
# BIG DATA – HYPE ODER CHANCE



Fuchs Dominik 16.05.2014

# Fahrplan

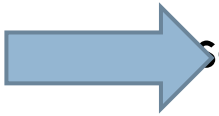


- Begriff Big Data
- Die „3“ V‘s
- Fallbeispiel Google
- „Was?“
- „Wie?“
- „Womit?“
- Fazit & Ausblick in die Zukunft

# Der Begriff Big Data



- Datenmengen, die zu groß sind um mit händischen oder klassischen Methoden bewältigt werden zu können
- Analyse & Überwachung von Nutzdaten (NSA-Affäre)
- Sammeln von Daten
- Erzeugung von nützlichen/gewinnbringenden Informationen



- Zusammenarbeit von Computer und dem schlichen Gehirn

# Relevanz des Begriffs



- Anstieg des Datenvolumens
  Stand 2011 wurde die Zettabyte Barriere geknackt(Eine 1 mit 21 Nullen)
- Daten als 4. Produktionsfaktor -> enorme wirtschaftliche Bedeutung
- Bsp.: Stockholm
  ->Verkehrsbewältigung

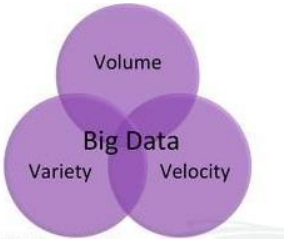
# Fallbeispiel Google



- Enormes Ausmaß an Suchanfragen täglich
- Google als riesige Datenplattform (Google Drive)
- Google als soziales Netzwerk (Google+)
- Auswertung und Analyse von Daten
- Kurze Latenzen pro Job

→ Google als Vorreiter im Umgang mit Big Data

# Die „3“ V‘s






- Velocity: konstant/nach oben unbeschränkt
  - Echtzeit
- Volume: Datenmenge verdoppelt sich alle 2 Jahre
  - Alle 2013 gespeicherten Daten auf iPads ergäbe eine 21.000 km lange Mauer
- Variety: 90% der Daten sind unstrukturiert
- Verbraucher: Pro Tag ca. 218.000 neue Internet-Nutzer

# „W“ Fragen für den Umgang mit Big Data

Für einen erfolgreichen Umgang mit Big Data ist es wichtig sich selbst die richtigen Fragen zu stellen -> Nur so kann ein strategisches Vorgehen erzielt werden

# „Was?“



- Schwierigste, aber zentralste Frage im Umgang mit Big Data
- Daten die nur rumliegen kosten Geld, deswegen ist es wichtig diese schnell und effizient zu verwerten
- Bsp.: Kundendaten
  - Alter, Umsatz, Einkaufszeit usw. liegen in einer Datenbank
  - Auswertungen um Umsatzprognosen oder Wachstumsraten graphisch darzustellen
  - Auswertungen um herauszufinden welche Produkte von Kunden welchen Alters gekauft werden
  - Ermittlung von persönlichen Vorlieben der Kunden
  - Zuschneiden der Werbung und Angebote speziell auf bestimmte Kundenkreise

# „Was?“ - Chancen



- Beispiele für Chancen im Umgang mit Big Data:
  - Optimierung und Personalisierung von Werbemaßnahmen und Steigerung von Cross Selling aufgrund von besserem Kunden- und Marktwissen
  - Besseres Risiko-Management in Zahlungs- und Handels-Strömen durch Entdeckung von Abweichungen und Unregelmäßigkeiten
  - Aufbau flexibler und intelligenter Abrechnungssysteme in der Versorgung (Strom, Wasser, Gas) und Telekommunikation
  - Erkennen von Abhängigkeiten und automatisierte Hypothesenbildung in Wissenschaft und Forschung

# „Wie?“



- Die zweite wichtige Frage, die sich mit den Mitteln, aber auch den limitierenden Faktoren befasst
- Problem: Motto „Never touch a running System“ wird oft zu streng verfolgt. Gerade hier gilt, wer wagt, der gewinnt!
- Mathematik
  - Kennen und Beherrschen von Algorithmen
  - Wahrscheinlichkeiten, Statistiken & Prognosen
    - Machine Learning
    - Natural Language Processing
- Polyglott
  - Welches ist die passende Programmiersprache?
    - Fixierung auf eine ist nicht effizient -> mehrschichtige Orientierung (Phyton, R, F# usw.)
  - Auch bei Datenbanken sollte man sich nicht auf eine versteifen

# „Wie?“ - Logs



- Logs – effizientes Mittel zur Analyse
  - Aktionen die User auf einem System durchführen werden protokolliert
    - Nicht die einzelne Tätigkeit, sondern der Fluss an Aktivitäten
  - Diese werden in relativ simplen, leicht lesbaren Logs abgespeichert, die interessante Informationen beinhalten
  - Beispiele:
    - Zeit die der Nutzer auf der Seite verbracht -> kurze Dauer = Kunde hat schnell das Interesse verloren
    - Produkte die der Nutzer angesehen hat -> Interessen herausfiltern (Empfehlungen)
    - Zugriffszeiten auswerten -> Peaks erkennen & vermeiden
    - Klick-Zähler -> simples Verfahren um beliebte Produkte zu ermitteln

# „Wie?“ – Nutzung sozialer Medien



- Nutzung sozialer Netzwerke
  - Sind extrem relevant um Vorlieben der Kunden zu analysieren
  - Haben heutzutage sehr hohe Verbreitung
- Beispiel Facebook:
  - Die „Likes“: Ermitteln der Interessen einer Person
  - Statusmeldungen: persönliche Beiträge (Orte und Personen)
  - Fotos: Kleidung, Hobbys
  - Grenzen -> Privatsphäre Einstellungen, mit Personen befreundet sein

moralische Grenze -> Privatsphäre der Leute wird missachtet

# „Wie?“ – Nutzung sozialer Medien



- Nutzung sozialer Netzwerke
- Beispiel: Xing
  - Kann benutzt werden um Ansprechpartner interessanter Firmen zu recherchieren
  - Voraussetzung hierfür ist ein Premium Account
  - Header der Personen mit Name, Titel, Firma, Ort ist auch für fremde einsehbar
  - Gute Suchunktion lässt die Suche nach relevanten Begriffen zu
    - Beispiel: Eingabe des Begriffes „Big Data“ in ein Freitextfeld lässt alle Nutzer erscheinen, die an irgendeiner Stelle im Profil den Terminus vermerkt haben
    - So können direkt Personen angesprochen werden, die Fachwissen über das gewünschte Thema haben

# „Womit?“

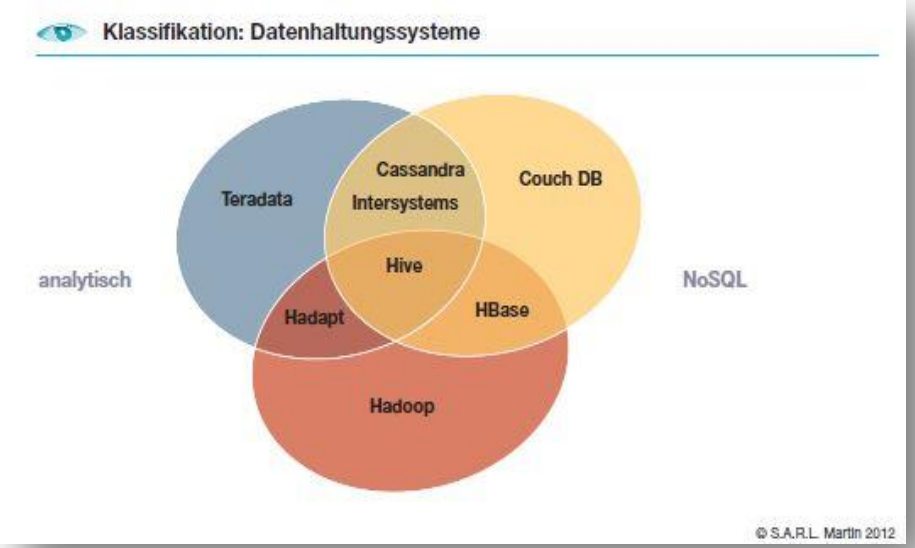


- Beim „Womit?“ handelt es sich um die leichteste, aber konkreteste Frage
- Es gibt Unmengen an Tools und Technologien (junger Markt -> viele Veränderungen)
- Verteilte Systeme
  - Gelten als der einzige Weg um flexibel und weit skalieren zu können -> zentrale Position im Umgang mit Big Data
  - Ist ein Zusammenschluss unabhängiger Computer -> präsentiert sich gegenüber dem Nutzer als ein System
  - Einzelne Einheiten kommunizieren über Nachrichtenaustausch
    - Asynchron: Vorteil ist, dass das Empfängersystem beim Funktionsaufruf nicht zwangsweise verfügbar sein muss
  - Verteilte Datenspeicher, verteilter Cache

# „Womit?“ - Datenhaltung



- Datenhaltung
  - Die Leistung von tra - ditionellen reicht nicht mehr aus um den steigen Daten- mengen gerecht zu werden
  - Hadoop: Framework für skalierbare verteilte Software
  - NoSQL: Datenbanken mit nicht relationalen Ansatz
  - Analytische Datenbanken:
    - Spalten-Orientierung
    - Daten-Komprimierung
    - In-Memory



# „Womit?“ - NoSQL



- Datenbanken mit nicht relationalen Ansatz
- Es gibt verschiedene Arten:
  - Key/Value Stores:
    - Hashtabelle -> Streualgorithmus zum Suchen bestimmter Objekte in großen Datenmengen
  - Document Stores:
    - Werden vorwiegend zur Speicherung von Dokumenten verwendet
    - Diese liegen redundant mit schwächere Struktur in den Stores vor
    - Man muss nicht auf Relationen achten, das nachträgliche Hinzufügen von Feldern ist unproblematisch
    - Greift beim Auslesen auf das MapReduce Verfahren zu
  - In-Memory Stores:
    - Eignen sich für das schnelle Laden und Speichern der Daten direkt aus dem Hauptspeicher
    - Verluste werden in Kauf genommen

# „Womit?“ – Map Reduce



- Programmiermodell für nebenläufige Berechnungen von Daten enormer Größe
- Baut auf Parallelität auf – um zu funktionieren werden mehrere Maschinen benötigt
- Baut auf verschiedenen Phasen auf:
  - Mapping Phase:
    - Eingehende Daten werden mit Basisberechungen durchsucht und lokal gespeichert
  - Reduce Phase:
    - Die Liste wird auf weniger Werte reduziert
  - Split Phase:
    - Hier werden die Daten auf die verschieden Cluster-Rechner aufgeteilt
  - Combine Phase:
    - Vor der Reduce Phase werden ähnliche Daten zusammengefügt

| | Traditional RDBMS | MapReduce |
|---|---|---|
| **Data Size** | Gigabytes | Petabytes |
| **Access** | Interactive and Batch | Batch |
| **Updates** | Read and write many times | Write once read many times |
| **Structure** | Static schema | Dynamic schema |
| **Integrity** | High | Low |
| **Scaling** | Nonlinear | Linear |



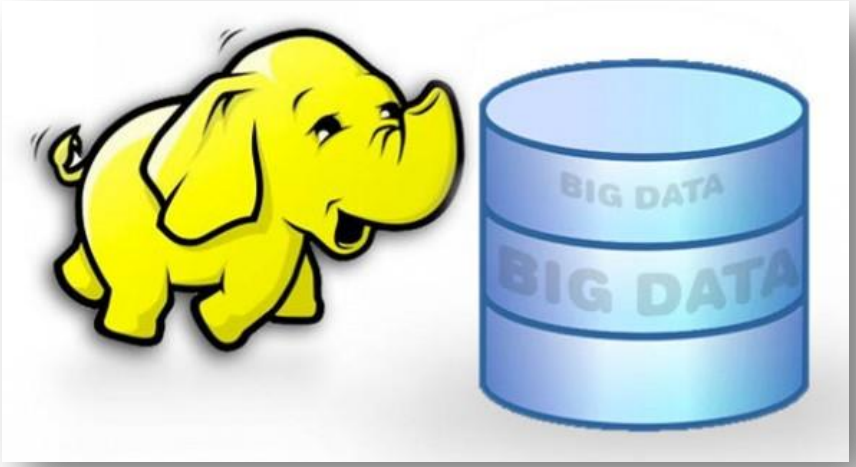
# Map Reduce vs. Traditionelle DBS

Hier werden die Unterschiede der verschiedenen Datenhaltungsansätze gegenübergestellt.

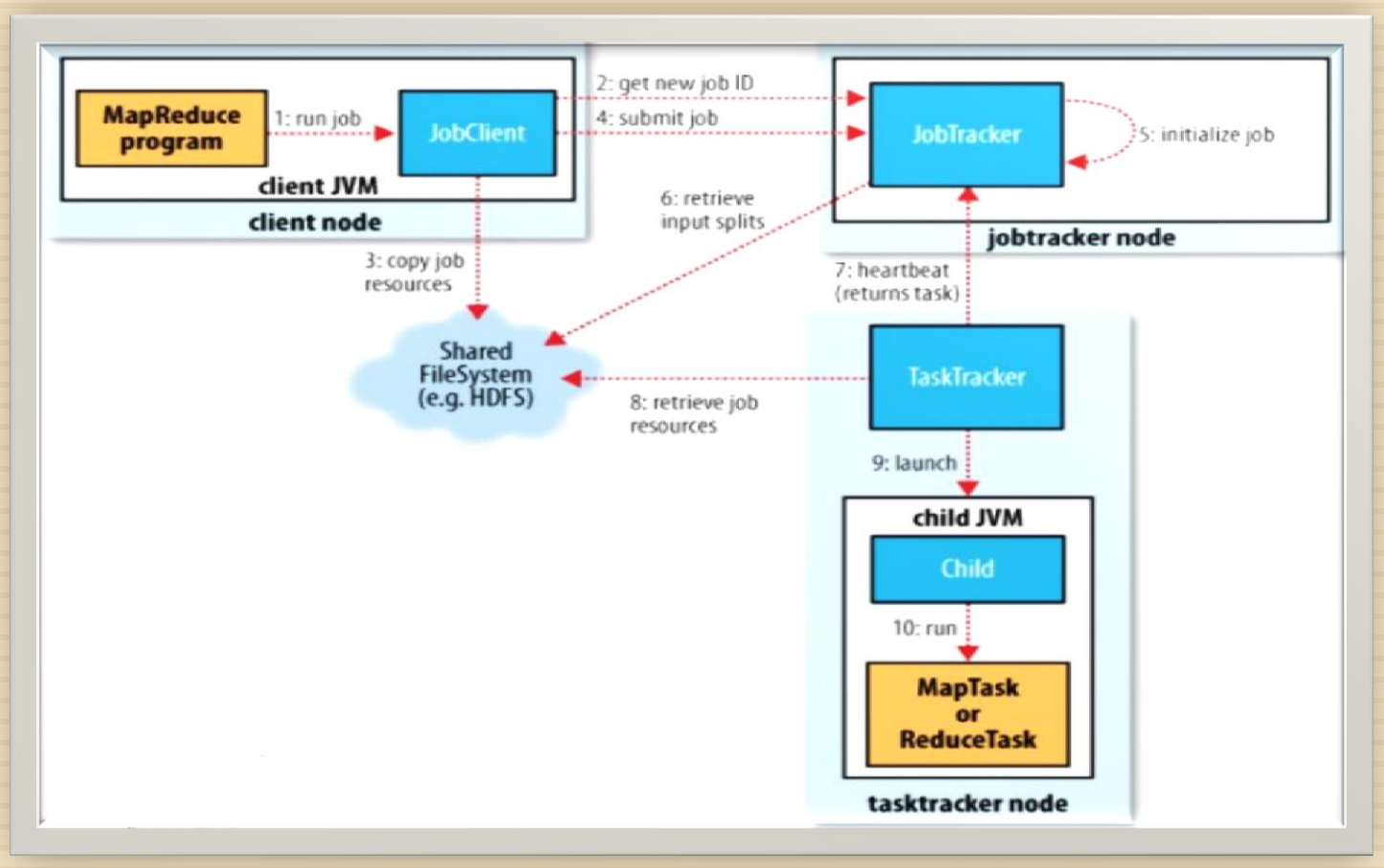
# „Womit?“ - Hadoop



- Implementierung die Map Reduce Konzept verwendet
- Open Source Projekt der Apache Software Foundation
- Hohe Fehlertoleranz (1-2 pro Job)
- Download nur unter Unix/Linux verfügbar
- Hadoop benötigt Java 1.6
- Für das Cluster wird ein gemeinsames Dateisystem benutzt –> HDFS= Hadoop Distributed File System



# Funktionsweise von Hadoop




JobClient, JobTracker und TaskTracker interagieren mit den Daten aus dem HDFS und verarbeiten diese. Hier finden sich auch die Phasen des MapReduce Konzeptes wieder.

# Marktentwicklungen – Trends



- Annäherung des SQL Marktes
  - NoSQL, New SQL und die RDBMS verschmelzen mehr und mehr ineinander
    - NewSQL verbindet die klassische SQL mit den Technologien, die für riesigen Datenmengen nötig sind
  - Es gibt keinen ultimativen Weg, der Markt ist relativ jung deswegen herrscht hier noch sehr viel Bewegung
- Veralten der Technologien
  - Bsp.: MapReduce
    - MapReduce gilt heute als bereits veraltet
    - Nicht mehr performant, neuere Technologien wie Spark sind effizienter
    - Kürzere Latenzen pro Job
    - In-Memory Support
    - Streaming Support

# Marktentwicklung - Umsatz



- Von 2012 bis 2017 soll sich der Umsatz von Big Data verzehnfachen (5,4$ Mrd. auf 53,4$ Mrd)
- Allerdings ist der Anteil des Big Data Umsatzes am gesamten Gewinn sehr klein
  - Bei den Top Unternehmen mit mehr als 100$ Mrd. macht er nur 0-1 % aus
  - Ausnahme ist lediglich TerraData mit 10%
- Man geht im Markt von großen Übernahmewellen der großen Firmen aus

# Persönliches Fazit



- Hype?
  - Ja und Nein
    - Ja, im Bezug auf den Begriff. Dieser wird wahllos für alles was im entferntesten mit Big Data zu tun hat verwendet
    - Nein, im Bezug auf die sich dahinter verbergenden Technologien. Hier gibt es für Unternehmen große Chancen und Potenziale (Kundenbindung)
- Die 3 Fragen „Was?“, „Wie?“ und „Womit?“ bieten eine guten Ansatz sich der Thematik zu nähern
  - Dabei ist das „Was?“ am wichtigsten, da diese Frage am wenigsten konkret ist und die beiden anderen Fragen auf sie aufbauen
- Sehr spannendes Thema, was sich in den nächsten Jahren weiterentwickeln wird

# Noch Fragen?!

Vielen Dank für eure Aufmerksamkeit!