Deployments

Für das Deployment von Big-Data-Anwendungen sind zwei verschiedene Analyseansätze zu unterscheiden:

- Erschließung der Daten und Ermittlung von Zusammenhängen für die weitere Formulierung von Business Cases
- Fortlaufende Produktion vorgefertigter Berichte oder Datenbestände zur weiteren Analyse durch den Anwender.

Das Deployment von Anwendungen der ersten Kategorie ist extrem individuell und entsprechend kaum in generalisierter Form darstellbar. Ein solcher Ansatz wird typischerweise in Form eines Data-Discovery-Projektes bei erstmaliger Erschließung neuer Big-Data-Quellen genutzt. Die Vorgehensweise dabei dürfte in der Regel ein »Sandbox«-Ansatz sein und sollte als Ergebnis einen oder mehrere Business Cases als Grundlage für weitere Big-Data-Projekte liefern. Dabei ist es durchaus legitim, dass als Ergebnis auch die eine oder andere betrachtete Big-Data-Datenquelle als nicht relevant verworfen wird.

Beim Deployment von Big-Data-Umgebungen spielt die konkrete Ausprägung der Merkmal (vgl. Abbildung 1) eine wichtige Rolle.

Die im Folgenden dargestellten Dimensionen eines Big-Data-Deployments beziehen sich im Wesentlichen auf Deployments des zweiten Analyseansatzes[55].

9.1.1 Datenvolumen

Ein zentraler und zugleich offensichtlicher Faktor bei der Betrachtung des Deployments von Big-Data-Umgebungen ist das zu berücksichtigende Datenvolumen. Unternehmen in nahezu jeder Branche sehen sich mit der Herausforderung von Informationsgewinnung aus großen Datenbeständen konfrontiert[56]. Dies zeigt: Nicht allein der Umfang der Daten ist für das Deployment von Big-Data-Lösungen relevant. Auch das enorme Wachstum der Datenbestände hat Einfluss auf Architekturentscheidungen.

Der Umfang und das Wachstum der zu betrachtenden Daten haben weitreichende Auswirkungen auf die

[55] Um ein Deployment der Big-Data-Applikationen zu ermöglichen, müssen bereits in der der Phase 2 (Readiness) des Vorgehensmodells die Hardware- und Software-Grundlagen gelegt werden, die später für den zuverlässigen Betrieb der Lösung notwendig sind.

[56] Vgl. dazu [BITKOM, 2012]

Architektur einer Big-Data-Umgebung. Der Speicherplatz muss entsprechend dimensioniert werden und Reserven für das prognostizierte Wachstum vorsehen. Um die Datenvolumina aus den verschiedenen Datenquellen transferieren zu können, sollten die Infrastruktur ausreichend angebunden sein. Die Verarbeitung der Datenbestände erfordert die entsprechende Rechenleistung. Und letztendlich sollte bei der Auswahl der Architektur Skalierungs- und Parallelisierungsoptionen in Betracht gezogen werden. Nicht selten bietet sich aufgrund dieser notwendigen Skalierungsanforderung eine Cloud-basierte Big-Data-Lösung an.

9.1.2 Datenvielfalt

Big Data kann aus verschiedensten Quellen, intern und extern, gespeist werden. Insbesondere für externe Quellen müssen hierfür im Deployment Aspekte wie Authentifizierung, Autorisierung, Verschlüsselung und Zugriffs- bzw. Antwortzeitverhalten des Datenanbieters berücksichtigt werden. Bei externen Quellen sind für das Deployment notwendige Vorlaufzeiten für Beantragungen, Genehmigungen und Bereitstellung der Daten durch den Drittanbieter zu berücksichtigen.

Der Strukturierungsgrad von Big Data variiert zwischen klassisch strukturierten Daten, z. B. Protokolldaten interner Applikationen und völlig unstrukturierten Bild- und Tondaten. Zwischen diesen beiden Extremen können alle Abstufungen semi-strukturierter Daten auftreten. Zum Beispiel prinzipiell strukturierte Daten, die aber Freitextfelder oder Bilder enthalten. Der Strukturierungsgrad der Daten ist ein wesentlicher Aspekt der Systemarchitektur in Bezug auf die einzusetzenden Werkzeuge und notwendigen Verarbeitungsschritte und damit auch entscheidend für das Deployment einer Big-Data-Anwendung.

Grundsätzlich gilt:

- Die Nutzung externer Datenquellen verursacht höhere Aufwände im Deployment als die Verwendung interner Datenquellen
- je unstrukturierter die Daten, desto mehr Verarbeitungsschritte sind notwendig und desto mehr Aufwand entsteht im Deployment.

9.1.3 Datenqualität

Ein weiterer, für die Architektur einer Big-Data-Lösung und für deren Deployment, relevanter Aspekt, sind die Dichte, Schärfe bzw. Unschärfe und Verlässlichkeit der Daten. Abhängig von diesen Parametern kann die Datenmenge festgelegt werden, die für die Gewinnung belastbarer Aussagen zu analysieren ist und daher benötigt wird, z. B. zur statistischen Ableitung von Trends.

Unter Dichte ist der Anteil der tatsächlich relevanten Informationen am Gesamtvolumen der Daten zu verstehen. Ein Beispiel hierfür können Protokolldateien einer internen operativen Anwendung sein, aus denen im Rahmen der aktuellen Analyse jedoch nur Ereignisse eines bestimmten Typs interessieren. Relevant für die aktuelle Analyse ist nun abzuschätzen, wie viele dieser Ereignisse benötigt werden, um daraus eine repräsentativen Trend abzuleiten. Ermittelt man nun noch den prozentualen Anteil dieser Ereignisse am Gesamtvolumen des Protokoll-Files kann man so die benötigte Menge an Rohdaten errechnen. Je höher die Informationsdichte der Daten ist, umso geringer kann die benötige Datenmenge angesetzt werden.

Mit Schärfe ist die Eindeutigkeit der Informationen gemeint. So verliert ein aus einer Sprachnachricht gefilterter Text an Informationsgehalt, wenn die Emotionen auf Basis der Stimmlage nicht als Information erfasst werden. Gleiches gilt z. B. für Mimik und Gestik bei Bildaufzeichnungen. Auch strukturierte Daten können eine mangelnde Schärfe aufweisen, wenn z. B. Eigenschaften in Freitextfeldern beschrieben werden und nicht mit Hilfe von Auswahlen oder definierten Wertebereichen.

Die Verlässlichkeit von Daten ist häufig abhängig von deren Quelle. In Ausnahmefällen können auch Transportwege einen Einfluss haben, z. B. durch Datenverluste. Im Zusammenhang mit externen Quellen denke man z. B. an Daten aus Bewertungsportalen und ihre mögliche Manipulation durch Dritte. Aber auch bei internen Applikationen schwankt die Verlässlichkeit der Daten in der Regel mit der direkten Relevanz für den die Daten erfassenden Benutzer bzw. der Güte der die Datenqualität sichernden Maßnahmen. In Abhängigkeit von der Quelle kann wiederum eine Gewichtung der Relevanz der Daten erfolgen und auch hier ein Datenvolumen bestimmt werden, das repräsentative Aussagen erlaubt.

Alle genannten Punkte sind architekturrelevant und bestimmen damit auch die Komplexität des Deployments einer Lösung.

9.1.4 Datenzugriff

Maßgeblich für die Komplexität des Zugriffs auf die Daten sind neben der Datenquelle selbst auch Aspekte wie Datenschutz und Datensicherheit. Während in der Entwicklung oft exemplarische Datenbestände verwendet werden können, muss im Deployment eine kontinuierliche Datenversorgung sichergestellt werden. Hierbei sind notwendige Vorlaufzeiten zur Etablierung der Zugriffsverfahren, insbesondere bei externen Datenbeständen, zu berücksichtigen. Auf Grund des Datenvolumens kommt der Kapazität und der Skalierbarkeit der Verbindungen eine besondere Bedeutung zu. Weiterhin auch rechtliche Aspekte bezüglich der Lokalität der Datenspeicherung und der zugriffsberechtigten Benutzerkreise einschließlich geeigneter Verschlüsselungsverfahren für die Datenübermittlung aus der Quelle.

Bei personenbezogenen Daten sind, unabhängig von der Datenquelle, erhöhte Datenschutzanforderungen zu berücksichtigen, die ebenfalls die Deployment-Aufwände negativ beeinflussen können.

9.1.5 Echtzeitverhalten

Die Anforderung an ein Echtzeitverhalten (Realtime) für Auswertungen innerhalb von Big-Data-Lösungen setzt ein Echtzeitverhalten der Eingangsdaten voraus. Realistischer Weise spricht man hierbei von einem Beinahe-Echtzeitverhalten (Near-Realtime), da die Verarbeitung der Daten eine gewisse Zeit in Anspruch nimmt. Die Optimierung des Auswertungsverhaltens von Near-Realtime auf Realtime steht für die meisten Anwendungsfällen in keinem Verhältnis zum technischen und damit finanziellen Aufwand.

Unabhängig davon ob ein Unternehmen in seiner bei Big-Data-Lösung Near-Realtime oder Realtime-Auswertungen realisieren möchte, nimmt die Aktualität der Eingangsdaten einen entscheidenden Einfluss auf die Gesamtarchitektur ein. Sehr häufig erzielt ein Unternehmen aus den gewonnenen, verknüpften und analysierten Informationen einen Wettbewerbsvorteil. Realtime Auswertungen und Trends helfen dem Unternehmen diese Erkenntnisse adhoc zu nutzen. Ein Luftfahrtunternehmen, welches Diagnoseinformationen der in der Luft befindlichen Flugzeuge erhält und so Wartungszyklen und Bodenzeiten minimieren kann, oder ein Finanzunternehmen, welches im Sekundentakt den Datenbestand auf betrügerische Vorgänge überprüft, sind auf die Aktualität der Eingangsdaten angewiesen.

Die hohe Frequenz der Eingangsdaten in Kombination mit dem großen Volumen stellt eine Herausforderung für Gesamtarchitektur da. Die Vielzahl der Eingangskanäle erfordert die Möglichkeit der Parallelisierung. Eine entsprechend dimensionierte Rechenleistung und Netzwerkbandbreite bilden die Grundlage einer Big-Data-Architektur.

9.1.6 Analytics

Die bereits benannten Eigenschaften der zu betrachtenden Daten, wie Strukturierungsgrad, Dichte etc. bedingen ggf. ein mehrstufiges Verfahren der Datenaufbereitung. Auch die Analyse selbst erfordert unter Umständen ein mehrstufiges Verfahren. Verfahren, die dabei zum Einsatz kommen können sind z. B. Mustererkennung, Bildanalytic, Spracherkennung, Data Mining etc. Entscheidend ist aber auch, ob der Endbenutzer ein vorkonfektioniertes, parametergesteuertes Resultat erwartet oder vielmehr eine Datenplattform für weitergehende, eigene Analysen. Die Implementierung unterschiedlicher Verdichtungs- und Analysestufen geschieht u.U. mit Hilfe unterschiedlicher Werkzeuge und bedingt dadurch einzelne Deployments von unterschiedlicher Art und Umfang. Die Anforderungen in Bezug auf Rechenleistung und Zugriffsgeschwindigkeit sind mit Big Data sehr hoch, insbesondere wenn Echtzeit-Verarbeitungsanforderungen[57] hinzukommen. Es kommen Technologien wie In-Memory-Verarbeitung, spaltenorientierte Datenbanken, Hadoop, etc. zum Einsatz und verlangen nach einer extrem skalierbaren Infrastruktur. Solchen extremen Skalierungsanforderungen bezüglich der Infrastruktur kann z. B. durch Nutzung einer Cloud begegnet werden. Bei der Entscheidung ob und welche Art von Cloud (Public, Private) zum Einsatz kommen kann, sind Datenschutz- und Datensicherheitsanforderungen zu berücksichtigen.

Je mehr Stufen der Datenaufbereitung und Analyse erforderlich sind, desto höher werden die Komplexität und damit die Aufwände für das Deployment.

9.1.7 Agile Vorgehensweise

Agile Vorgehensweisen sind heutzutage in der klassischen Softwareentwicklung und in Business-Intelligence-Projekten weit verbreitet. Insbesondere bei BI-Projekten sind kurze Teilprojektzyklen bzw. Sprints notwendig, da die betriebswirtschaftlichen Fragestellungen der Fachabteilungen der Geschäftsdynamik unterliegen. In Bezug auf Big-Data-Lösungen stellen agile Vorgehensweisen (vgl. S. 39) die Unternehmen vor Herausforderungen.

Reports und Self-Service BI

Bei der Entwicklung von Auswertungen im Rahmen einer Big-Data-Lösung entdecken Unternehmen oft neue und zuvor nicht direkt erkannte Zusammenhänge Die Kombination von verschiedenen Datenquellen, die Gewinnung von Erkenntnissen aus diesen Verknüpfungen ziehen neue Fragestellung nach sich, die in Reports abgebildet werden. Eine Herausforderung in diesem Zusammenhang ist die Bereitstellung der analytischen Fähigkeit der Gesamtlösung. Gerade Fachbereiche, die mit den Datenmengen konfrontiert werden, benötigen meist eine Abstraktionsschicht oder eine Qualifizierung der Datenbestände.

Architektur

Bei einer agilen Vorgehensweise besteht, gerade bei Big-Data-Projekten, die Gefahr, eine nicht weitreichende Architektur zu wählen. Weitblick über etwaige Neuanforderungen und Erweiterungen sollte in die Gesamtlösung einfließen. Hierdurch wird gewährleistet, dass das Gesamtbild nicht verloren geht und zukünftige Anforderungen umgesetzt werden können, ohne einen Umbau der Architektur zu veranlassen. Durch kurze Sprints entstandene Fehlentscheidungen in Big-Data-Projekten lassen sich aufgrund der Größe der Gesamtlösung meist nur schwer und aufwändig korrigieren.

Infrastruktur

Bedingt durch agile Entwicklungszyklen ergeben sich gerade bei Big-Data-Lösungen starke Auswirkungen auf die Bereitstellung von Infrastruktur, Applikationen und deren Entwicklung und Betrieb relevanten Mitarbeitern. Werden beispielsweise in einem Scrum-Sprint kurzfristig neue Datenquellen angebunden, werden kurzfristig

[57] z. B. für Anwendungen zum Monitoring sozialer Netzwerke

Ressourcen benötigt. Entscheidet sich ein Unternehmen beispielsweise in einer Iteration zusätzlich zur bereits integrierten Social Web Analyse von Twitter eine neue Datenquelle wie Facebook anzubinden, sollte die notwendigen Datenverbindung, der erforderliche Speicherplatz und die Rechnerkapazität kurzfristig zur Verfügung stehen.

9.1.8 Big-Data-Factory

Je nach Wiederverwendungsgrad einer Big-Data-Lösung kann die Gesamtarchitektur und Infrastruktur in einem Factory-Ansatz gebündelt werden. Hierbei erkennen Unternehmen den Mehrwert, eine Fragestellung zu einem späteren Zeitpunkt erneut zu analysieren. Möchte beispielsweise ein Automobilunternehmen bei der Markteinführung eines neuen Modells die Reaktionen im Social Web analysieren, werden kurzfristig und wahrscheinlich auch nur temporär die relevanten Ressourcen benötigt. Nach einem definierten Zeitraum wird die Big-Data-Lösung bis zur nächsten Markteinführung nicht mehr beansprucht. Die gesamte Architektur und Infrastruktur wird in einer Big-Data-Factory gekapselt und kann beliebig gestartet oder gestoppt werden.

Weiterer Vorteil dieses Ansatzes im Vergleich zu einem individuellen Deployment ist die Vervielfältigungsmöglichkeit der Factory. Bei parallelen Markteinführungen kann der Automobilhersteller verschiedene Instanzen betreiben und die Daten entsprechend analysieren.

Vorrausetzung für den Einsatz einer Big-Data-Factory ist die entsprechende Verfügbarkeit der notwendigen Ressourcen. Hierzu zählen die Anbindung der Datenquellen, die Infrastruktur, die Applikationen und die zugehörigen Mitarbeiter.

9.1.9 Reifegrad der Enterprise Architecture

Wesentlicher Faktor beim Deployment einer Big-Data-Lösung ist der jeweilige Reifegrad der Enterprise Architecture in einem Unternehmen.[58] Meist setzt der Aufbau und Betrieb einer Big-Data-Lösung voraus, dass das Unternehmen einen ganzheitlichen Blick auf ihre Informationstechnologie hat. Gerade für einen Big-Data-Factory-Ansatz ist ein gewisser Abstraktionsgrad notwendig. Sowohl die erforderliche Agilität der Infrastruktur und der Prozesse sowie die skalierbare Architektur stellen das Rückgrat eines Big-Data-Projekts da.

9.2 Betriebsmodelle

Für den Betrieb von Big-Data-Lösungen, insbesondere unter Berücksichtigung der extremen Anforderungen an die Skalierbarkeit der Systeme, die aus den Dimensionen des Deployments abgeleitet sind, bieten sich die in der Abbildung 12 dargestellten Betriebsmodelle an.[59]

Je nach Anforderung sind verschiedene Betriebsmodelle für Big Data möglich.

Die im Abschnitt 9.1 beschriebenen zwei Kategorien von Big-Data-Anwendungen sind auch und besonders für das Betriebsmodell relevant. Kategorie eins erlaubt in der Regel lediglich das Betriebsmodell »Infrastruktur als Service«, denn bereits die eingesetzte Software entspricht nicht notwendigerweise den üblichen Standards und ihr Einsatz ist eher einem Test als einem fortlaufenden Betrieb gleichzusetzen.

[58] Vgl. [EAM, 2011]
[59] Der Betrieb von Big-Data-Lösungen wird in der Phase 8 (Optimierung) des Big-Data-Vorgehensmodells adressiert (vgl. Unterabschnitt 5.2.8).

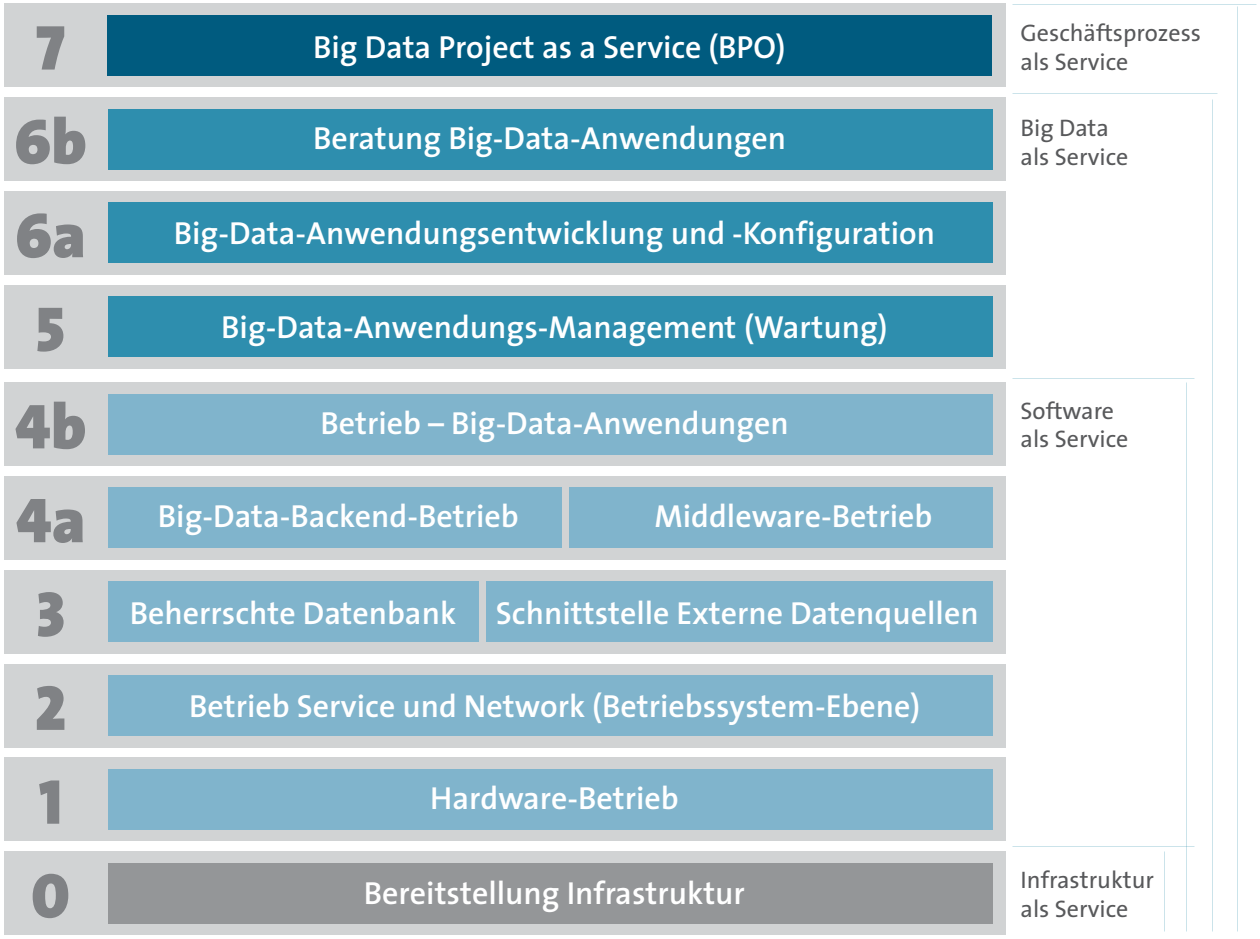


Abbildung 12: Betriebsmodelle für Big Data

9.2.1 Infrastruktur als Service

Die Beschaffung, Programmierung/Konfiguration und Implementierung erfolgt durch das eigene Unternehmen. Auch der Betrieb der Rechner, und die Wartung der Software werden in Eigenregie durchgeführt. Lediglich die Hardware wird als Infrastruktur-Service zugekauft.

Dieses Modell bietet sich an, wenn Experten im Bereich Hardware/Betriebssysteme, Software, Anwendungsbetrieb und Entwicklung zur Verfügung stehen, und die Lizenzen als Vermögenswerte (Assets) geführt werden. Kapazitätsengpässe im Rechenzentrum motivieren zur externen Vergabe des Infrastruktur-Service.

9.2.2 Software als Service

Die Beschaffung, von Hard- und Software und das Hosting wird als Software-Service zugekauft. Der Lieferant implementiert auch die in Eigenregie erstellte/beschaffte Software. Die Programmierung/Konfiguration und Implementierung erfolgt durch das eigene Unternehmen.

Dieses Modell bietet sich an, wenn unternehmensspezifisches Know-How nicht an Dritte herausgegeben werden soll. Es erfordert ein hohes Maß an Koordination mit dem Service-Lieferanten. Auch ist die Erweiterung des eigenen Kompetenzteams um externe Berater denkbar.

9.2.3 Big Data als Service

In diesem Modell werden alle Aspekte der IT zugekauft. Die Geschäftsanforderungen werden mit dem Lieferanten besprochen, der im Gegenzug über vereinbarte Medien (z. B. Entscheidungsvorlagen, Berichte) das eigene Unternehmen beliefert.

Dieses Modell bietet sich an, wenn die erwarteten Informationstypen grundsätzlich einheitlich sind (z. B. Statistiken von Konsumentenverhalten) und diese nur auf einen Service/auf eine Marke bezogen werden sollen. Der Lieferant kann hier Synergien mit ähnlichen Projekten herstellen.

9.2.4 Geschäftsprozess als Service

In diesem Fall führt ein Lieferant Geschäftsprozesse im Namen des eigenen Unternehmens aus.

Dieses Modell bietet sich an, wenn der Zweck des Big-Data-Projektes wiederkehrende Maßnahmen (z. B. Werbekampagnen) sind, oder über definierte Angebote direkt[60] an Konsumenten übermittelt werden. Dies kann z. B. aufgrund vorhandener Netzwerkstrukturen effizienter geliefert werden, als vom eigenen Unternehmen.

[60] z. B. in einem sozialen Netzwerk

10 Anhang

10.1 Privacy Impact Assessment – Checkliste

A – Projekt/Idee

A.1	Beschreibung des Projekts:
A.2	Wie sieht das Endprodukt aus?
A.3	Zu beteiligende Organisationseinheiten/ Personen:

Die Beschreibung soll so detailliert wie möglich sein. Dies soll den aktuell Beteiligten sowie den später hinzukommenden ermöglichen, ein genaues Bild vom Projekt und Endprodukt zu bekommen. Auch die beteiligten Abteilungen sind hier aufzuführen.

B – Personenbezogene Daten

Ein Privacy Impact Assessment hat den Schutz der persönlichen Daten zum Ziel. Daher ist bei jedem Projektvorhaben zu Beginn und dann im weiteren Verlauf zu prüfen, ob personenbezogene Daten verwendet werden.

B.1	Personenbezogene Daten	■ Name, Anschrift ■ E-Mailadresse, Passwort ■ Geb.-Datum ■ Bankverbindung, Kreditkartendaten ■ Rufnummern ■ Rechnungsdetails ■ _____ ■ _____
B.2	Besondere personenbezogene Daten	■ Religion ■ Ethnische Herkunft ■ Politische Meinung (Partei- oder Gewerkschaftszugehörigkeit) ■ Sexualleben ■ Gesundheitsdaten Nach der EU-Grundsatz-VO sind folgende Daten besonders zu beachten und hinsichtlich möglicher Risiken zu bewerten: ■ Daten über Kinder ■ Genetische und biometrische Daten ■ Standortdaten ■ Personenbezogene Daten aus umfangreichen Datenbeständen (Big Data)
B.3	Wie erfolgt die Datenverarbeitung?	☐ Erheben ☐ Speichern ☐ Übermitteln Beachte: Übermittlung ist die Datenweitergabe an einen Dritten, nicht an einen Auftragsdatenverarbeiter ☐ Verändern
B.4	Wo erfolgt die Datenverarbeitung?	☐ im Unternehmen ☐ im Konzern Beachte: kein Konzernprivileg im Datenschutz ☐ in der EU/im EWR ☐ außerhalb der EU
B.5	Wohin sollen die Daten gegeben werden?	☐ im Unternehmen ☐ im Konzern Beachte: kein Konzernprivileg im Datenschutz ☐ in der EU/im EWR ☐ außerhalb der EU
B.6	Meinung der Betroffenen[61]	

[61] Art. 33 Abs. 4 EU VO

C – Grundlage der Datenverwendung

Aufgrund des sogenannten Verbotes mit Erlaubnisvorbehalt dürfen Daten nur verwendet werden, wenn hierfür eine Rechtsgrundlage gegeben ist[62].

C.1	Einwilligung des Betroffenen	☐ liegt vor ☐ liegt nicht vor → weiter mit C. 2
C.2	Eigener Geschäftszweck	☐ ja; folgender _____ ☐ nein → weiter mit C. 3
C.3	Abwägung berechtigtes Interesse der verantwortlichen Stelle und schutzwürdiges Interesse des Betroffenen Bei der Abwägung ist sowohl das Prinzip der Datensparsamkeit als auch die Möglichkeit zur Anonymisierung zu berücksichtigen.	☐ berechtigte Interesse der verantwortlichen Stelle überwiegt ☐ schutzwürdiges Interesse des Betroffenen überwiegt → Datenverwendung nur mit Einwilligung oder wenn C. 4 gegeben ist
C.4	Anonymisierung der Daten	Beschreibung des Anonymisierungsverfahrens (ggf. auf gesondertem Blatt darstellen): _____ ☐ durch internes Gutachten bestätigt ☐ durch externes Gutachten bestätigt

D – Risiken der Datenverwendung

An jedem Meilenstein des Projekts ist zu überprüfen, ob die erforderliche Datenverarbeitung Risiken des Datenverlustes oder des Datenmissbrauchs birgt. Diese Risiken bedeuten nicht das Ende des Projekts, sondern führen zur Klärung und Prüfung möglicher Maßnahmen (vgl. E – Maßnahmen zur Risikominimierung).

D.1	Projektzeitpunkt	☐ Planung ☐ Abstimmung ☐ Umsetzung
D.2	Mögliches Risiko	☐ Datenverlust Kurze Beschreibung: _____ ☐ Datenmissbrauch Kurze Beschreibung: _____ ☐ Missbrauch durch Kombination mit Zusatzwissen (beim Dritten oder Auftragsdatenverarbeiter) Kurze Beschreibung: _____ ☐ anderes Risiko Kurze Beschreibung: _____
D.3	Grad des Risikos	Hohes Risiko: Konsultationspflicht (Art. 34 Abs. 2 EU VO)

[62] Die Regelungen sind anzupassen, wenn die endgültige Fassung der EU-Datenschutzverordnung vorliegt.

E – Maßnahmen zur Risikominimierung

E.1	Maßnahmen	☐ höherer Aggregierungs-/Anonymisierungsgrad Kurze Beschreibung: _____ ☐ technische Maßnahme Kurze Beschreibung: _____ ☐ andere Maßnahme Kurze Beschreibung: _____

10.2 Big-Data-Maturity-Modell

Nr.	Level	Charakteristika
0	Legacy Applications	■ Das Unternehmen hat bislang keine Big-Data-Aktivitäten begonnen oder Erfahrungen in diesem Umfeld gesammelt.
1	Big-Data-Initiativen	■ Erste Big-Data-Initiativen wurden gestartet. ■ Überlegungen zur Verwendung von Big Data zur Optimierung von Geschäftsbereichen oder der IT-Infrastruktur werden angestellt. ■ Erste Big-Data-Kompetenzen sind vorhanden.
2	Big Data Center of Excellence	■ Im Unternehmen sind bereits weitgehende Big-Data-Initiativen gestartet worden. ■ Erste Verantwortlichkeiten für die Umsetzung der Big-Data-Strategie sowie zur Verankerung der zukünftigen Big-Data-Projekte sowie der neuen Daten (Governance) wurden festgelegt. ■ Big-Data-Projekte zur Optimierung von Geschäftseinheiten sind etabliert.
3	Process Modeling	■ Erste Big-Data-Projekte befinden sich vor dem Abschluss. ■ Die strategischen Planungen zu Big Data gehen in konkrete Geschäftsprozess-Implementierungen über. ■ Big Data ist fester Bestandteil der Unternehmensstrategie. ■ Erste Big-Data-Lösungen und Ansätze wurden im Rahmen von PoCs erfolgreich implementiert und getestet.
4	Big Data Execution	■ Erste Big-Data-Aktivitäten wurden bereits erfolgreich durchgeführt. ■ Erste Geschäftsprozesse wurden durch Big Data optimiert oder sogar aufgesetzt. ■ Big Data ist Grundlage für die zukünftige Entwicklung neuer Geschäftsprozesse sowie der Optimierung bestehender Geschäftsprozesse. ■ ILM ist als integrale Aufgabe zur Pflege und Optimierung von Big Data etabliert.
5	Optimizing Big Data	■ Big Data ist ein integraler Bestandteilteil der Unternehmensstrategie. Konzepte wie Big-Data-Governance und ILM sind etabliert. ■ Die IT-Infrastruktur nutzt die Möglichkeiten durch Big Data aus, wichtige Elemente des Enterprise Information Managements sind auf Basis von Big-Data-Technologien abgebildet. ■ Viele Geschäftsprozesse basieren auf Big Data. ■ Herausforderungen sind die Optimierung der Big-Data-Infrastruktur und Geschäftsprozesse.

10.3 Aufbau eines Big Data Center of Excellence

Unternehmen streben danach, Entscheidungen immer besser zu unterstützen und dafür die verfügbaren Informationen umfassender auszuwerten. Bei einer Ausweitung der Datentypen und quellen wird es für die Fachabteilungen schwieriger, die erforderliche Big-Data-Kompetenz selbst aufzubauen und weiterzuentwickeln.

Auf einer bestimmten Stufe (vgl. Abbildung 9) wird es sinnvoll sein, die Experten für die Datenanalyse in den Fachabteilungen durch den Aufbau eines Big Data Center of Excellence zu unterstützen.

Ein Big Data CoE versetzt das Unternehmen in die Lage, mit höherer Geschwindigkeit und Konsistenz auf den Informationsbedarf der Fachabteilungen zu reagieren. Letztlich erhalten die Big-Data-Initiativen im Unternehmen mit einem CoE eine sinnvolle organisatorische Absicherung.

Das Big Data CoE ist ständiger Impulsgeber, unterstützt Geschäftsbereiche und IT bei der Entwicklung und Umsetzung einer Big-Data-Strategie und bündelt die mobilisierbaren Ressourcen mit Blick auf die Unternehmensziele. Das CoE unterstützt den bestmöglichen Einsatz von Informationen als Vermögenswert.

Zu seinen Aufgaben zählen auch die Bereitstellung der Tools sowie die Weiterbildung der Mitarbeiter aus den Fachabteilungen für deren zielgerichtete Nutzung.

Aufgaben eines Big Data CoE

Das Big Data CoE ist eine Fachabteilung, die technisches und fachliches Wissen zusammenführt, um Informationen bereitzustellen, analytische Fähigkeiten zu verbessern und Informationslösungen anzubieten. Es verbindet effizient Systeme und Daten und versetzt die Organisation in die Lage, Big Data konsistent und effektiv im Gesamtunternehmen einzusetzen.

Zusammengefasst bietet das Big Data CoE bestmögliche Unterstützung bei der konsistenten Nutzung von Daten, Werkzeugen, Techniken, gemeinsamen Architekturen, bewährten Vorgehensweisen und Ressourcen sowie beim Kompetenzaufbau der Fachabteilungen. Es orchestriert alle am Informationsmanagement Beteiligten, trägt zur Akzeptanz unternehmensweiter Datenstandards bei und erhöht den Businessbeitrag der Investitionen in Big Data.

Vorteile eines Big Data CoE

Mit einem etablierten CoE lassen sich einige der besonderen Herausforderungen von Big Data adressieren:

- Erhöhung der Effektivität der Experten
- Einführung einer Governance (Strukturen und Prozesse)
- Co-Nutzung von Technologieinvestitionen
- Verbesserung der Nutzerunterstützung in den Fachabteilungen
- Konsistentere und genauere Antworten auf geschäftliche Fragestellungen

Das Big Data CoE trägt zu einer einheitlichen Sicht auf Entscheidungssituationen bei, indem es bestimmte Ressourcen, Prozesse und Architekturen zentralisiert. Es verhindert Redundanzen innerhalb des Unternehmens, die nicht selten zu inkonsistenten Sichten führen. Die Fachabteilungen werden zu Konsumenten gemeinsamer Prozesse und müssen nicht mehr alle Datenanalyse allein durchführen; sie fokussieren sich auf Prozesse, die speziell für ihre Abteilung relevant sind und stützen sich bei ihren Entscheidungen auf für alle verbindliche Daten.

Effektivitätsverbesserung der Wissensarbeiter

Fachabteilungen können ihre Produktivität deutlich erhöhen, indem sie die Ausbildung und Unterstützung ihrer Wissensarbeiter verbessern und sie dabei unterstützen, Daten besser zu interpretieren und neue Erkenntnisse zu gewinnen. Gegenwärtig verbringen Analysten oft 50 bis 70 Prozent ihrer Zeit mit Aufgaben außerhalb der Datenanalyse. Ohne angemessenes Training zum Verständnis der Daten und der Werkzeuge entwickeln viele Analysten

eigene Wege zur Datenakquisition. Die Nachteile eines solchen Vorgehens sind offensichtlich: hoher Aufwand, fehlende Wiederholbarkeit, mangelhafte Konsistent für andere Benutzer.

Ein Big Data CoE kann Bildungs- und Zertifizierungsprogramme entwickeln, um den Wissensarbeitern ein angemessenes Training für Daten- und Werkzeugnutzung zu bieten. Analysten entwickeln sich so von technischen Datensammlern zu Lösungsanbietern weiter.

Die Vereinheitlichung der Werkzeuge über die Big-Data-Anwendungen hinweg, ermöglicht es, die Wartungskosten zu reduzieren und bessere Vertragskonditionen mit den Werkzeuglieferanten auszuhandeln.

Weitere Aufgaben von Big Data CoE

Unternehmensspezifisch sollten auch folgende Aspekte beachtet werden:

- Big-Data-Governance:
  Big-Data-Governance sollte integraler Teil eines unternehmensweiten Information Governance Programms sein. Für Big-Data-Aspekte wie z. B. Social Web sind Verantwortliche zu benennen.

- Integration von Aufgaben:
  Abstimmung von Big Data mit den Information Governance Disziplinen wie Metadaten, Masterdaten, Datenqualität, Datenschutz, Datensicherheit.

- Optimierungsstrategien:
  Entwicklung von Konzepten, was bei der Umsetzung von Big-Data-Projekten zu beachten ist – Metadaten, Datenqualität, Information Lifecycle Management – und wie diese zu einer kontinuierlichen Optimierung der Big-Data-Infrastruktur beitragen.

- Policies:
  Wie müssen Organisationen und Prozesse angepasst werden und wie müssen sich Mitarbeiter verhalten, wenn Daten und Applikationen für Big Data integriert werden.

- Daten als Assets:
  Da Big Data bzw. die vorhandenen Daten zunehmend als unternehmerisches Gut verstanden werden, gilt es, diese Bedeutung für die einzelnen Unternehmensbereiche herauszuarbeiten.

- Schlichter:
  Big Data kann dazu führen, dass es innerhalb eines Unternehmens unterschiedliche Interessen und Meinungen gibt, wie mit möglichen Daten umgegangen werden kann. So kann beispielsweise die Marketingabteilung Interesse an der Nutzung sensitiver Kundendaten für Marketingaufgaben haben, während die Rechtsabteilung deshalb Kundenklagen befürchtet. Das Big Data CoE muss hier als Schlichter auftreten und unterschiedliche Interessen austarieren.

10.4 Technische und organisatorische Ansätze für eine anonyme Verarbeitung und Speicherung von personenbezogenen Daten

Kontext des Praxisbeispiels

Der Abschnitt 10.4 beschreibt ein Praxisbeispiel für die Anonymisierung: Es werden technische und organisatorische Ansätze für eine wirksame Anonymisierung von personenbezogenen Massendaten für deren Verwendung innerhalb von Big-Data-Anwendungen erörtert. Darunter fällt auch der Begriff der Standortdaten und umfasst dabei jegliche Ortsinformationen, die durch eine Nutzung von Diensten in öffentlichen Kommunikationsnetzen, wie beispielsweise Mobilfunk- oder WLAN-Netzen sowie durch die Nutzung spezifischer satellitengestützter Ortungsdienste, entstehen können.

Hierbei wird auch die Möglichkeit zur Einbeziehung von Langzeitaussagen berücksichtigt. Die vorgestellten Ansätze orientieren sich dabei an den Anforderungen des BDSG, TKG und TMG sowie der bisherigen Rechtsprechung zu diesem Thema. Gleichzeitig werden auch aktuelle Entwicklungen entsprechender EU-Verordnungen berücksichtigt.

Je umfassender und feingranularer die Datenbasis, desto mehr Möglichkeiten bestehen für die Kombination der einzelnen Datensätze und desto vielfältiger werden folglich die Schlüsse, die daraus gezogen werden können. Vor diesem Hintergrund ist eine möglichst breite und aggregationsfreie Speicherung der Daten «wünschenswert«. Darin besteht aber zugleich die Gefahr, dass die zur Verfügung stehenden Informationen so umfangreich sind, dass sich trotz einer wirksamen Anonymisierung der einzelnen Datensätze, aus der Kombination vieler Datensätze doch wieder individuelle Eigenschaften oder Profile ableiten lassen und somit ein De-Anonymisierung möglich wird.

Der Fokus aller im Abschnitt 10.4 vorgestellten Ansätze liegt dementsprechend in der Generierung einer möglichst umfassenden Datenbasis als Grundlage für diverse (potenzielle) Big-Data-Anwendungen unter Berücksichtigung aller datenschutzrechtlichen Anforderungen. Der grundlegende Anonymisierungsansatz ist dabei sehr allgemein gehalten und kann somit für jegliche personenbezogene Daten Anwendung finden. Die weiteren Ausführungen gehen auf Maßnahmen gegen De-Anonymisierung sowie Möglichkeiten zur Bildung von Langzeitaussagen ein. Da diese Argumentationen stark von der zugrundeliegenden Datenbasis abhängen, wird hier aus Gründen der Übersichtlichkeit im Speziellen auf besondere Herausforderungen bei der Verarbeitung von Standortdaten eingegangen.

Eingeordnet in eine beispielhafte Gesamtarchitektur können alle hier vorgestellten Prozesse innerhalb der Übergangs- bzw. Transformationsschicht, also zwischen den ursprünglichen Rohdatenquellen und der finalen Datenbasis, für verschiedene Big-Data-Anwendungen eingeordnet werden.

Überblick

Die nachfolgenden Erörterungen zeigen dabei zunächst, wie durch Kombination passender technischer und organisatorischer Maßnahmen eine kennzahlenbasierte Anonymisierung realisiert werden kann, so dass die resultierenden anonymen Datensätze auch weiterhin untereinander in Bezug gesetzt werden können (vgl. Unterabschnitt 10.4.1). Im Anschluss wird darauf eingegangen, welche besonderen Herausforderungen bei einer langfristigen Speicherung von Standortdaten alleine aufgrund ihrer speziellen Charakteristik bestehen und wie dabei dennoch sichergestellt werden kann, dass keine direkte (vgl. Unterabschnitt 10.4.2) oder indirekte (vgl. Unterabschnitt 10.4.3) De-Anonymisierung auf Grundlage der Datenbasis erfolgt. Im Abschnitt 10.4.4 wird gezeigt, dass trotz aller datenschutzrechtlich notwendigen Einschränkungen dennoch die Möglichkeit besteht, wichtige Aussagen über längere Zeiträume zu berechnen. Ein abschließendes Fazit (vgl. 10.4.5) fasst die Erkenntnisse zusammen.

10.4.1 Umsetzung einer kennzahlenbasierten Anonymisierung

Das wesentliche Potenzial von Big-Data-Anwendungen liegt in der Möglichkeit auf eine möglichst breite Datenbasis zuzugreifen und anhand jeweils passender Kombination der vorhandenen Daten Antworten (Aussagen) auf diverse Fragestellungen liefern zu können. Die Bezugsmöglichkeit zwischen verschiedenen Datensätzen spielt somit eine entscheidende Rolle. Das bedeutet: Datensätze, die ursprünglich anhand einer bestimmten Kennzahl unter einander zuordenbar waren, sollen auch im anonymisierten Zustand anhand einer entsprechend Kennzahl zuordenbar bleiben.

Neben dem grundsätzlichen Weglassen aller Attribute, die an sich einen Rückschluss auf das Individuum zulassen würden[63], ist für eine effektive Anonymisierung entscheidend, dass keine Rückschlussmöglichkeit anhand der anonymisierten Kennzahl auf die ursprüngliche personenbezogene Kennzahl besteht. Nach dem BDSG ist eine zulässige Anonymisierung gegeben, wenn ein Rückschluss nur mit einem »unverhältnismäßig« hohen Aufwand möglich wäre.

[63] z. B. Nummer des Personalausweises

Um einen solchen unverhältnismäßig hohen Aufwand sicherzustellen, sind technische und häufig auch organisatorische Maßnahmen notwendig.

So ist eine personenbezogene Kennzahl nach Verschlüsselung durch ein geeignetes Hash-Verfahren zunächst ein Pseudonym. Nur wenn sichergestellt werden kann, dass ein Rückschluss auf den Ausgangswert einen unverhältnismäßig hohen Aufwand darstellt, gilt die verschlüsselte Kennzahl auch als anonymisiert. Dies stellt insbesondere eine Herausforderung dar, wenn der für das Hash-Verfahren eingesetzte Schlüssel für einen gewissen Zeitraum konstant sein muss.

Theoretisch wäre hier ein Rückschluss möglich, indem das Hash-Verfahren (ggf. wiederholt) durchlaufen und dabei jeweils der Input (personenbezogene Kennzahl) und der erzeugte Output (verschlüsselte Kennzahl) in einer sogenannten Referenzliste zusammen gespeichert werden. Der Einsatz wirksamer organisatorischer Maßnahmen kann den Aufwand für die Erzeugung einer solchen Zuordnungstabelle deutlich erhöhen. Eine unternehmensinterne Vergabe entsprechender Zugriffsrechte begründet nicht unbedingt den geforderten unverhältnismäßigen Aufwand, da die Trennung von Kennzahl und Pseudonym nicht in hinreichendem Umfang gegeben wäre. Eine organisatorische Aufteilung auf zwei Unternehmen könnte einen entsprechenden unverhältnismäßigen Aufwand darstellen. Dabei muss jedoch technisch sichergestellt werden, dass innerhalb des jeweiligen Unternehmens entweder ausschließlich die personenbezogene Kennzahl oder ausschließlich die verschlüsselte Kennzahl bekannt ist, also unterschiedliche Datenbestände mit jeweils nur einer Version der Kennzahl vorliegen.

Ist diese Voraussetzung erfüllt, kann von einem unverhältnismäßigen Aufwand für eine De-Anonymisierung ausgegangen werden.

10.4.2 Sicherung gegen direkte De-Anonymisierung

Bei der Verarbeitung von Daten besteht ggf. die Gefahr, dass es unabhängig von einer erfolgreich anonymisierten Kennzahl (vgl. Kap. 10.4.1) zu einer direkten De-Anonymisierung kommen kann, sofern eine einzelne Information eindeutig oder relativ eindeutig einem einzelnen Individuum bzw. einer kleinen Gruppe von Individuen zugeordnet werden kann. So könnte man in einer sehr dünn besiedelten Region Standortdaten, die beispielsweise durch die Nutzung eines GPS- oder Mobilfunkservices vorliegen, mit Daten aus öffentlichen Verzeichnissen[64] kombinieren und damit einen Rückschluss auf die tatsächliche Identität einer Person ziehen.

Um eine solche Zuordnung auszuschließen, muss technisch für jeden Standort sichergestellt werden, dass die anfallenden Standortdaten stets von einer gewissen Mindestanzahl an Individuen stammen, bevor sie für die Speicherung innerhalb der finalen Datenbasis oder für eine Nutzung in weiteren vorgelagerten Verarbeitungsstufen (vgl. Kapitel 10.4.4) freigegeben werden. Entscheidend ist dabei auch die Festlegung eines bestimmten Zeitraums (Prüfzeitraum), innerhalb dessen die entsprechende Mindestanzahl erreicht werden muss. Liegen am Ende des Prüfungszeitraums nicht genügend Standortinformationen von verschiedenen anonymen Kennzahlen vor, müssen alle bis dahin angefallenen Daten verworfen oder in einen allgemeineren Kontext überführt werden. Zur Verallgemeinerung kann dabei beispielsweise die Genauigkeit der Ortsinformation gesenkt werden. Dieses Verfahren kann bei jeglicher Art von personenbezogenen Daten Anwendung finden, um direkte De-Anonymisierung zu vermeiden.

[64] wie z. B. dem Melderegister

10.4.3 Sicherung gegen indirekte Re-Anonymisierung

Wie im Abschnitt 10.4.1 erörtert, liegt das Potenzial von Big-Data-Anwendungen in einer für den jeweiligen Anwendungsfall spezifischen Kombination von Daten. Dieses Potenzial ist umso ergiebiger, wenn die anonymen Kennzahlen bzw. die verwendeten Schlüssel zur Erzeugung dieser Kennzahlen für einen gewissen Zeitraum konstant sind. Das heißt: Datensätze, die ursprünglich personenbezogen waren, dann anonymisiert wurden, sollen auch im anonymisierten Zustand entsprechende, über Kennzahlen verknüpfte Aussagen ermöglichen. Dies kann erreicht werden, indem über einen gewissen Zeitraum dieselbe nicht personenbezogene Kennzahl kombinierbar ist.

Durch die Bezugsmöglichkeit besteht bei der Verarbeitung und Speicherung entsprechend anonymisierter Daten jedoch gleichzeitig die Gefahr, dass deren Kombination ein (relativ) eindeutiges Muster erzeugt. Eine De-Anonymisierung wäre nun indirekt möglich, sobald das erzeugte Muster ein eindeutiges, in der Realität nachvollziehbares Profil einer bestimmten Person erzeugt. Diese Gefahr steigt mit zunehmendem Datenumfang im Zeitverlauf, gerade unter dem Aspekt, dass damit die Möglichkeit zur Identifizierung wiederkehrender Muster immer wahrscheinlicher wird.

Sobald die Anonymität aufgehoben ist, können alle Daten des Individuums identifiziert und dadurch ggf. unerlaubt weitere personenbezogene Informationen gewonnen werden. Umfang und Detailgrad dieser Informationen und somit auch das Missbrauchspotenzial steigen dabei ebenfalls mit zunehmender Dauer der Beobachtungsmöglichkeit.

Die Zeitdauer, innerhalb derer eine Zuordnung zwischen einzelnen Datensätzen erfolgen kann, spielt folglich sowohl für die Möglichkeit der indirekten De-Anonymisierung durch Profilbildung als auch für den Umfang der anschließenden Missbrauchspotenziale eine entscheidende Rolle. Eine sinnvolle zeitliche Beschränkung der Bezugsmöglichkeit einzelner Datensätze stellt somit ein wichtiges Grundprinzip für eine anonyme Verarbeitung von Standortdaten dar. Technisch wird dieser Anforderung durch einen regelmäßigen Wechsel des Schlüssels zur Erzeugung der anonymen Kennung entsprochen. Die so entstehende Datenbasis ermöglicht Big-Data-Anwendungen zwar weiterhin eine Kombination der anonymisierten Standortdaten, dies aber nur über einen begrenzten, für Profilbildung sowie Missbrauch unzureichenden, Zeitraum (kurzfristiger Bezugszeitraum).

10.4.4Ermöglichung von Langzeitaussagen

Der zuvor erörterte kurzfristige Bezugszeitraum (Kap. 10.4.3) ist nicht nur für die finale Speicherung, sondern auch für jegliche vorgelagerte Berechnung maßgeblich. Das bedeutet: Jegliche Aussagen auf Grundlage von personenbezogenen Daten können maximal für den kurzfristigen Bezugszeitraum getroffen werden. Dabei spielt es keine Rolle, an welcher Stelle[65] die Berechnung dieser Aussagen erfolgt.

Eine durchgängige Anwendung dieses Prinzips stellt die Anonymität des Individuums bei der Nutzung von personenbezogenen Daten sicher. Aber gerade Aussagen auf Grundlage von langfristigen Beobachtungen sind häufig sehr wertvoll für Big-Data-Anwendungen[66]. Auswertungen über längere Zeiträume ermöglichen dabei häufig sehr viel realistischere Aussagen, da sie das Erkennen und Berücksichtigen von Unregelmäßigkeiten[67] erlauben.

Die Anforderung, langfristige Aussagen trotz eines beschränkten Bezugszeitraums zu ermöglichen, wird durch das Grundprinzip der Erzeugung von aggregationsbasierten Langzeitindizes gelöst. Dabei spielt die

[65] vorgelagert zur Erzeugung der Datenbasis oder auf Grundlage der finalen Datenbasis
[66] z. B. um anhand von Standortdaten stark frequentierte Orte erkennen zu können
[67] z. B. besonderen Ereignissen

Tatsache, dass eine bestimmte Langzeitaussage stets einen einzelnen aggregierten Wert darstellt, eine wichtige Rolle. Anstatt diesen Wert direkt aus einer Vielzahl, über einen langen Zeitraum gesammelter personenbezogener Daten abzuleiten, erfolgt die Berechnung auf Grundlage mehrerer bereits aggregierter Werte, die jeweils anhand eines einzelnen kurzfristigen Bezugszeitraums ermittelt wurden. Im Rahmen der technischen Umsetzung wird dabei für eine bestimmte, vorher zu definierende Fragestellung anhand der Daten für jeden kurzfristigen Bezugszeitraums genau ein aggregierter Wert ermittelt. Dieser Wert repräsentiert eine statistische Häufigkeits- oder Wahrscheinlichkeitsaussage für den jeweiligen kurzfristigen Bezugszeitraum. Durch Verschlüsselungstechniken können diese Kurzzeitaussagen anschließend über einen längeren Zeitraum in Bezug gesetzt werden. Auf Grundlage der Relationen zwischen mehreren Kurzzeitaussagen können schließlich die gesuchten Langzeitaussagen abgeleitet werden.

Im Ergebnis handelt es sich bei allen Langzeitaussagen somit immer um einzelne, wiederum aggregierte Werte, die jeweils eine statistische Häufigkeit oder Wahrscheinlichkeit repräsentieren. Zum Erhalt der Anonymität ist dabei nicht nur die technische sondern auch eine organisatorische Trennung zwischen Kurzzeit- und Langzeitberechnungen wichtig. Dabei ist davon auszugehen, dass auch hier erst eine Aufteilung der Berechnungsschritte auf zwei Unternehmen den nötigen Tatbestand des unverhältnismäßigen Aufwands erfüllt.

Die angesprochenen Verschlüsselungstechniken ermöglichen nun diese Aussagen in Bezug zu den, entsprechend Kapitel 10.4.3, in kurzen Zeitabständen wechselnden anonymen Kennungen der anderen personenbezogenen Daten zu setzen.

10.4.5 Fazit

Die im Abschnitt 10.4 vorgestellten Ansätze zeigen, dass eine sinnvolle Kombination aus technischen und organisatorischen Maßnahmen eine Anonymisierung von personenbezogenen Daten erlaubt, bei der auch im anonymisierten Zustand Bezugsmöglichkeiten zwischen den einzelnen Datensätzen erhalten bleiben. Im Fall von Standortdaten können darüber hinaus durch eine ortsbasierte Filterung sowie einen regelmäßigen Wechsel des Anonymisierungsschlüssels auch direkte bzw. indirekte Rückschlüsse auf einzelne Individuen anhand der Datenbasis effektiv verhindert werden. Trotz dieser datenschutzrechtlich notwendigen Einschränkungen können auf Grundlage von Wahrscheinlichkeitsberechnungen dennoch wertvolle Langzeitaussagen getroffen werden. Somit kann gezeigt werden, dass es möglich ist, unter Einbehaltung aller datenschutzrechtlichen Bestimmungen eine genügend umfassende Datenbasis für diverse (potenzielle) Big-Data-Anwendungen zu realisieren.

10.5 Quellen

[BITKOM, 2012] Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Leitfaden des BITKOM, Berlin 2012. http://www.bitkom.org/files/documents/BITKOM_LF_big_data_2012_online(1).pdf, abgerufen am 13.02.2013

[BSI, o.J.] Informationsschrift »Das Ende der Anonymität? Datenspuren in modernen Netzen«, https://www.bsi.bund.de/ContentBSI/Publikationen/Studien/anonym/dasEndederAnonymitaet.html (Abruf am 15.02.2013)

[EAM, 2011] Enterprise Architecture Management – neue Disziplin für die ganzheitliche Unternehmensentwicklung. Leitfaden, BITKOM 2011, http://www.bitkom.org/files/documents/EAM_Enterprise_Architecture_Management_-_BITKOM_Leitfaden.pdf

[EGAG 2012] Experton Group AG: MultiClient-Studie »Big Data 2012-2015«

[HP/Vertica, 2013]. The New Math: Return on Information (ROI) http://www.vertica.com/industries/, abgerufen am 13.02.2013

[RoScho, 2000] Roßnagel, Alexander; Scholz, Philip: Datenschutz durch Anonymität und Pseudonymität. In: »MultiMedia und Recht«, 2000, S. 721-731

[SOM, 2013] SevenOne Navigator Mediennutzung 2012. Studie der SevenOne Media GmbH, 2012. https://www.sevenonemedia.de/research_mediennutzung_navigator-mediennutzung

[TDWI, 2012] Whitepaper »Building the Business Intelligence Competency Center«, http://tdwi.1105cms01.com/whitepapers/2012/09/hp_building-the-business-intelligence-competency-center.aspx?tc=page0, abgerufen am 13.02.2013

[TDWI, 2013] TDWI Best Practices Report, »Achieving Greater Agility with Business Intelligence«, http://tdwi.org/research/2013/01/tdwi-best-practices-report-achieving-greater-agility-with-business-intelligence.aspx?tc=page0 (Abruf am 15.02.2013)

[DeGhe, 2004] Dean, Jeffrey; Ghemawat, Sanjay (Google, Inc.) (2004): »MapReduce: Simplified Data Processing on Large Clusters«, http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/de//archive/mapreduce-osdi04.pdf, Abruf am 07.03.2012

10.6 Autoren

Arnd Böken, Graf von Westphalen Rechtsanwälte Partnerschaft

Susanne Dehmel, BITKOM e.V.

Guido Falkenberg, Software AG

Norbert Franke, arvato systems GmbH

Dr. Thomas Jansen, DLA Piper UK LLP

Dr. Holger K. von Jouanne-Diedrich, Atos IT Solutions and Services GmbH

Ralf Konrad, T-Systems International GmbH

Holm Landrock, Experton Group AG

Dr. Mario Lenz, Empolis Information Management GmbH

Wulf Maier, Hewlett-Packard GmbH

Dr. Mark Mattingley-Scott, IBM Deutschland GmbH

Dr. Michael May, Fraunhofer IAIS Institut für Intelligente Analyse- und Informationssysteme

Bernd Mußmann, Hewlett-Packard GmbH

Dr. Andreas Ribbrock, Teradata GmbH

Barbara Schmitz, Telefónica Germany GmbH & Co. OHG

Christian Valerius, Hewlett-Packard GmbH

Jonathan Ukena, Telefónica Germany GmbH & Co. OHG

Dr. Carlo Velten, Experton Group AG

Friedrich Vollmar, IBM Deutschland GmbH

Dr. Mathias Weber, BITKOM e.V.

Weiter wirkten an der Entwicklung des Leitfadens mit

Jörg Bartel, IBM Deutschland GmbH

Christian Glatschke, Splunk Services Germany GmbH

Robert Guzek, Fujitsu Technology Solutions GmbH

Dr. Peter Katko, Ernst & Young GmbH Wirtschaftsprüfungsgesellschaft

Rolf Klapproth, Hewlett-Packard GmbH

Dr. Holger Kisker, Forrester Research GmbH & Co. KG

Daniel Leinius, Hewlett-Packard GmbH

Axel Mester, IBM Deutschland GmbH

Boris Andreas Michaelis, SAP Deutschland AG & Co. KG

Jürgen Urbanski, T-Systems International GmbH.

10.7 Sachwortregister

Der Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. vertritt mehr als 2.000 Unternehmen, davon über 1.200 Direktmitglieder mit etwa 140 Milliarden Euro Umsatz und 700.000 Beschäftigten. Hierzu gehören fast alle Global Player sowie 800 leistungsstarke Mittelständler und zahlreiche gründergeführte, kreative Unternehmen. Mitglieder sind Anbieter von Software und 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