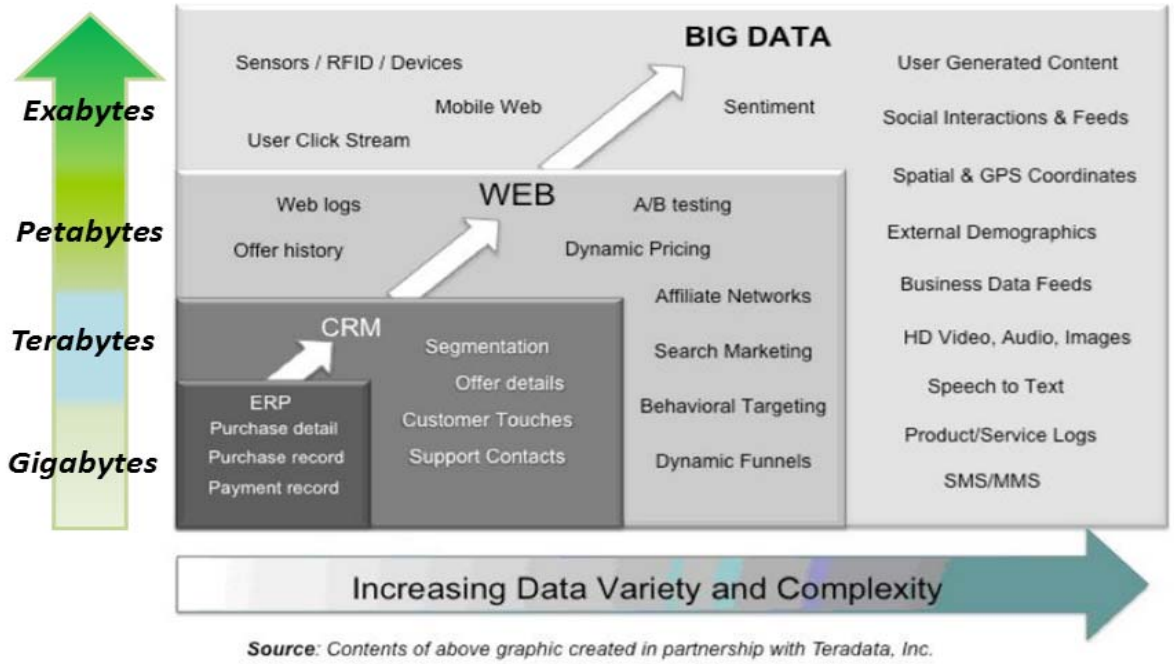
New Trends in Big Data

Prof. Dr. E. Rahm
und Mitarbeiter

WS 2013/14

UNIVERSITÄT LEIPZIG

Massives Wachstum an Daten



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Gartner:

- pro Tag werden 2.5 Exabytes an Daten generiert
- 90% aller Daten weltweit wurden in den 2 letzten Jahren erzeugt.

Datenproduzenten: Soziale Netze, Smartphones, Sensoren ...

Big Data Challenges

Volume	Skalierbarkeit von Terabytes nach Petabytes (1K TBs) bis Zettabytes (1 Milliarde TBs)
Variety	variierende Komplexität: strukturiert, teilstrukturiert, Text / Bild / Video
Velocity:	Near-Realtime, Streaming
Veracity:	Vertrauenswürdigkeit
Value	Erzielen des (wirtschaftl.) Nutzens durch Analysen

Potentiale für Big Data-Technologien

- Daten sind Produktionsfaktor ähnlich Betriebsmitteln und "Humankapital "
- Essentiell für viele Branchen und Wissenschaftsbereiche
- Valide Grundlage für zahlreiche Entscheidungsprozesse
  - Vorhersage/Bewertung/Kausalität von Ereignissen
- Kurzfristige Analysen von Realdaten im Geschäftsleben
- Beispiele
  - Nutzungsanalyse auf Web-Sites
  - Empfehlungsdienste (Live Recommendations)
  - Analyse/Optimierung von Werbe-Massnahmen

Neuartige Anwendungen für Big Data Analytics

Verkehr und Logistik

Hausautomatisierung

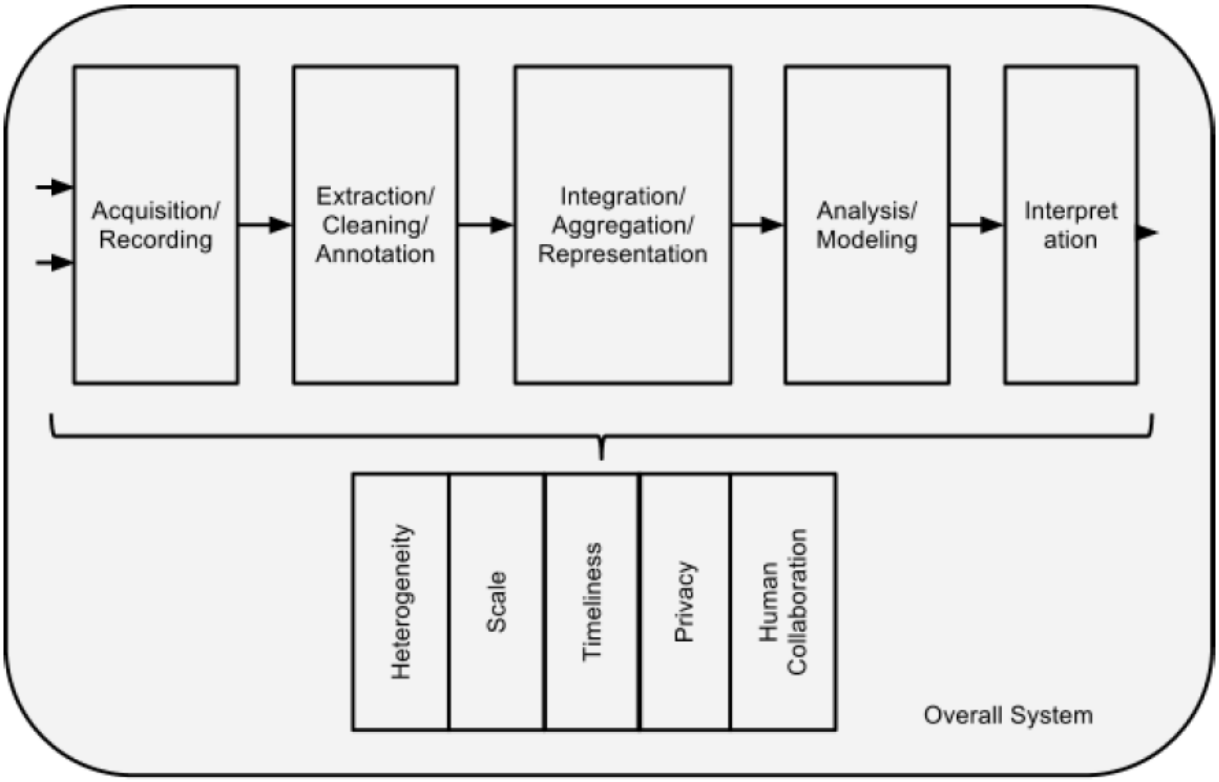
Lebenswissenschaften

Marktforschung

Digital Humanities

Ressourcenmanagement („smarter cities/planet“)

Big Data Analysis Pipeline



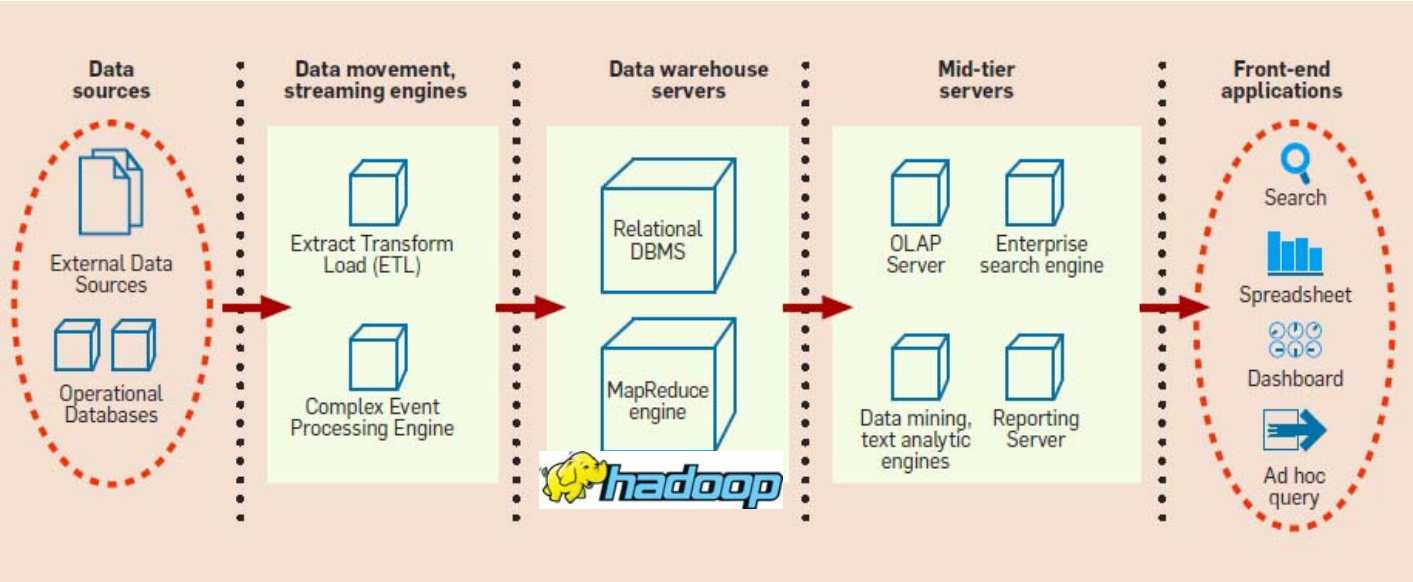
Source: *Agrawal et al: Big Data: Challenges and Opportunities*, 2011

Architekturalternativen

- Data Warehouse Appliances
  - Column Store, In-Memory-Optimierungen
  - Parallele DB-Vearbeitung mit vielen Knoten/Cores, Spezial-Hardware, z.B. FPGA (Netezza)
- Massiv skalierbare Cloud-Architekturen
  - Nutzung von NoSQL Data Stores
  - Frameworks zur automatischen Parallelisierung datenintensiver Aufgaben (MapReduce / Hadoop)

- Kombinationen: DWH + Cloud/Hadoop

Analyse-Pipeline



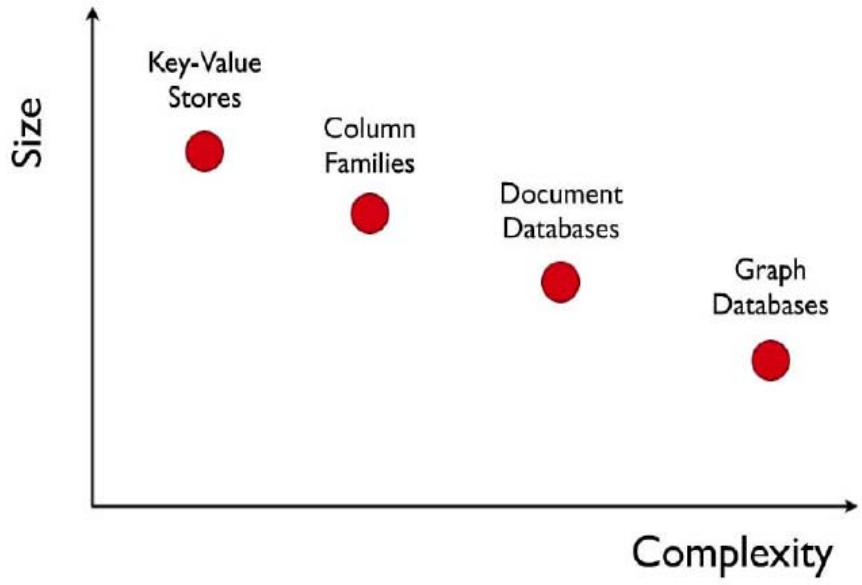
- Datenvorverarbeitung und Datenintegration
- Unterstützung von Stream-Daten und Cloud-Infrastrukturen (Hadoop)

Probleme relationaler Datenbanken

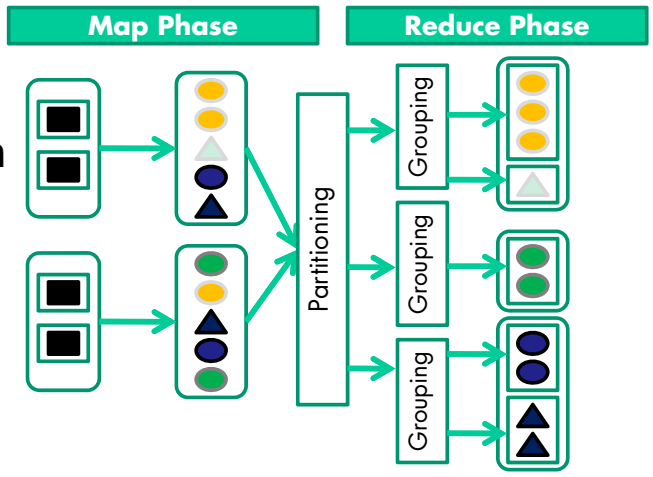
- Schema-getrieben ("Schema First")
  - weniger geeignet für semi-strukturierte (Web-) Daten
  - zu starr für irreguläre Daten
- relativ hohe Kosten, v.a. für Parallele DBS (kein Open-Source System)
- Skalierbarkeitsprobleme für Big Data (Web Scale)
  - Milliarden von Webseiten
  - Milliarden von Nutzern von Websites und sozialen Netzen
- ACID aufwändig / strenge Konsistenz nicht immer erforderlich

- Entwicklung seit ca. 2009
  - Ursprünglicher Fokus: moderne “web-scale” Datenbanken
- Merkmale
  - nicht-relational
  - open-source
  - verteilt, horizontal (auf große Datenmengen) skalierbar
  - schema-frei, Datenreplikation
  - einfache API
  - eventually consistent / BASE (statt ACID)
  - fehlende Standardisierung
- Zunehmende Koexistenz mit SQL
  - “NoSql" wird als “Not only Sql“ interpretiert

Grobeinordnung NoSQL-Systeme

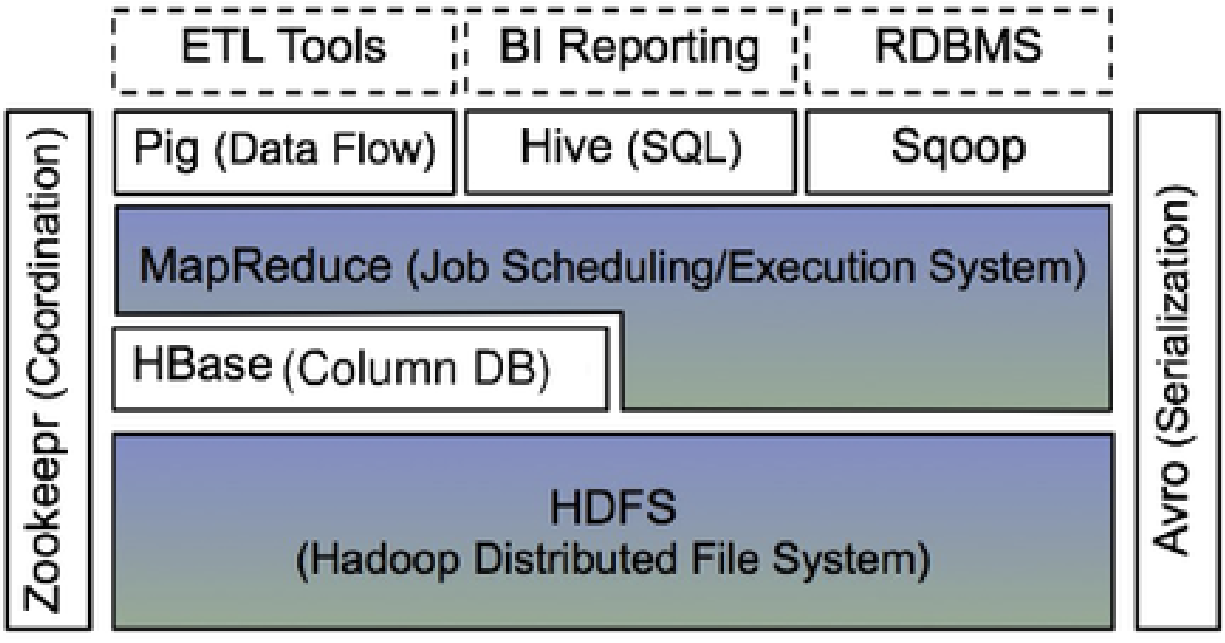


MapReduce



- Framework zur automatischen Parallelisierung von Auswertungen auf großen Datenmengen
  - Entwicklung bei Google
  - Apache Open-Source-Implementierung: Hadoop
- Nutzung v.a. zur Verarbeitung riesiger Mengen teilstrukturierter Daten in einem verteilten Dateisystem
  - *Konstruktion Suchmaschinenindex*
  - *Clusterung von News-Artikeln*
  - *Spam-Erkennung …*

Hadoop Ökosystem



Hadoop Ökosystem

- Zunehmende Unterstützung für SQL-Anbindung
  - Cloudera Impala
  - Apache Drill
  - Scoop: JDBC-Konnektor für Bulk-Datentransfer
- Unterstützung für Stream-Daten (Sensordaten, Twitter, Logs etc) : Flume
- Unterstützung für Graph-Daten: Giraph

Google: Trend zu massiv verteilten Datenbanken

- 2003/04 Google Filesystem (GFS), Map-Reduce
  - Basis für Apache HDFS, Hadoop
- 2006: Google BigTable
  - Basis für HBase (2008), Facebook-Nutzung (2010+)
- 2012: Neues verteiltes SQL/ACID-fähiges DBS **Spanner**
  - Ziel: Millionen Knoten mit über verschiedene Data-Center verteilten Daten
  - Basis für unternehmenskritische Anwendungen, v.a. Online-Werbung (Google **F1**)

SEMINAR

UNIVERSITÄT LEIPZIG

Seminarziele

- Beschäftigung mit einem praxis- und wissenschaftlich relevanten Thema
  - kann Grundlage für Abschlussarbeit oder SHK-Tätigkeit sein
- Erarbeitung + Durchführung eines Vortrags unter Verwendung wissenschaftlicher (englischer) Literatur
- Diskussion
- Schriftliche Ausarbeitung zum Thema
- Hilfe und Feedback durch zugeteilten Betreuer

Seminar: Anrechnungsmöglichkeiten

- Masterstudium
  - Teil der Module *Moderne Datenbanktechnologien*
  - *Seminarmodul (*oder *Masterseminar)*
- Bachelorstudium
  - *Seminarmodul (*oder *Bachelorseminar)*

UNIVERSITÄT LEIPZIG

Scheinvergabe / Modulprüfung

- selbständiger Vortrag mit Diskussion (ca. 45 Minuten)
  - Abnahme der Folien durch Betreuer
- schriftliche Ausarbeitung (ca. 15 Seiten)
  - Abnahme der Ausarbeitung durch Betreuer
  - Ausarbeitung soll zum Vortragstermin vorliegen (Vorträge ab Januar 2014)
- aktive Teilnahme an allen Vortragsterminen
- Modul-Workload: 30h Präsenzzeit,
  120 h Selbststudium

Seminar (3)

- Themenzuordnung
  - Koordinierungstreffen mit Betreuer bis **spätestens 4.11.2013**
  - ansonsten verfällt Seminaranmeldung
  - freiwilliger Rücktritt auch bis max. 4.11.2013

- Vortragstermine
  - 5x Montags, P702, ab **6. 1. 2014**
  - max. 2 Doppelstunden ab 9:15 Uhr (bis max. ca 12:30 Uhr)

UNIVERSITÄT LEIPZIG

Komplex	Betreuer	max. #Themen	Termin	Studenten
Large-scale Datenanalyse (Facebook Scuba, Google Dremel, Apache Drill / Spark)	Kolb	3	6.1.	Häberlein, Jacob
SQL & Big Data (Shark, Cloudera Impala, Clustrix)	Wartner/ Nentwig	3		
OLTP & Big Data (Google Spanner /Google F1, HStore/VoltDB, HyPer)	Wartner/ Nentwig	3	6.1.	Recknagel/ Schultz
Graphdatenbanken und –anfragesysteme (MS Trinity, MS Horton, Facebook TAO, Facebook Unicorn)	Petermann	3	13.1.	Prochotska, V. Rudolph Draeger
Arraydatenbanken (SciDB, TimeArr)	Arnold	2	20.1.	Kohlmeyer
Streaming-Daten & Complex Event Processing (Übersicht, StreamDB, StreamBase, Ester)	Arnold/ Sehili	2	20.1.	Tilch Pötschke
Next Generation Sequencing	Gross	2		
GPU-basierte Datenanalyse	Gross/ Sehili	2	27.1.	Konrad, Buchmeier