Big Data

Das Öl des 21. Jahrhunderts[1]

Jonathan Gruber

HTWK Leipzig

19. April 2017

[1] Gartner Company, 2010

Inhaltliche Gliederung

1. Was ist Big Data?
2. Technologien & Wertschöpfungskette
3. Anwendungen
4. Kritik & Schattenseiten von Big Data
5. Fazit & Ausblick

1 Hintergrund & Begriff

- Begriff *Big Data*
- Einführende Beispiele
- Beschreibungsmodelle
- Historie von Big Data

Die digitale Ära

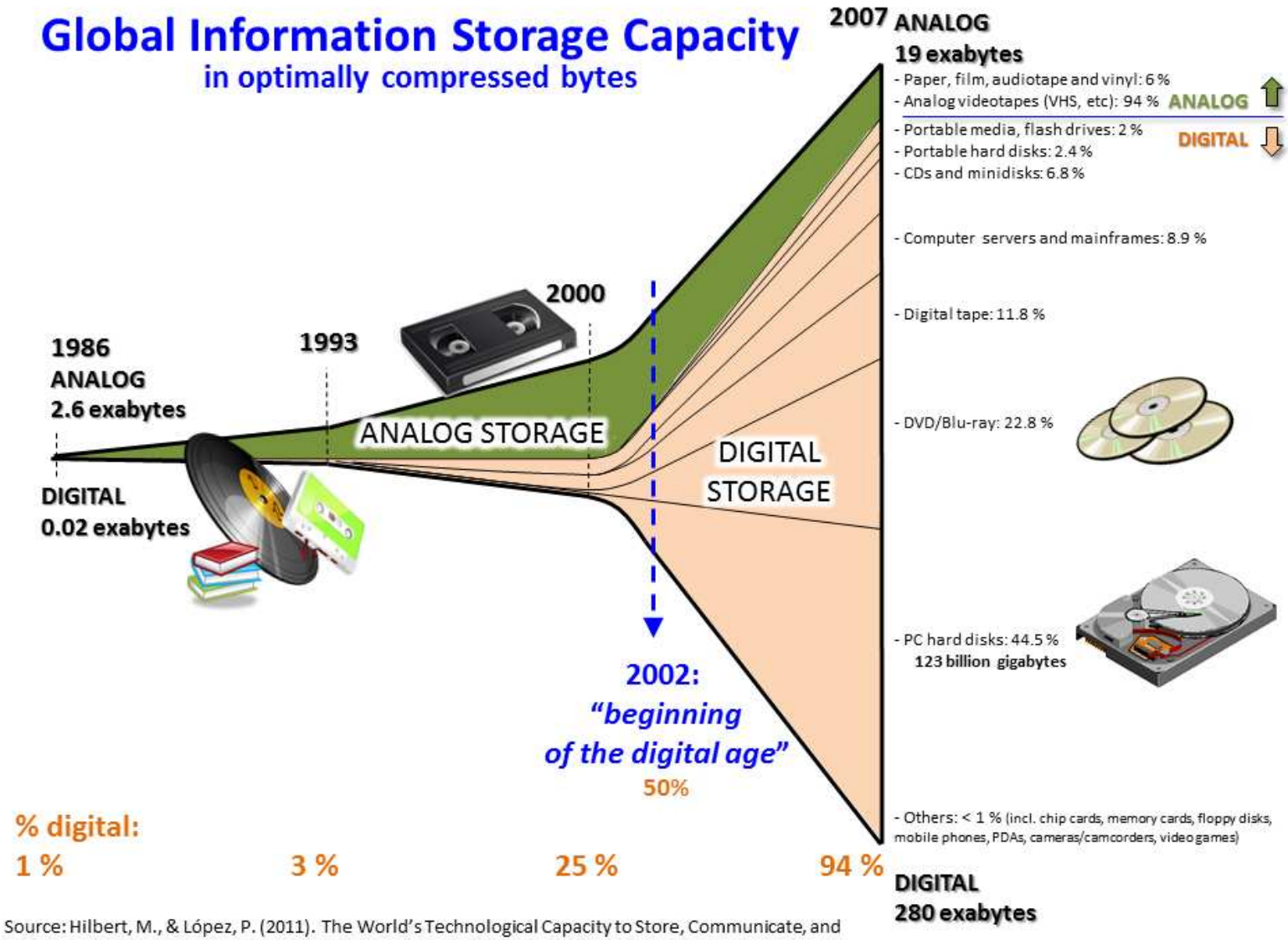
Wir sind im Zeitalter riesiger, digitaler Datenmengen angelangt:

- Google: mehr als 3 Mrd. Suchanfragen pro Tag. Verarbeitung von 1 Petabyte (10^{15} Bytes) täglich.
- Facebook: 10 Mio. Fotos pro Stunde
- Twitter: 400 Mio. Tweets pro Tag
- Enorme Menge an Sensordaten von Kameras, Smartphones, wissenschaftlichen Messgeräten, RFID-Chips u. a. Geräten des *Internets der Dinge*

Alle zwei Tage werden **1,8 Zettabyte** (10^{21} Bytes) an Daten generiert (Stand: 2011). Das entspricht der entspricht der **gesamten** Datenmenge der menschlichen Zivilisation bis 2003.

Quelle: [10]

Rasantes Wachstum



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Was ist Big Data?

- **Big Data**: Datensätze, die so gigantisch groß sind, dass sie mit herkömmlichen Datenverarbeitungssystemen nicht bewältigt werden können
- Daten sind dynamisch und typischerweise schwach bis gar nicht strukturiert
- Typischerweise werden Echtzeit-Analysen benötigt
- Das Datenvolumen *kann* über die Zeit wachsen
- Viele Herausforderungen: Enorme Anforderungen an Netzwerk-Infrastruktur, Speicher- und Rechenkapazität

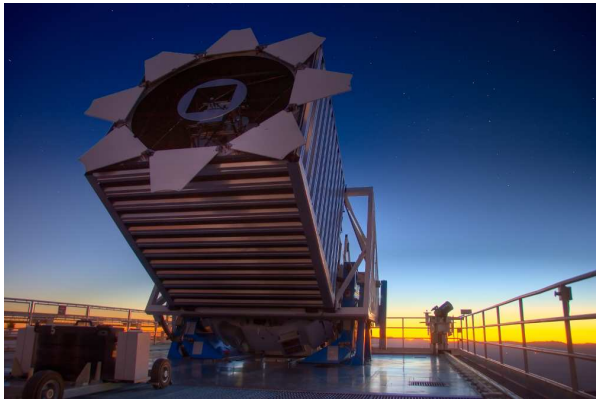
Aber: Noch ist der Begriff nur vage definiert und wandelt sich kontinuierlich. Es existieren mehrere Betrachtungsweisen, die unterschiedliche Aspekte hervorheben.

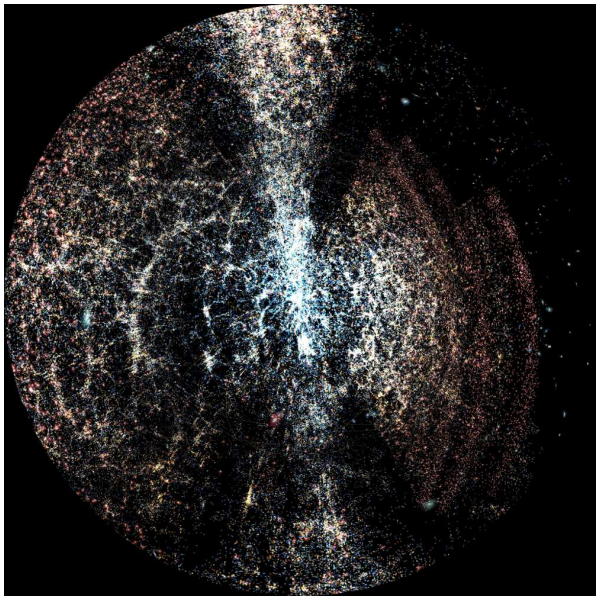
Zweck und Nutzen von Big Data

- Grundlegend: Erlangung neuer Erkenntnissen auf Grundlage der Daten (**Wertschöpfung**)
- Business Intelligence: Besseres Verständnis von Geschäftsprozessen, Prozessoptimierung, usw.
- Industrie 4.0 (Automatisierung)
- **Vorhersage** von Entwicklungen, Trends, Interessen (Amazon, Netflix, etc.)
- Marketing: Analyse von Benutzerverhalten, Erfolg Werbekampagnen
- **Forschung** (Medizin, Biologie, Meteorologie, Astronomie, uvm.)

Fazit: zumeist ökonomische Interessen, aber auch viele wissenschaftliche Anwendungen.

Beispiel 1: Sloan Digital Sky Survey (SDSS)





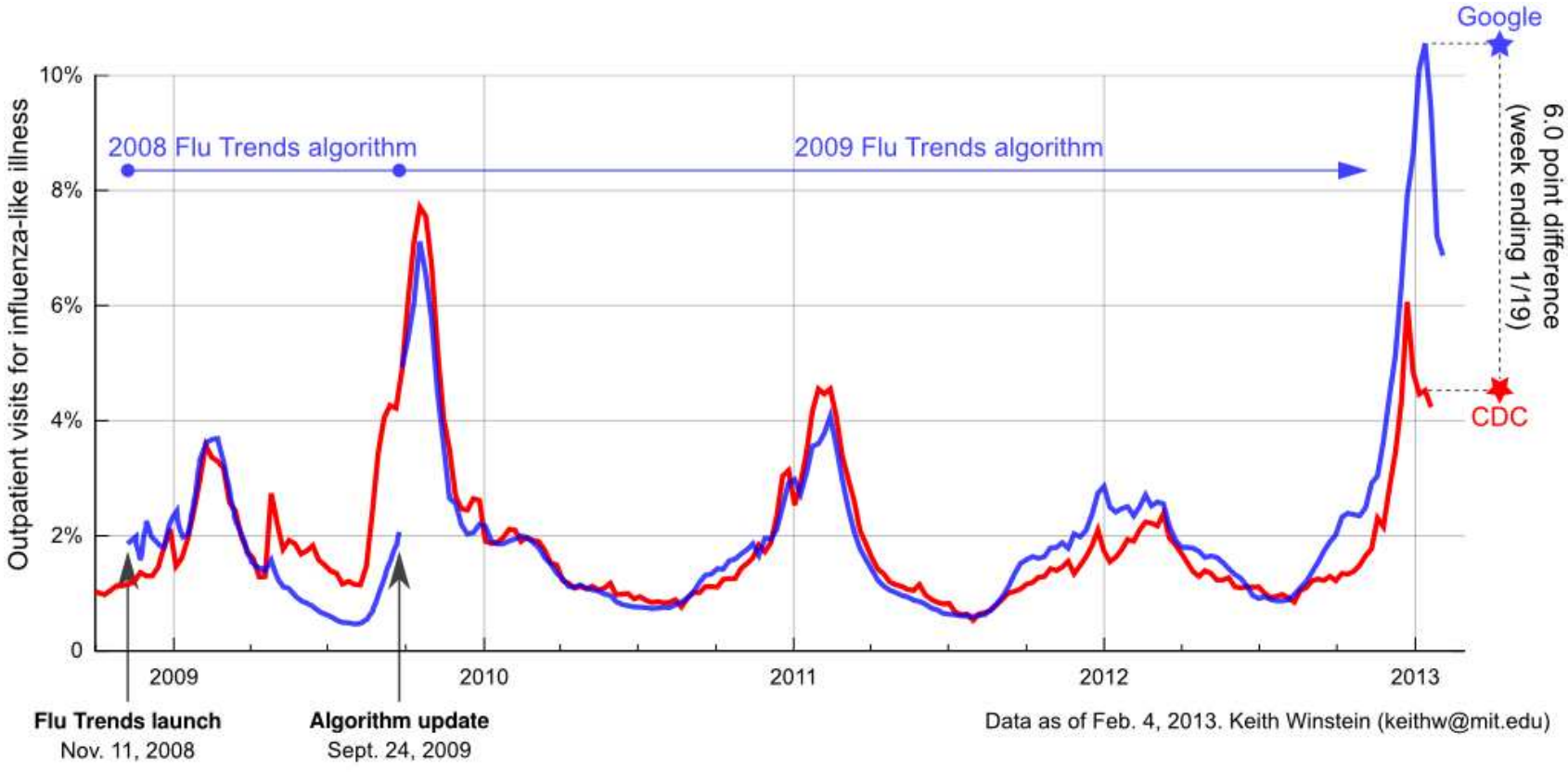
- Leistungsstarkes Teleskop in New Mexico (USA) zur *Durchmusterung*[2] des Sternenhimmels
- Gemeinschaftsprojekt der USA, Japan, Südkorea und Deutschland
- Beginn der Datenaufzeichnung im Jahr 2000: 200 GB Daten pro Nacht
- Bis 2011 entstehen so 140 TB Daten ($\approx$ 35% des Sternenhimmels) [13]
- 2016: Nachfolger *Large Synoptic Survey Telescope* sammelt diese Datenmenge (140 TB) alle **5 Tage**!

[2] Systematische Durchsuchung und Katalogisierung des Himmels

Beispiel 2: H1N1-Virus I

- Jedes Jahr erkranken weltweit 3 - 5 Millionen Menschen an saisonaler Grippe. Davon sterben 250.000 - 500.000 [11].
- 2009: Experten befürchten bis zu 10 Millionen Opfer durch neues H1N1-Virus (anfangs keine Impfung).
- *Jeder* Grippefall muss den Behörden gemeldet werden. Diese sind jedoch überfordert und in ihren Auswertungen immer etwa **zwei Wochen** hinterher.
- Idee von Google: Auswertung enorm vieler Suchanfragen (50 Mio. häufigste Suchwörter), um Verbreitung der Pandemie nahezu in Echtzeit analysieren zu können.

Beispiel 2: H1N1-Virus II



Sources: http://www.google.org/flutrends/us, CDC ILInet data from http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html,
Cook et al. (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic.

Beispiel 3: Frühgeburten I

- WHO: Etwa 10% aller Babys kommt zu früh auf die Welt. Tendenz: steigend.
- Hauptgrund für **Kindersterblichkeit** unter 5 Jahren. Weltweit jährlich 1 Mio. Tote (Stand 2015) [12].
- Ärzte z. T. machtlos bei zu später Diagnose von Infektionen.
- Idee von **Dr. Carolyn McGregor**[3]: Aufzeichnung der Vitaldaten der Babys durch Sensoren.
- Pro Sekunde werden 1.200 Datenpunkte gesammelt.

[3] University of Ontario, Institute of Technology, Canada

Beispiel 3: Frühgeburten II

- Mustersuche mit mathematischen Modellen zur Erkennung von Infektionen.
- Ergebnis: Nicht „Flattern“ der Vitaldaten, sondern paradoxerweise eine Stabilisierung dieser ist eindeutiger Hinweis auf Gefahr!
- Heute: Feststellung von Infektionen **24 Stunden im Voraus** mit hoher Wahrscheinlichkeit möglich.

Begriff: Das V-Modell I

3 Vs – Traditionelles Modell

Bereits 2001 von Laney, einem Analysten des Marktforschungsunternehmens Gartner in einem Forschungsbericht [9] definiert:

- **Volume**: Der riesige Umfang der generierten und gespeicherten Daten im Peta-, Exa- und Zettabytebereich (10^{15} - 10^{21} Bytes).
- **Velocity**: Die enorme Datenrate und ausreichende Geschwindigkeit bei der Datenverarbeitung und -analyse.
- **Variety**: Die Vielzahl unterschiedlicher Datenformate, -typen und -quellen.

4 Vs – Erweitertes Modell

- **Veracity**: Glaubwürdigkeit der Daten. Wie akkurat sind diese?

Begriff: Das V-Modell II

Außerdem: 2011 Erweiterung durch die International Data Corporation (IDC) [7], die Sinnhaftigkeit von Big Data zu unterstreichen:

- **Value**: Big Data hat großes Potential ist in seiner *Gesamtheit* wertvoll. Auf Mikroebene jedoch nur geringe Informationsdichte.

Begriff: Mehr – Unscharf – Korrelationen

Alternatives (aber ähnliches) Beschreibungsmodell aus Mayer-Schönberger und Cukier [10] (2013):

Mehr Paradigmenwechsel: Nicht mehr einige, wenige relevante Daten messen, sondern alles Sammeln, dann analysieren.

Unscharf Nicht alle Datenpunkte müssen akkurat sein: die Makroebene gleicht die Mikroebene aus.

Korrelation Relevant sind Zusammenhänge, nicht Kausalitäten. Nicht das *Warum* ist entscheidend, sondern das *Was*.

Korrelation impliziert keine Kausalität

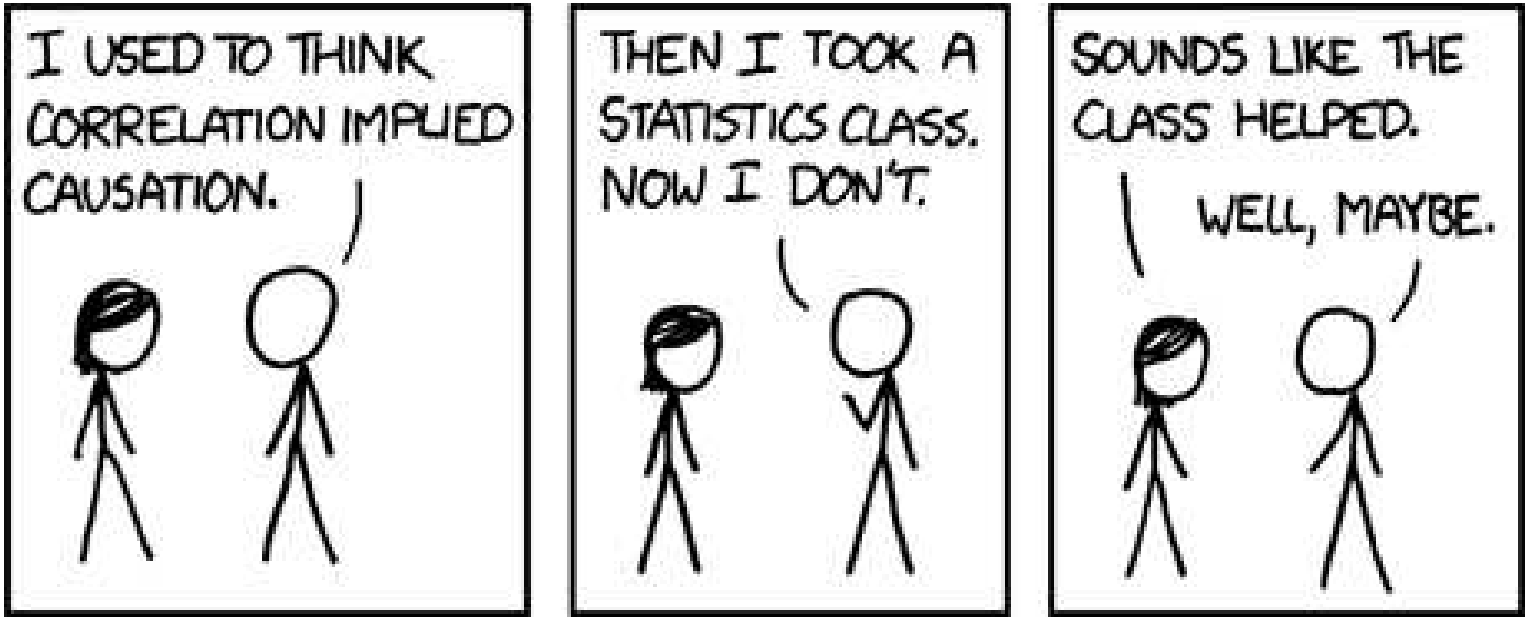


Abbildung: https://xkcd.com/552/

„Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'."

Quantität kann Qualität schaffen

Abbildung: „*The Horse in Motion*“ – Eadweard Muybridge, 1878

Big Data aus Anwendersicht

Das V-Modell beschreibt Big Data aus technischer Sicht. Im Gegensatz dazu beschreibt das F-Modell die **Anwendersicht**:

Fast Das Big-Data-System sollte Ergebnisse möglichst schnell bereit stellen. Flaschenhals: Heterogenität der Daten, verfügbare Ressourcen, Problemkomplexität.

Flexible Es muss mit geringem Aufwand möglich sein, das System an veränderte Bedingungen anzupassen (weitere Datenquellen, neue Algorithmen oder statistische Modelle).

Focused Nur relevante Datenquellen können ausgewählt werden.

„The Fourth Paradigm“

Jim Gray, 2007: Ein **fundamentaler Paradigmenwechsel** ist hinsichtlich Rechnerarchitektur und Datenverarbeitung notwendig [8]:

1. Experimentelle Naturwissenschaft:
   Beschreibung natürlicher Phänomene
2. Theoretische Naturwissenschaft:
   Kopernikus, Galilei, Newton, usw.
3. Simulationswissenschaft (Computational Science):
   Simulation komplexer Phänomene
4. **Data-Intensive Science**
   - Wissenschaftler überwältigt mit Datenflut
   - Neue Werkzeuge werden benötigt
   - Big Data als mögliche Lösung

2 Technologien & Wertschöpfungskette

- Herausforderungen für Big Data
- Grundlegende Technologien
- Wertschöpfungskette
  - Datengenerierung
  - Datenerfassung
  - Datenspeicherung
  - Datenanalyse

Herausforderungen für Big Data I

Big Data steht vielen Herausforderungen gegenüber, denen bisherige Technologien nicht gewachsen sind:

- Infrastruktur: Hohe Anforderung an Netzwerk, Rechen- und Speicherkapazität
- Datenrepräsentation: heterogene, unstrukturierte Daten müssen zur effizienten Verarbeitung geeignet repräsentiert werden
- Redundanzvermeidung / Kompression
- *Data life cycle*: Was wird gespeichert, was verworfen?

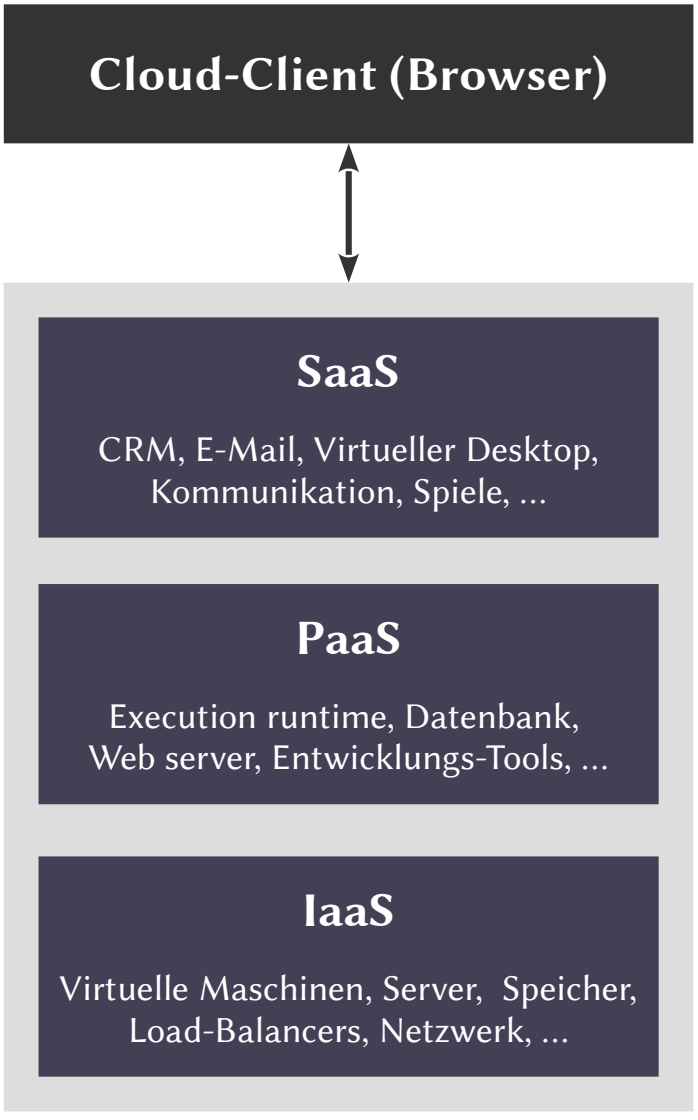
Herausforderungen für Big Data II

- Skalierbarkeit / Erweiterbarkeit
- Datenvertraulichkeit
- Energiemanagement
- Kooperation: Interdisziplinäre Forschung ist für die Entwicklung notwendig

Cloud Computing I

- Eng verknüpft mit Big Data
- Stellt die riesigen, benötigten Speicher- und Rechenkapazitäten als Dienstleistung zur Verfügung (**Infrastruktur**)
- Realisiert Skalierbarkeit
- **Verteilte Dateisysteme** ermöglichen Speicherung riesiger Datenmengen
- Die Entwicklung von Cloud Computing treibt Big Data voran und umgekehrt

Cloud Computing II

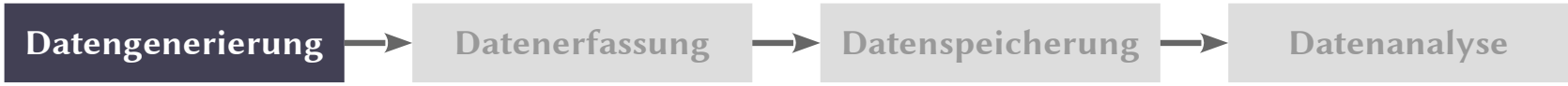


Rechenzentren (Data centers)

- Gewinnen in Zeiten von Big Data an großer Bedeutung
- Müssen Hardware-Infrastruktur bereitstellen für die Datenerfassung, -verarbeitung und -organisation
- Benötigen Hochgeschwindigkeitsnetzwerk (potentieller Flaschenhals)
- Senkung der operationalen Kosten
- Backups
- *Soft capacities*: Unterstützung der Entscheidungsträger, Erkennung von Problemen in betriebswirtschaftlichen Abläufen usw.

Fazit: Zunehmend mehr Verantwortung für die Rechenzentren

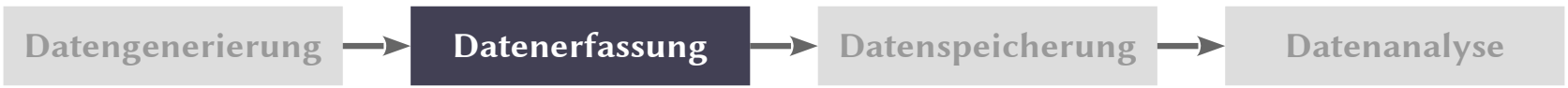
1. Phase – Datengenerierung



Datenquellen

- World Wide Web
- Kommunikationsdaten
- Geschäftsdaten
- Internet der Dinge
- Biomedizinische Daten
- Naturwissenschaftliche Experimente
- uvm.

2. Phase – Datenerfassung



Gliederung in drei Teilschritte:

1. Datenaneignung
2. Datenübertragung
3. Datenvorbehandlung (*data pre-processing*)

2. Phase – Datenerfassung: Datenaneignung

Zunächst: Auslesen der **rohen Daten** aus verschiedenen Quellen.

Typische Quellen / Methoden

- Auslesen von Logdateien (z. B. Web Server)
- Sensordaten: Audio, Video, meteorologische Daten, Druck, Vibration usw.
- Netzwerk-Traversierung mit Webcrawlern (Suchmaschinen, soziale Netzwerke)
- Mobile Daten (Smartphones)
- Wissenschaftliche Messgeräte
- ...

2. Phase – Datenerfassung: Datenübertragung

- Anschließend: Übertragung der Daten zum Rechenzentrum (oder andere Speicherinfrastruktur)
- Entscheidend: Übertragungsgeschwindigkeit, Performance

Inter-RN Übertragung

- Datenfluss zwischen (zusammenhängenden) RN

Intra-RN Übertragung

- Datenfluss *innerhalb* des RN
- Reorganisation der Daten als Vorbereitung für die Analyse oft notwendig

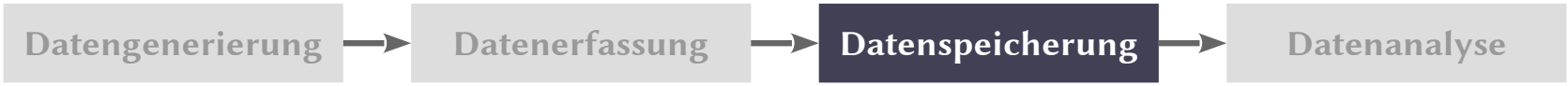
2. Phase – Datenerfassung: Datenvorbehandlung

- Notwendig, denn die Daten variieren in Rauschen (*Noise*), Redundanz und Konsistenz
- Datenqualität in manchen Anwendungen kritisch
- Kein triviales Unterfangen

Ziele

- **Datenintegration**: Kombination verschiedener Quellen, Datenextraktion und -transformation
- **Datenbereinigung**: Identifkation fehlerhafter oder unvollständiger Datensätze. Fehlerkorrektur durch Löschen oder Modifizieren der Daten
- **Redundanzvermeidung & Kompression**: Vermeidung unnötiger Datenübertragung durch Filter und Kompression

3. Phase – Datenspeicherung



Derzeitige Speichersysteme für Big Data können in drei Level mit zunehmender Abstraktion kategorisiert werden:

1. Verteilte Dateisysteme
2. NoSQL-Datenbanksysteme
3. Programmiermodelle

Verteilte Dateisysteme

- Grundlage für jegliche Big-Data-Speichersysteme
- Nach Jahren Entwicklung bereits recht ausgereift
- Dateien werden in Chunks (Fragmente) aufgeteilt und auf mehrere Knoten (*Nodes*) im Netz verteilt $\rightarrow$ Skalierbarkeit
- Redundanz ermöglicht Ausfallsicherheit
- Vertreter: Google File System (GFS), Hadoop Distributed File System (HDFS), Haystack (Facebook)

CAP-Theorem: Eigenschaft verteilter Systeme

Brewer, 2000 [1]: Maximal zwei der folgenden Eigenschaften können von einem verteiltem System realisiert werden:

- **C**onsistency: Dateninkonsistenz bei Serverausfall möglich (Daten über mehrere Server repliziert)
- **A**vailability: Alle Anfragen müssen stets, auch bei Serverausfällen, beantwortet werden
- **P**artition Tolerance: Ausfalltoleranz des Gesamtsystems bei Ausfall von Servern / Subnetzen

Beispiele

- AP – DNS oder Cloud Computing: Hohe Verfügbarkeit und Ausfallsicherheit. Teilweise: *eventual consistency*
- CA – RDBMS: Partitionierung von untergeordneter Bedeutung
- CP – Banking-Anwendungen: Konsistenz enorm wichtiger, als Ausfallsicherheit

Google File System (GFS) I

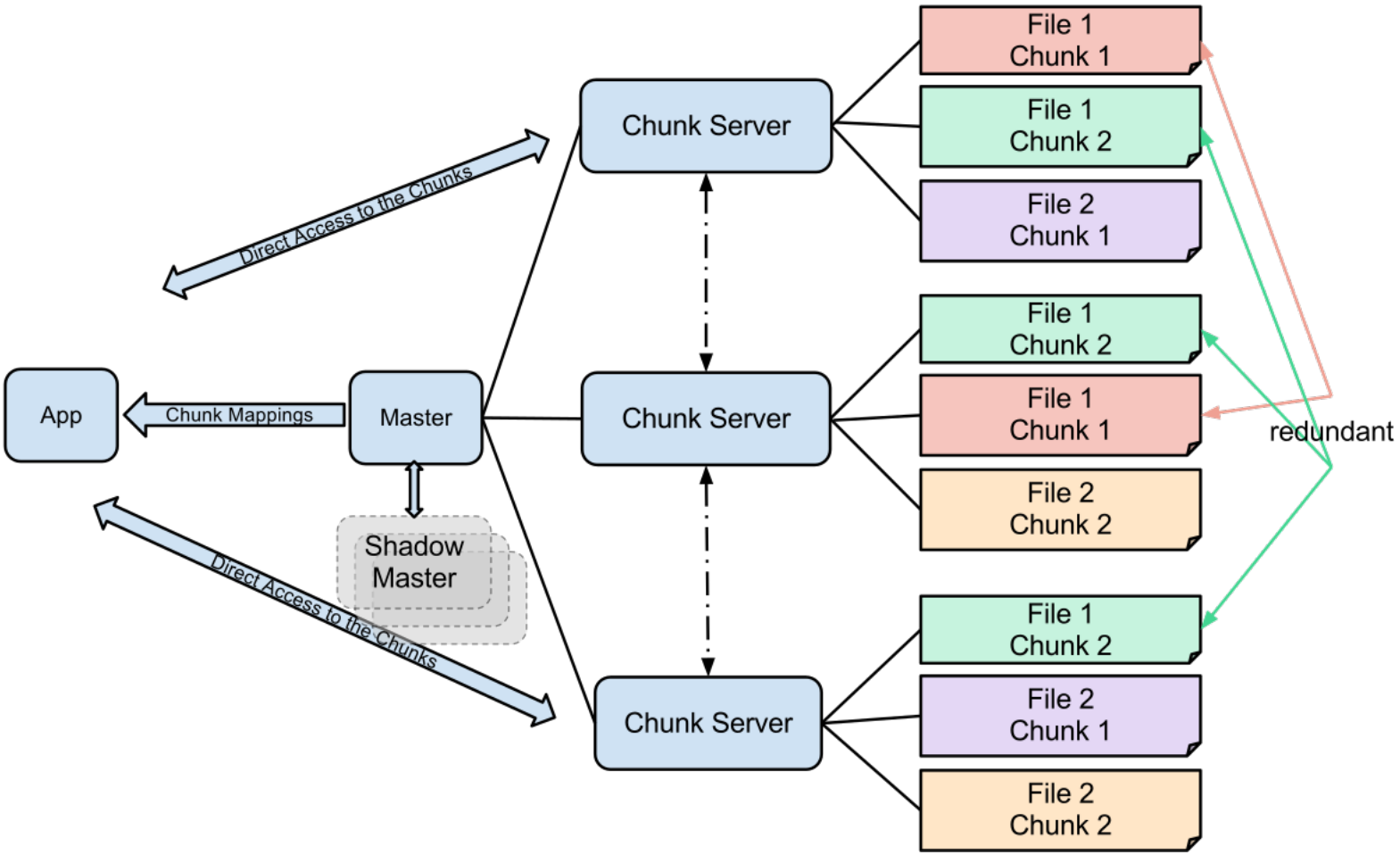


Abbildung: Konzeptioneller Aufbau von GFS

Google File System (GFS) II

- Proprietäres, verteiltes Dateisystem von Google (Linux)
- Optimiert für hohe Datendurchsätze und häufiges Lesen der Dateien. Löschen nur selten.
- Gut skalier- und erweiterbar
- Master und *n* Chunkserver. Chunkgröße: 64 MB
- Jedes Fragment mindestens drei mal **redundant** gespeichert
- Alle Anfragen gehen durch Master: dieser kennt Metadaten der Dateien und deligiert Zugriff auf diese (Mapping)
- Periodische „*heart-beat messages*“ des Masters an alle Chunkserver, um Metadaten aktuell zu halten

NoSQL-Datenbanksysteme

- Klassische relationale Datenbanksysteme (RDBMS) sind für Big Data ungeeignet
- Vielzahl an Neuentwicklungen, sog. *Not only SQL* Datenbanken. Teilweise in Kombination mit RDBMS.
- Gelten als **Schlüsseltechnologie** für Big Data [2]
- Klare Abgrenzung zu bisherigem relationalem Modell. Manche ACID[4]-Eigenschaften werden (bewusst) verletzt.
- Fähig gigantische Datenmengen zu verarbeiten
- Schemafreiheit
- Simple API
- Unterstützung für viele Programmiersprachen

[4] Atomicity, Consistency, Isolation, Durability

Verschiedene NoSQL-Datenbanksysteme I

Key-Value-Datenbanken

- einfaches Datenmodell: „primitiver“ Wert wird eindeutigem Schlüssel zugeordnet
- Gute Erweiterbarkeit und schnelle Antwortzeiten
- Wegbereiter Amazon Dynamo: Hohe Serververfügbarkeit durch redundante Datenhaltung auf n Servern
- Weitere Vertreter: Berkeley DB ($\rightarrow$ Oracle NoSQL), Redis

Verschiedene NoSQL-Datenbanksysteme II

Dokumentenorientierte Datenbanken

- Wie Key-Value-Datenbank, aber komplexe Werte (Objekte) möglich
- Unterstützung für horizontale Skalierbarkeit
- Oft RESTful HTTP-API
- Vertreter
  - MongoDB: Binary JSON (BSON)
  - Apache CouchDB: SQL Support
  - SimpleDB: JSON

Verschiedene NoSQL-Datenbanksysteme III

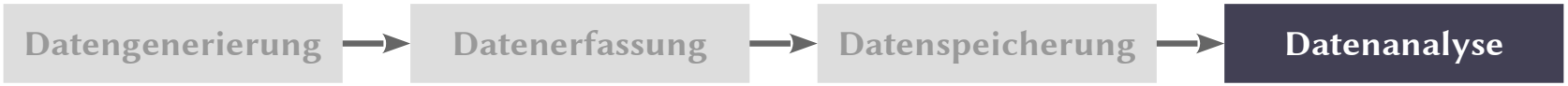
Spaltenorientierte Datenbanken

- Inhalte werden physikalisch nicht zeilen- sondern spaltenweise abgespeichert
- Spalten und Zeilen werden horizontal im Cluster verteilt
- Vorteile: Hohe Effizienz beim Datenzugriff, leichte Aggregatbildung $\rightarrow$ *Data Warehouse*
- Wegbereiter: Google **BigTable** (proprietär)
- Weitere Vertreter: Apache Cassandra (Facebook, jetzt Open Source), HBase (Bestandteil Hadoop), Hypertable

Relationale vs. NoSQL-Datenbanken

	RBDMS	**NoSQL**
Paradigma	relational	nicht-relational, verteilt
Schema	statisch	dynamisch
Daten	strukturiert, homogen	unstrukturiert, heterogen
Query-Sprache	SQL	verschiedene
Komplexe Abfragen	ja	eher nicht
Skalierbarkeit	vertikal	vertikal & horizontal

4. Phase – Datenanalyse



- Traditionelle Datenanalyse
- Big-Data-Datenanalyse
- Big-Data-Plattform
- Programmiermodelle

Datenanalyse – Methoden I

Traditionelle Datenanalyse (Auswahl)

- Clusteranalyse: finden von Ähnlichkeitsstrukturen (Clustern)
- Korrelationsanalyse: Identifikation von Zusammenhängen
- Regressionsanalyse: Modell zur Bestimmung der Beziehung zwischen Variablen, Prognose von Werten
- Statistische Analyse: Wahrscheinlichkeitstheorie
- **Data-Mining**
  - Erkennen von von Mustern, Querverbindungen und Trends
  - Extraktion versteckter „Werte“
  - Einsatz von künstlicher Intelligenz, Machine Learning

Datenanalyse – Methoden II

Big-Data-Analysemethoden

- Bloom Filter: äußerst effiziente Hashtabelle
- Hashing / Tries (Präfixbäume): effizientes Lesen / Schreiben
- Index: reduziert Festplattenzugriffe, jedoch zusätzlicher Speicherbedarf
- **Parallel Computing** → MapReduce

Programmiermodelle: MapReduce I

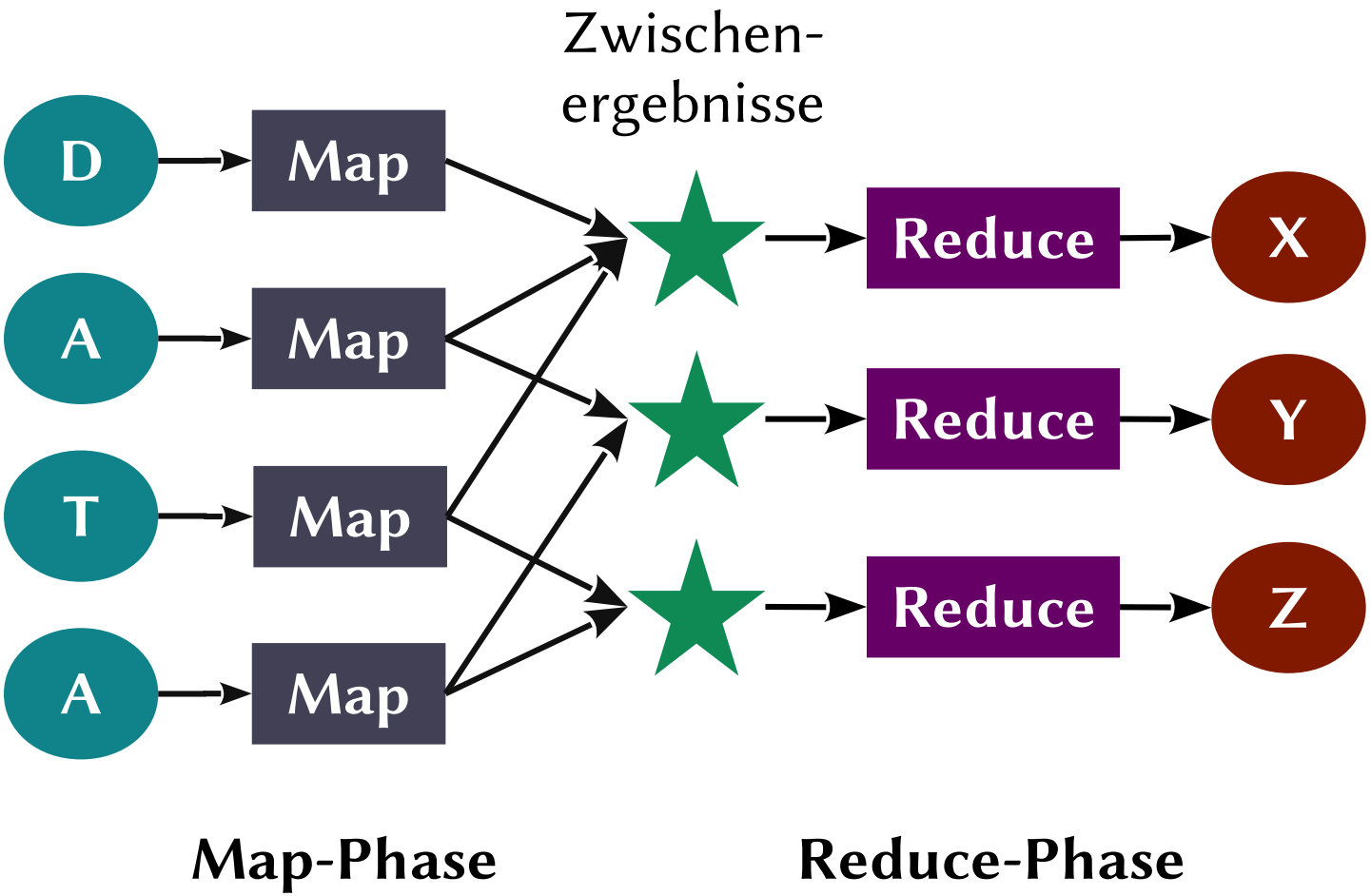
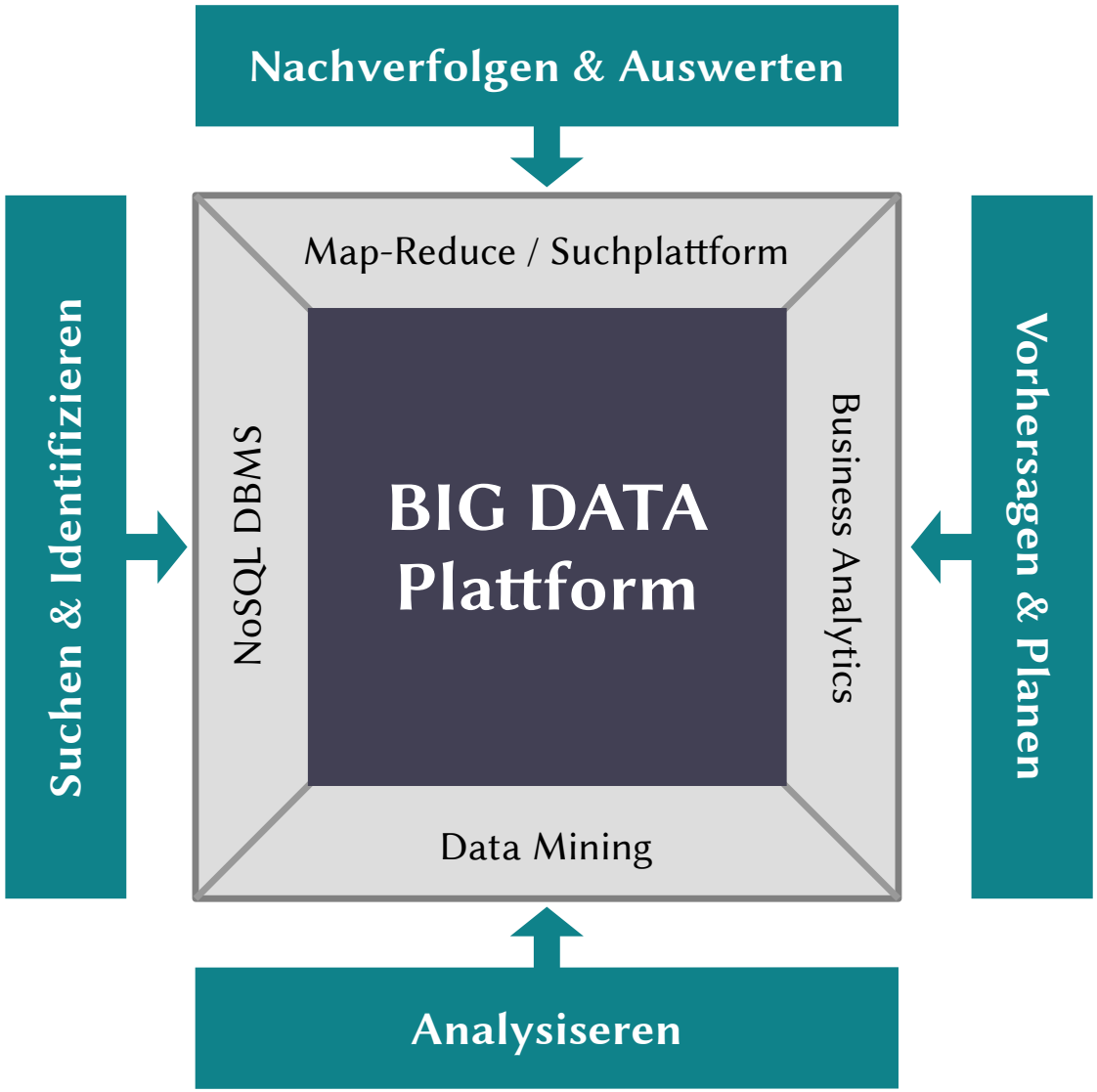


Abbildung: Konzept von MapReduce

Programmiermodelle: MapReduce II

- Finden von Informationen in verteilten Systemen durch Parallelverarbeitung
- 2010 von Google eingeführt: Daten im Petabytebereich
- Vermeidung von konkurrierenden Zugriffen durch Erzeugung und Weiterverarbeitung neuer Daten
- Drei Phasen
  - **Map**: Partitionierung und Abarbeitung der Daten nach fachlichen Kriterien und temporäre Speicherung der Zwischenergebnisse
  - Shuffle: Zuordnung Map-Zwischenergebnisse zu Reduce-Knoten
  - **Reduce**: Rekombination der Zwischenergebnisse zu Gesamtlösung

Komponenten einer Big-Data-Plattform [6]



Apache Hadoop

- Beliebtes Open-Source Framework, das eine Big-Data-Plattform realisiert
- Setzt auf handelsübliche Hardware
- Ausfälle dieser werden erwartet und entsprechend abgefangen
- Komponenten
  - Hadoop Common: gemeinsame Bibliotheken und Werkzeuge
  - Hadoop Distributed File System (HDFS): verteiltes Dateisystem
  - Hadoop YARN[5]: Ressourcen-Management
  - Hadoop MapReduce: Implementierung von MapReduce

[5] *Yet Another Resource Negotiator*

3 Big-Data-Anwendungen (Auswahl)

- Business & E-Commerce
- Social Media
- Gesundheitswesen
- Crowdsourcing
- **Internet der Dinge**: Smart Cities

Business & E-Commerce

- **Marketing**: Vorhersage von Kundenverhalten, Erschließung neuer Business-Modelle
- **E-Commerce**: Wettbewerbsfähige Preise durch Analyse konkurrierender Anbietern (z. B. Amazon), Realisierung von Suche in gigantischen Produktdatenbanken
- **Personalwesen**: bessere Planung von Einstellungen (und Kündigungen)
- **Finanzwesen**: Analyse des Kaufverhaltens, Finanztransaktionen zur Einschätzung der Kreditwürdigkeit (Scoring)

Social Media

- Betrachtung von Instant Messaging, Mikroblogging, Fotos usw.
- Analytische Methoden zum Finden von Korrelationen in Netzwerkstruktur
- Zwei Grundmethodiken:
  - inhaltliche Analyse: Sprach- und Textanalyse
  - strukturelle Analyse: Graphentheorie (Benutzer als Knoten, soziale Verknüpfungen als Kanten)
- Ziel: Analyse der Interessen, Beziehungen, Verhaltensmuster, Demografie usw. der Mitglieder des Netzwerks

Gesundheitswesen

(vgl. Einleitung)

- Analyse und Vorhersage von Pandemien (H1N1)
- Auswertung medizinischer Daten (Vitaldaten von Frühgeborenen)
- ...

Crowdsourcing

- Begriff angelehnt an *Outsorcing*
- Kollektive Problemlösung durch Verteilen von Teilaufgaben an freiwillige Helfer
- Beispiele
  - Google Maps „Local Guides“ mit Punktesystem
  - BOINC[6]: Software-Plattform für verteiltes Rechnen. Freiwillige Bereitstellung der eigenen Rechenleistung zur Lösung komplexer Probleme.

[6] Berkeley Open Infrastructure for Network Computing

Internet der Dinge (IdD)

- engl. *Internet of Things* (IoT)
- Definition der ITU[7]: „*A global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things [...]*“
- Gliederung in drei Ebenen: Sensor-, Netzwerk- und Anwendungsebene
- Eigenschaften:
  - Simple (numerische) oder auch komplexe (multimediale) Daten
  - Heterogenität
  - Zeit und Ort wichtig
  - Viel redundante, unnötige Information

[7] International Telecommunication Union

Städte der Zukunft: Smart Cities

Smart Cities [5] I

- IBM: Bis 2050 werden 70% der Menschheit in Städten leben [3]
- **Smart City**: Sammelbegriff für gesamtheitliche Entwicklungskonzepte für moderne Städte
- Grundidee: intelligente Ressourcenplanung, frühzeitiges Erkennen von Problemen, Vorhersage von Ereignissen, ...
- Technologien: Sensornetzwerke, Internet der Dinge, Cloud Computing, Big Data
- Aktuelle Forschungsprojekte: z. B. „*Morgenstadt*“[8] der Fraunhofer-Gesellschaft

Smart Cities [5] II

Ziele

- Verkehrssteuerung
- Stadtplanung
- Überwachung von Umweltfaktoren
- Besseres Energiemanagement (*Smart Grid*)
- Öffentliche Sicherheit

Kontroverse

- **Datenschutz** und Transparenz muss gewährleistet werden
- verantwortungsvoller Umgang mit den Daten, klare rechtliche Regelungen
- Missbrauch ausschließen

[8] http://www.morgenstadt.de

4 Kritik & Schattenseiten

- Datenschutz
- Ethische Fragestellungen
- Gesellschaftliche Verantwortung

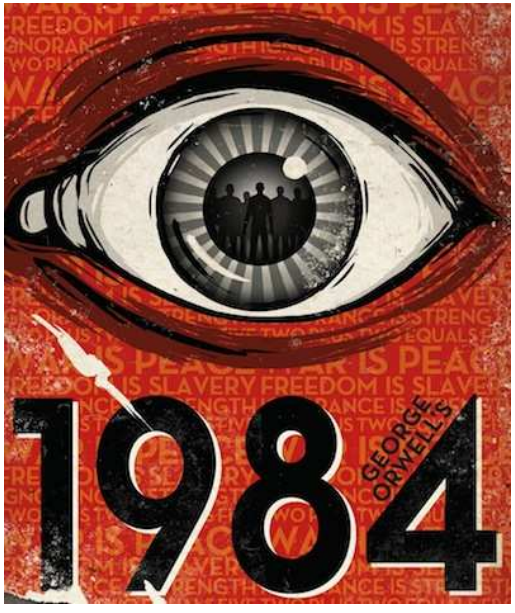
Problematischer Datenschutz I

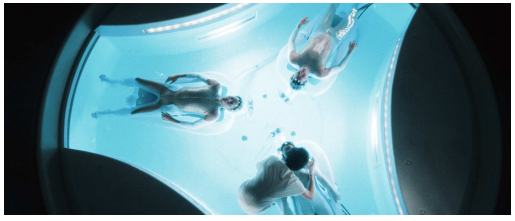
- Großes Problem: Gefahr des **„gläsernen Bürgers“**
- Jede Interaktion im **Internet** hinterlässt Spuren und wird aufgezeichnet (Social Media, Messenger, Clouddienste, Google Analytics, Weblogs, Benutzerdaten wie Kaufverhalten, usw.)
- Kaum **Regulierungen**, die Politik zu langsam: Unternehmen sammeln nahezu ohne Einschränkung enorme Datenmengen und verwerten diese
- Fehlende **Normen**, keine **Transparenz**: selten wird differenziert, was gesammelt werden sollte. Häufig blinde Sammelwut und Goldgräberstimmung.

Problematischer Datenschutz II

- Wer liest schon Datenschutzbestimmungen wirklich?
- **Informationelle Selbstbestimmung** ad absurdum: Akzeptieren Sie die Bestimmungen oder gehen Sie
- Auch **„anonymisierte Daten“** können z. T. im Nachhinein wieder entschlüsselt und Personen zugeordnet werden können [4]

Ethische Fragestellungen I





- Im Extremfall: Gefahr für die freie Meinungsäußerung / den **freien Willen**?
- Entwicklung hin zu **Überwachungsstaat** à la 1984? Die NSA-Enthüllungen von 2013 (Snowden) geben sehr zu Denken.
- Kampf gegen Terrorismus führt zur hochsensiblen Beschneidungen elementarer **Persönlichkeitsrechte**.
- Schon jetzt: *U. S. Department of Homeland Security* verwendet Big-Data-Analysen, um Passagiere am Flughafen bei Auffälligkeiten in deren Online-Historie gesondert abzuhandeln [14].

Ethische Fragestellungen II

- Ebenso problematisch: **Entsolidarisierung** bei Versicherungen.

Gesellschaftliche Verantwortung

- **Aber:** Die Gesellschaft ist nicht machtlos!
- Aktive Mitgestaltung und Einbringung in Entwicklungsprozess
- Wir selbst können entscheiden, wie mit unseren Daten umgegangen wird

Agenda

- Politisches Engagement zeigen
- Verantwortlichkeit einfordern
- Menschliche Handlungsfreiheit schützen
- Experten beratend ins Boot holen
- Datenmonopole vermeiden

5 Fazit & Ausblick

- Zukünftige Herausforderungen
- Ausblick
- Zusammenfassung

Zukünftige Herausforderungen für Big Data [2] I

Theoretische Hürden

- Ganzheitliche und rigorose **Definition** des Begriffs *Big Data*: formale Beschreibungen und theoretische Modelle.
- **Standardisierung**
  - Evaluierungs- und Benchmarkmethodiken für Datenqualität und Performance des Systems
  - *Theoretische* Validierung und Optimierung mit mathematischen Modellen
  - Abfragesprachen
- **Evolution** der Datenverarbeitung
  - Fokusverlagerung von rechenintensiven Ansätzen, hin zu datenorientierten Verfahren
  - Algorithm Engineering

Zukünftige Herausforderungen für Big Data [2] II

Praktische Hürden

- Performance der Echtzeit-Analyse
- Effiziente **Konvertierung der Datenformate** aufgrund der Heterogenität der Daten
- Unvermeidlicher **Datentransport** über das Netzwerk oft Flaschenhals
- **Datenverarbeitung**: Wege um Daten wiederzuverwenden, neu zu organisieren und Ausnutzung auch fehlerhafter Datensätze

Security & Datenschutz

- Grundproblem: Bisherige Schutzmechanismen auf Big Data kaum anwendbar, aber elementar wichtig.
- Neue, effiziente **Kryptoverfahren** müssen entwickelt werden

Ausblick: Wie geht es weiter? I

- Big Data wird viele **Lebensbereiche nachhaltig verändern**: ökonomische, gesellschaftliche und alltägliche Aspekte.
- Die Art, wie wir zukünftig **Denken und Entscheidungen** treffen, wird sich radikal wandeln [10].
  Aber: Big Data kann menschliches Denken nicht ersetzen!
- **Fokusverlagerung** in der IT: Technologie war früher treibende Kraft, nun sind die *Information* (die Daten) entscheidend.
- Hadoop ist erfolgreich, aber die theoretischen Grundlagen (2006) sind überholt. **Neue mächtigere Systeme** werden noch größere Datenvolumen mit noch höherer Diversität und Komplexität beherrschbar machen [2].

Ausblick: Wie geht es weiter? II

- Benutzerfreundliche **Visualisierung** von Ergebnissen
- Big Data fördert die **interdisziplinäre Fusion** vieler Felder in der Wissenschaft

Zusammenfassung

- In Big Data schlummert nachweislich **riesiges Potential**!
- **Großes Interesse** von Wirtschaft, Wissenschaft und Behörden
- Es existiert bereits eine Vielzahl entsprechender Technologien und Big Data wird seit Jahren **mit Erfolg praktisch eingesetzt**
- Dennoch steckt *Big Data* noch in den **Kinderschuhen** und Bedarf **weiterer Forschung** und Standardisierung, um künftige Hürden zu überwinden
- Der **Datenschutz** ist ein sensibles Thema und muss kritisch hinterfragt werden!

Diskussion

Big Data

Tatsächlich das „*Öl des 21. Jahrhunderts*“?

Quellenverzeichnis I

[1] Eric A. Brewer. „Towards Robust Distributed Systems“. In: *Proceedings of the Nineteenth Annual ACM Symposium on Principles of Distributed Computing*. PODC ’00. Portland, Oregon, USA: ACM, 2000, S. 7–. ISBN: 1-58113-183-6. DOI: 10.1145/343477.343502. URL: http://doi.acm.org/10.1145/343477.343502.

[2] Min Chen, Shiwen Mao und Yunhao Liu. „Big Data: A Survey“. In: *Mob. Netw. Appl.* 19.2 (Apr. 2014), S. 171–209. ISSN: 1383-469X. DOI: 10.1007/s11036-013-0489-0. URL: http://dx.doi.org/10.1007/s11036-013-0489-0.

[3] IBM Deutschland. *„IBM 5 in 5“: Innovationen, die unser Leben verändern werden*. Dez. 2013. URL: http://www-03.ibm.com/press/de/de/pressrelease/42779.wss (besucht am 17. 04. 2017).

Quellenverzeichnis II

[4] Andreas Dewes und Stephanie Rohde. *Es wird immer schwieriger, sich zu schützen.* Deutschlandfunk. Jan. 2017. URL: http://www.deutschlandfunk.de/datensicherheit-es-wird-immer-schwieriger-sich-zu-schuetzen.694.de.html?dram:article_id=377536 (besucht am 16. 04. 2017).

[5] Klaus-Peter Eckert und Radu Popescu-Zeletin. „Smart Data als Motor für Smart Cities“. In: *Informatik-Spektrum* 37.2 (2014), S. 120–126. ISSN: 1432-122X. DOI: 10.1007/s00287-014-0769-5. URL: http://dx.doi.org/10.1007/s00287-014-0769-5.

Quellenverzeichnis III

[6] Johann-Christoph Freytag. „Grundlagen und Visionen großer Forschungsfragen im Bereich Big Data“. In: *Informatik-Spektrum* 37.2 (2014), S. 97–104. ISSN: 1432-122X. DOI: 10.1007/s00287-014-0769-5. URL: http://dx.doi.org/10.1007/s00287-014-0769-5.

[7] John Gantz und David Reinsel. *Extracting Value from Chaos*. 2011. URL: http://www.emcgrandprix.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

[8] Tony Hey. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft External Research. 2016. URL: http://fiz1.fh-potsdam.de/volltext/fhpotsdam/10445.pdf (besucht am 15. 04. 2017).

Quellenverzeichnis IV

[9] Douglas Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety.* Techn. Ber. META Group, Feb. 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[10] Viktor Mayer-Schönberger und Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work and Think.* UK: John Murray Publishers, 2013.

[11] World Health Organization. *Influenza (Seasonal). Fact sheet.* 2016. URL: http://www.who.int/mediacentre/factsheets/fs211/en/ (besucht am 13. 04. 2017).

Quellenverzeichnis V

[12] World Health Organization. *Preterm Birth. Fact sheet.* 2016. URL: http://www.who.int/mediacentre/factsheets/fs363/en/ (besucht am 14. 04. 2017).

[13] Dominik Ryżko u. a. *Machine Intelligence and Big Data in Industry.* 19. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-30315-4.

[14] Chris Strohm. *Predicting Terrorism From Big Data Challenges U.S. Intelligence.* Okt. 2016. URL: https://www.bloomberg.com/news/articles/2016-10-13/predicting-terrorism-from-big-data-challenges-u-s-intelligence (besucht am 16. 04. 2017).