MATERNA
Information & Communications

NOSQL im BigData-Land: Von Analog-Istan nach Dig-Italien

Frank Pientka, Bonn, 14.06.2012

MATERNA

Vorstellung des Referenten: Frank Pientka

- Dipl.-Informatiker (TH Karlsruhe)
- Senior Architekt in Dortmund
- heise.de/developer/Federlesen-Kolumne
- Über 20 Jahre Erfahrung mit Datenbanken
- Veröffentlichungen und Vorträge dazu

MATERNA

MATERNA-Unternehmensgruppe

Dr. Winfried Materna Helmut an de Meulen

155 Mio. € Umsatz (2011, vorläufig)

1.300 Mitarbeiter

Gegründet 1980

MATERNA

Die Reiseroute

- Analog-Istan: Ein Blick zurück. Wie wichtig ist Physik?
- Die Herausforderungen: Schwimmen lernen im Datenstrom
- Dig-Italien: Besuch in einem unbekannten Land?
  - Leute: Neue Daten(-Banken) benötigt?
  - Sprache: Was ist NOSQL, MR?
  - Land: Welche Einsatzgebiete gibt es?
- Bewertung
- Ein Blick voraus

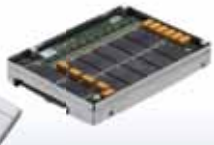
MATERNA

Analog-Istan: Warum IBM keine Platten mehr herstellt

4. September 1956, die erste Festplatte IBM 350 (5 MB)

2012

Hitachi Ultrastar SSD400S.B 400GB



2009

Western Digital WD Caviar Green 2TB

MATERNA

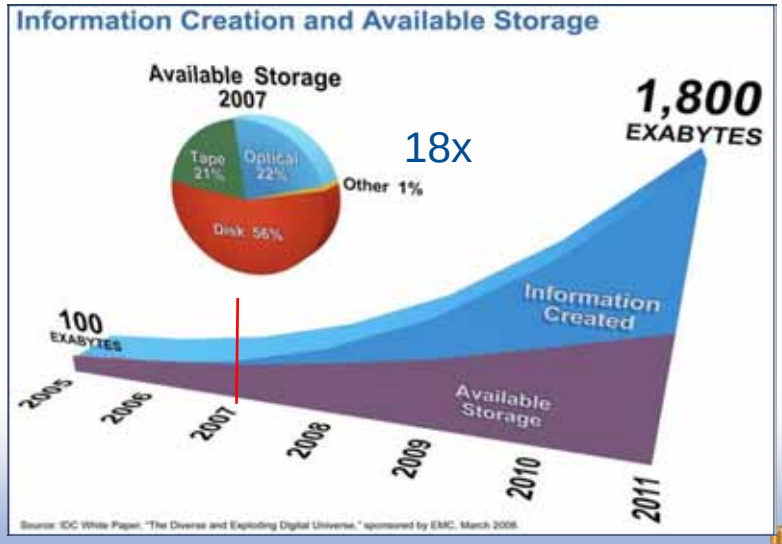
Datenexplosion

- 1,3 RFID tags in 2005 30 Billion today
- 2 Billion internet users by 2011
- 4,6 Billion mobile phones world wide
- Capital market data volumes grew 1750% (2003-06)
- Twitter processes 7 Terabytes of data per day
- World Data Centre for Climate 220 terabytes of Webdata, 9 Petabytes of additional data
- Facebook processes 10 Terabytes of data per day

MATERNA

Datenexplosion - das digitale Universum

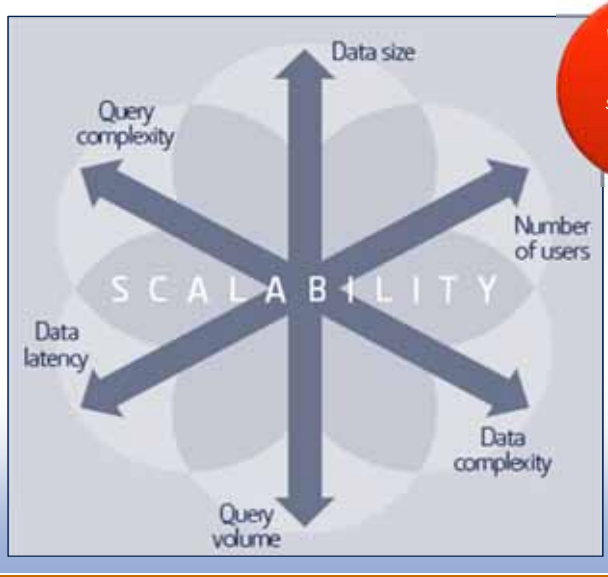
Information
Overkill
Overload



Die erzeugte Datenmenge (gespeichert, gesendet) übersteigt die aktuellen Speicherkapazitäten

MATERNA

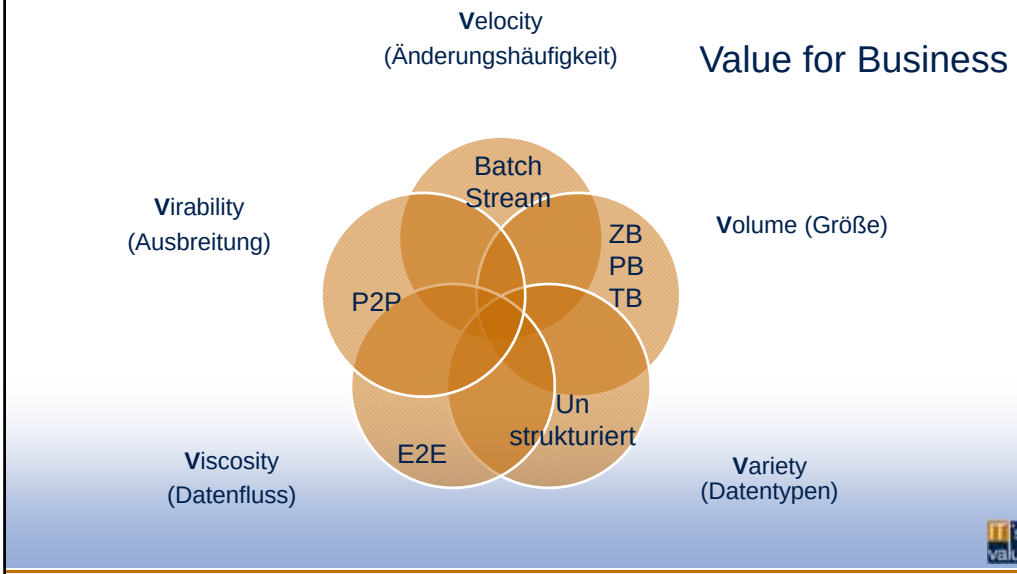
Datenwachstum ist ein mehrdimensionales Problem



Wachsende Daten wirken sich auf alle Bereiche aus!!!

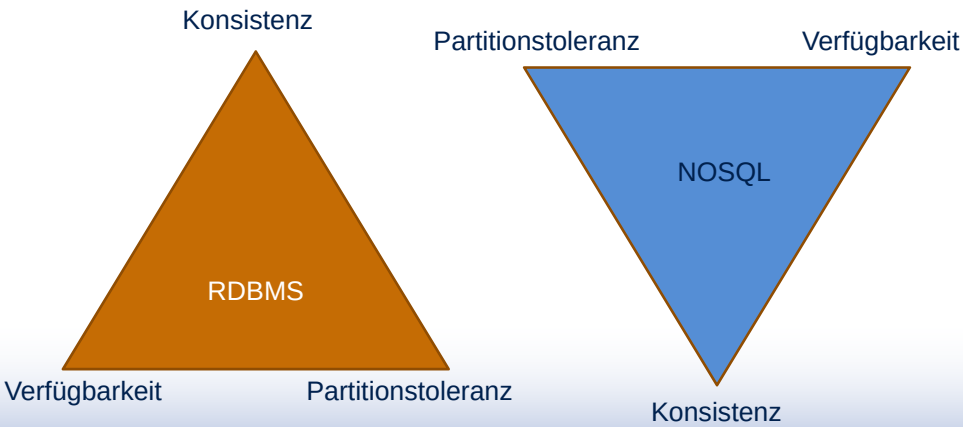
MATERNA

Fünf Vs bei Big Data



MATERNA

RDBMS vs NOSQL Schwerpunkte: Alles anders?



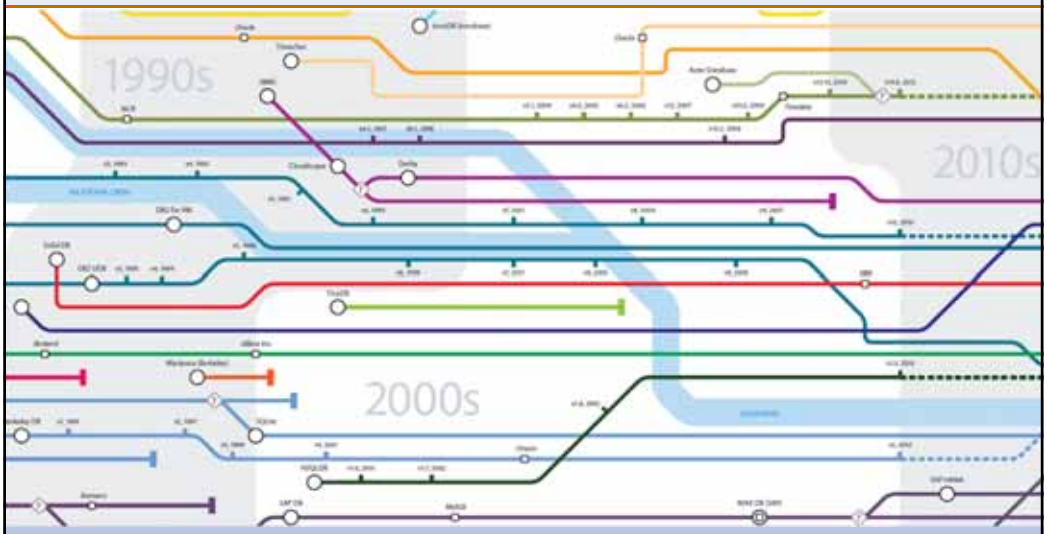
MATERNA

Die Reiseroute

- Analog-Istan: Ein Blick zurück. Wie wichtig ist Physik?
- Die Herausforderungen: Schwimmen lernen im Datenstrom
- Dig-Italien: Besuch in einem unbekannten Land?
  - Leute: Neue Daten(-Banken) benötigt?
  - Sprache: Was ist NOSQL, MR?
  - Land: Welche Einsatzgebiete gibt es?
- Bewertung
- Ein Blick voraus

MATERNA

An relationalen Flüssen und Codd-Seen….



www.hpi.uni-potsdam.de/naumann/projekte/rdbms_genealogy

MATERNA

Eine Ära geht zu Ende…

“The relational model of the 70s is not necessarily the answer”
(*Michael* Stonebraker, 2007)

MATERNA

Von Analog-Istan ins wilde Dig-Italien…

MATERNA

Verbindung von un-/strukturierten Inhalten

MATERNA

Die Reiseroute

- Analog-Istan: Ein Blick zurück. Wie wichtig ist Physik?
- Die Herausforderungen: Schwimmen lernen im Datenstrom
- Dig-Italien: Besuch in einem unbekannten Land?
  - Leute: Neue Daten(-Banken) benötigt?
  - Sprache: Was ist NOSQL, MR?
  - Land: Welche Einsatzgebiete gibt es?
- Bewertung
- Ein Blick voraus

MATERNA

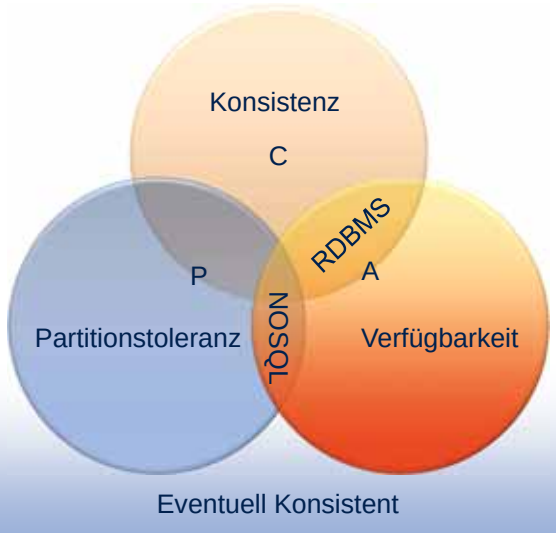
Kennzeichen von NOSQL-Systemen

N★SQL

- 1998 Carlo Strozzi No SQL, 2009 Johan Oskarsso Not only SQL
- Andere Datenstrukturen: einfach, geschachtelt, hierarchisch, vernetzt
- Andere Verteilung: CAP-Theorem
- Andere Verarbeitung: MR
- Andere Protokolle, Austauschformate: REST, JSON
- Partielle Konsistenz, statt verteilte Transaktionen: BASE vs ACID
- Mehr Flexibilität, weniger Kontrolle, Einschränkungen: Schemafrei, kein RI
- Weniger Verwaltungs-, mehr Programmieraufwand: Keine Sperren, Standard-QL
- Horizontale, lineare Skalierbarkeit: Standardhardware

MATERNA

CAP-Theorem: nur 2-aus3 (Brewer 2000)



MATERNA

ACID versus BASE Transaktionen

ACID:
Atomic
Consistent
Isolated Transactions
cannot interfere with each other.
Durable

BASE:
Basic **A**vailability
Soft-state
Eventual consistency

Verfügbarkeit wichtiger als Konsistenz für schwächere Transaktionsanforderungen

"Your Coffee Shop doesn't use Two-Phase-Commit"
(Gregor Hohpe, 2005)

MATERNA

NOSQL-Systeme

N★SQL

- Kommen nicht aus den Forschungslabors
- Aus der Praxis des Webs für das Web
- Amazon, Google, Facebook als Ideengeber
- OpenSource verfügbar und erweiterbar (Ökosystem, Partnerschaften)
- Developed-by-Community vs Designed-by-Company/ Committee

MATERNA

NOSQL-Produkte

Key Value Stores	Document Databases	BigTable Clones	Graph Database	Data Grid Caching
VoltDB	CouchOne	HBASE	neotechnology	redhat
DataStax	CouchDB	Cassandra		vmware
redis	10gen			
	mongoDB			
	Basho			

MATERNA

NotOnlySQL-Datenbanken

N★SQL

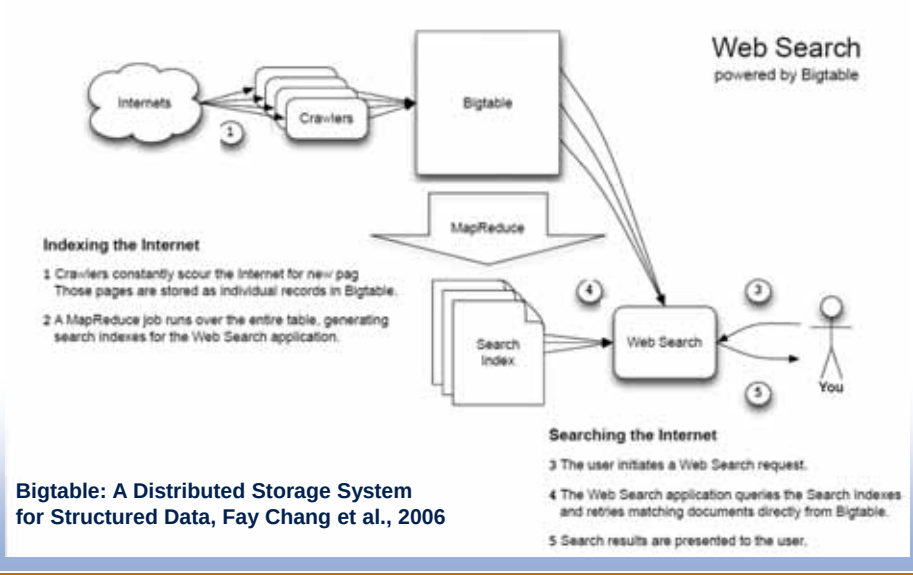
Gab es das nicht schon mal?

- 60er hierarchisch: IMS, LDAP, XML (Tamino, Xindice)
- 70er netzwerkartig: CODASYL
- 80er Hash-DB: xDBM Ken Thompson, **BerkleyDB**
- 80er dokumentenorientiert: Lotus Notes, Jackrabbit
- 90er objektorientiert: db4o, Versant
- 90er objekt-relational: ORACLE, DB2, mySQL

➔Hybride, temporale Datenbanken SQL:2011
Was ist die meist-installierte SQL-Datenbank?

MATERNA

Der Einsatz von BigTable bei Google



Bigtable: A Distributed Storage System for Structured Data, Fay Chang et al., 2006

MATERNA

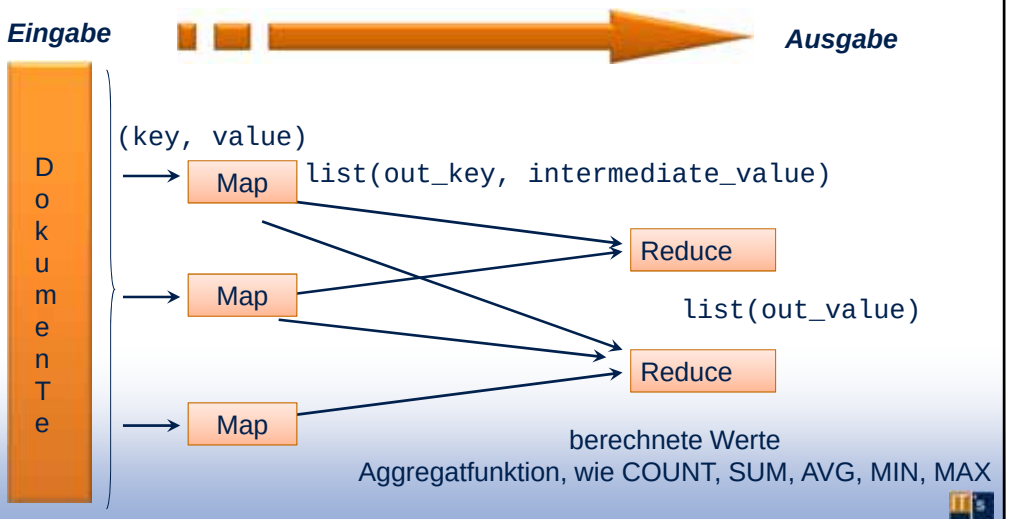
Sorting the World with MR

Quelle: (WIRED MAGAZINE 16.07)

MATERNA

MapReduce-Verfahren

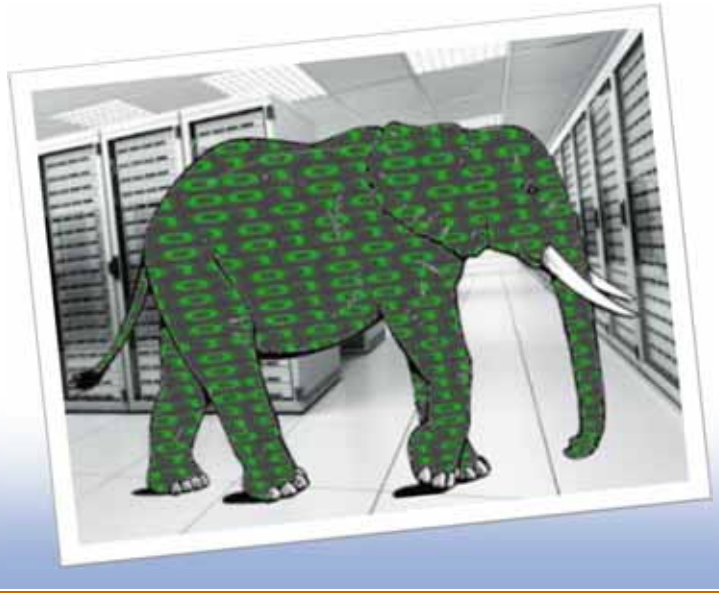
2 Phasen parallel ausgeführt, wie Pipes&Filter (UNIX)



berechnete Werte
Aggregatfunktion, wie COUNT, SUM, AVG, MIN, MAX

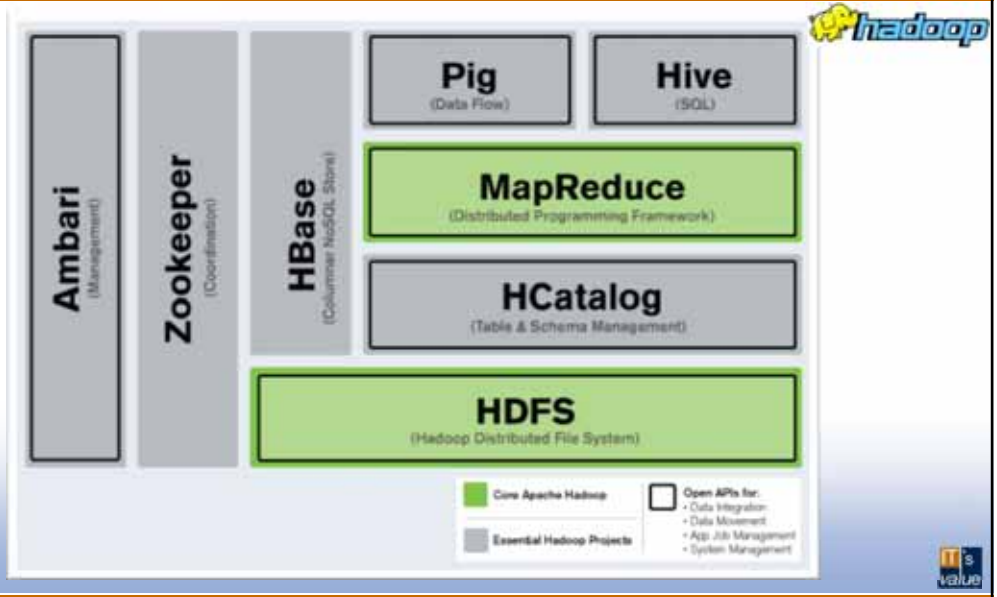
MATERNA

Der Elefant hat den Laden verlassen …



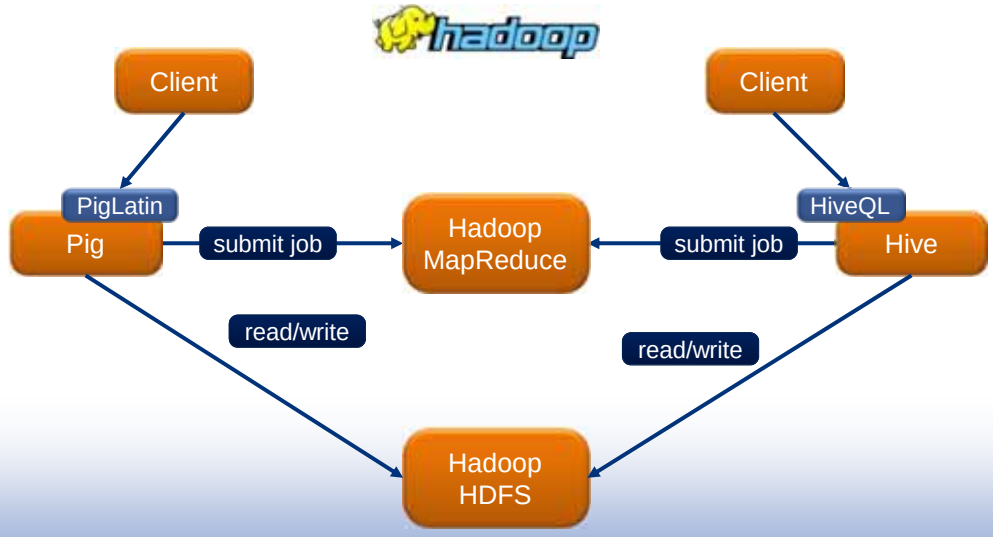
MATERNA

Das Hadoop-Ökosystem (De-Facto-Standard)



MATERNA

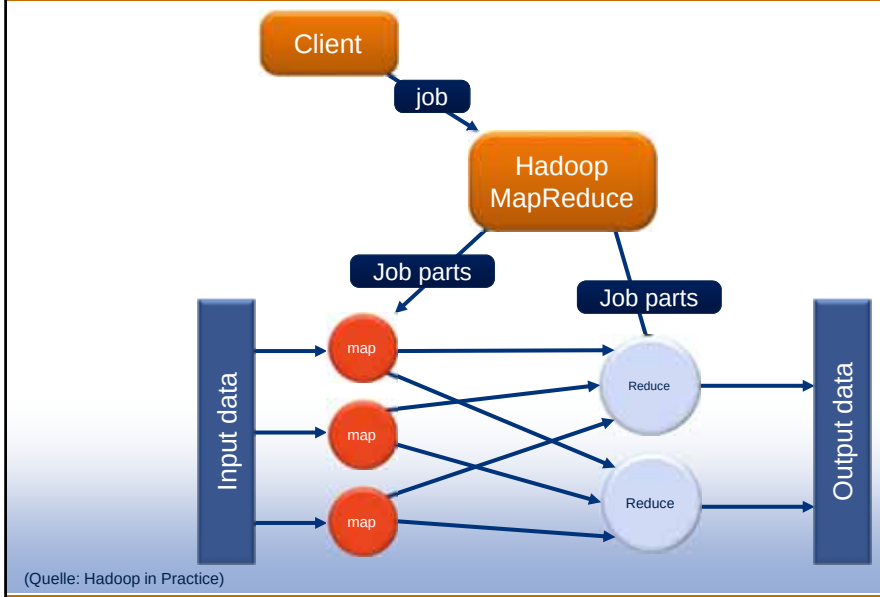
Wie hängen Teile zusammen?: MR, HDFS, Pig, Hive



(Quelle: Hadoop in Practice)

MATERNA

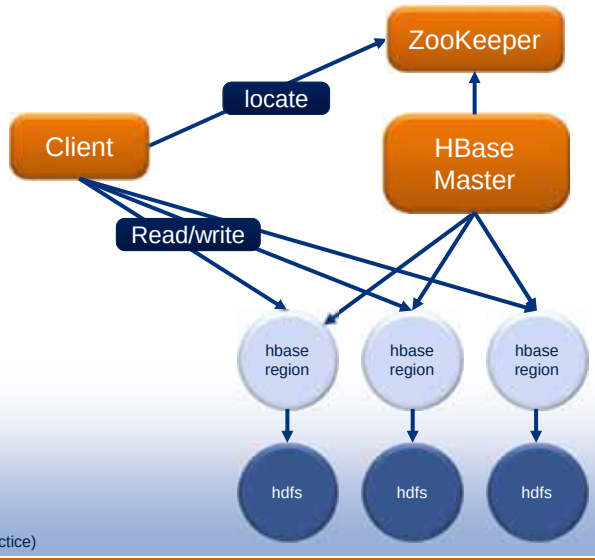
MapReduce bei Hadoop



(Quelle: Hadoop in Practice)

MATERNA

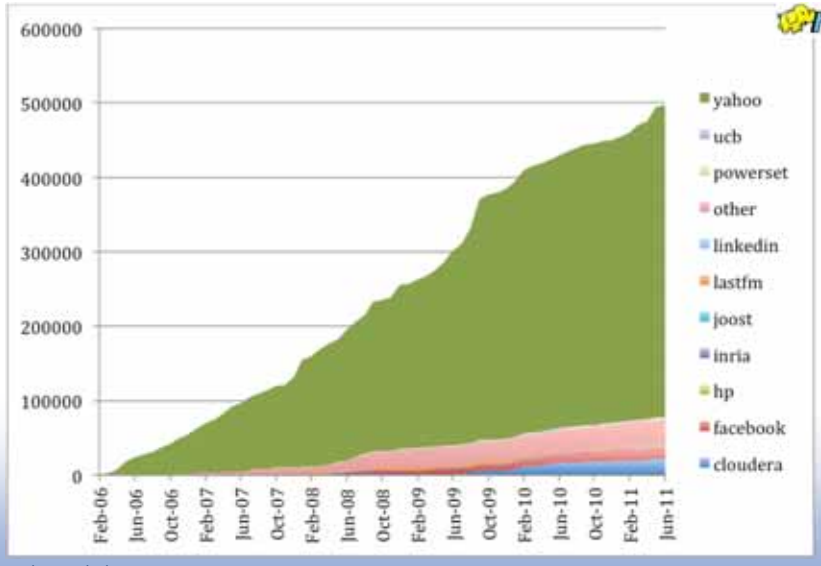
Wie hängen Teile zusammen?: Zookeeper, HBase (SpaltenDB), HDFS



(Quelle: Hadoop in Practice)

MATERNA

Wer entwickelt Hadoop?



(Quelle: Hadoop in Practice)

MATERNA

Der BigData-Hadoop-Zoo

Sqoop

Hortonworks

Greenplum

Pig

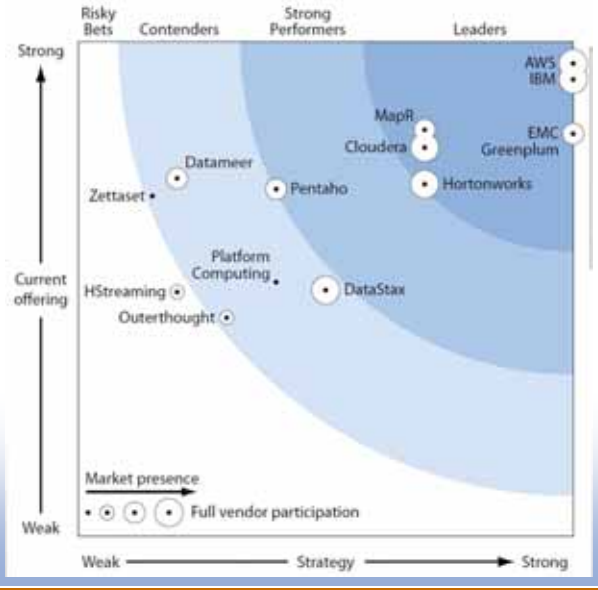
HIVE

HBASE

cloudera

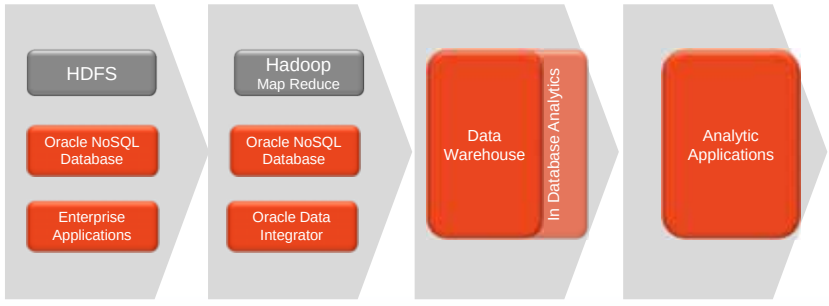
MATERNA

The Forrester Wave: Enterprise Hadoop Solutions, Q1 2012



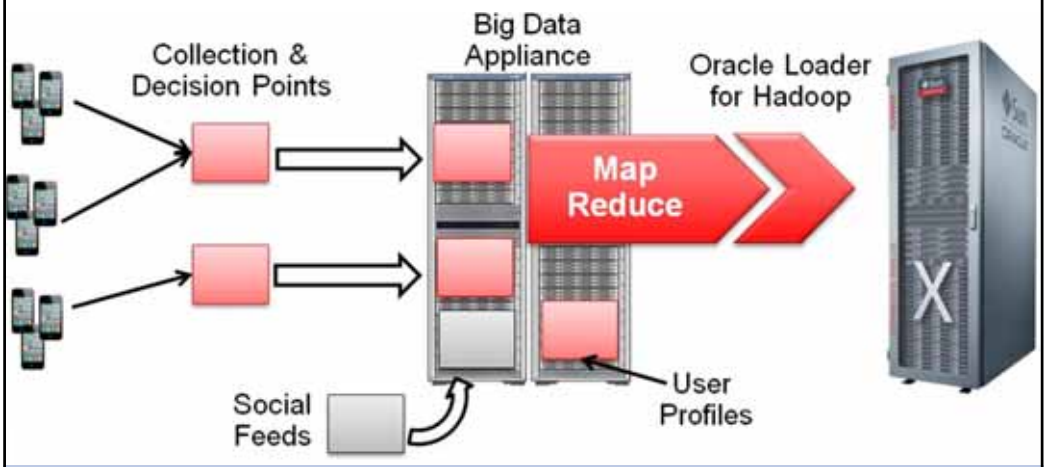
MATERNA

Oracle Big Data Appliance: Connectoren, Lader, NOSQL



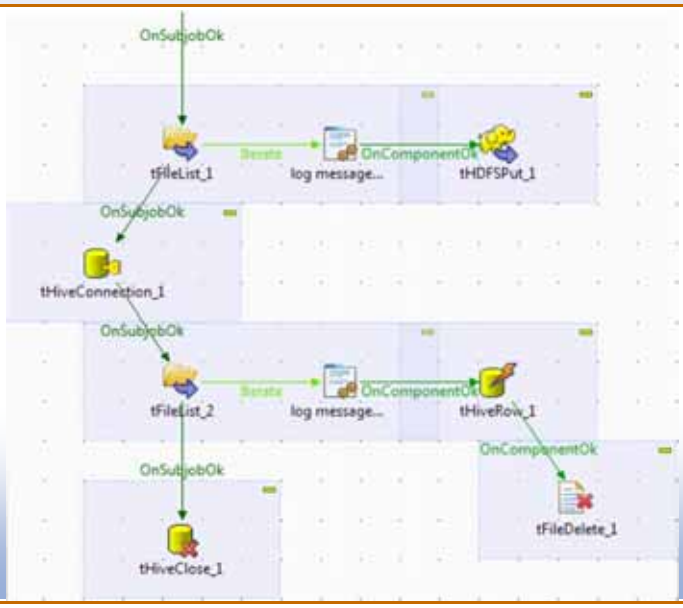
MATERNA

Einsatzmöglichkeit Oracle Big Data Appliance



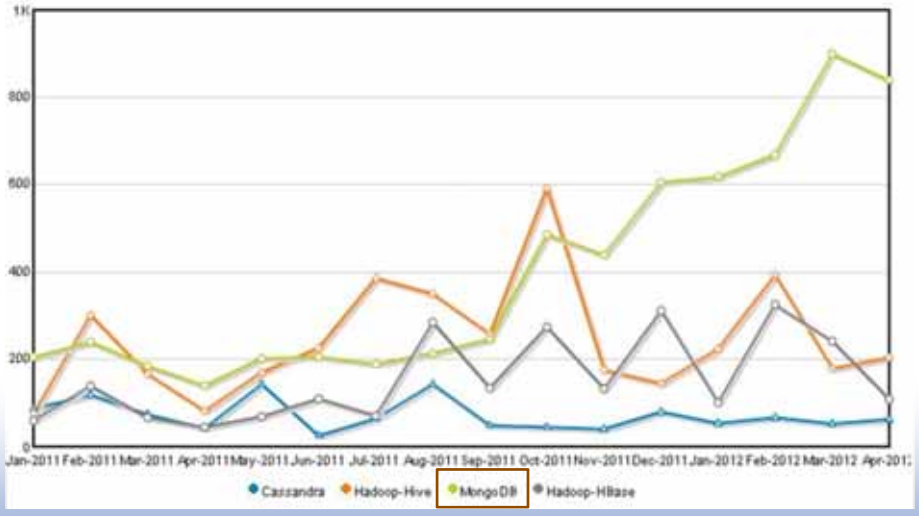
MATERNA

Datenladen in HDFS und Hive mit Talend Open Studio



MATERNA

Monthly Downloads of Top Four Big Data Connectors (Jaspersoft)



Jaspersoft Big Data Index (JBDI)

MATERNA

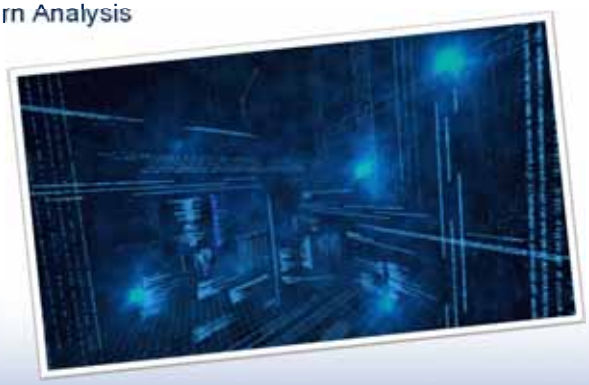
Die Reiseroute

- Analog-Istan: Ein Blick zurück. Wie wichtig ist Physik?
- Die Herausforderungen: Schwimmen lernen im Datenstrom
- Dig-Italien: Besuch in einem unbekannten Land?
  - Leute: Neue Daten(-Banken) benötigt?
  - Sprache: Was ist NOSQL, MR?
  - Land: Welche Einsatzgebiete gibt es?
- Bewertung
- Ein Blick voraus

MATERNA

Big Data Use Cases

- Recommendation Engine
- Marketing Campaign Analysis
- Customer Retention and Churn Analysis
- Social Graph Analysis
- Capital Markets Analysis
- Predictive Analytics
- Risk Management
- Rogue Trading
- Fraud Detection
- Retail Banking
- Network Monitoring
- Research And Development
- Archiving



MATERNA

Wirtschaftlicher Nutzen von BigData-Technologien

McKinsey&Company

US healthcare	Europe public sector	Gobal personal location data	US Retail	Manufacturing
$ 300 billion value per year ~ 0,7% annual procuctivity growth	$ 250 billion value per year ~ 0,5% annual procuctivity growth	$ 100 billion + revenue für service providers Up to $700 billion value to end users	60+ % increase in net margin possible 0,5% - 1% annual procuctivity growth	Up to 50% decrease in product development, assembly costs Up to 7% decrease in in working capital

MATERNA

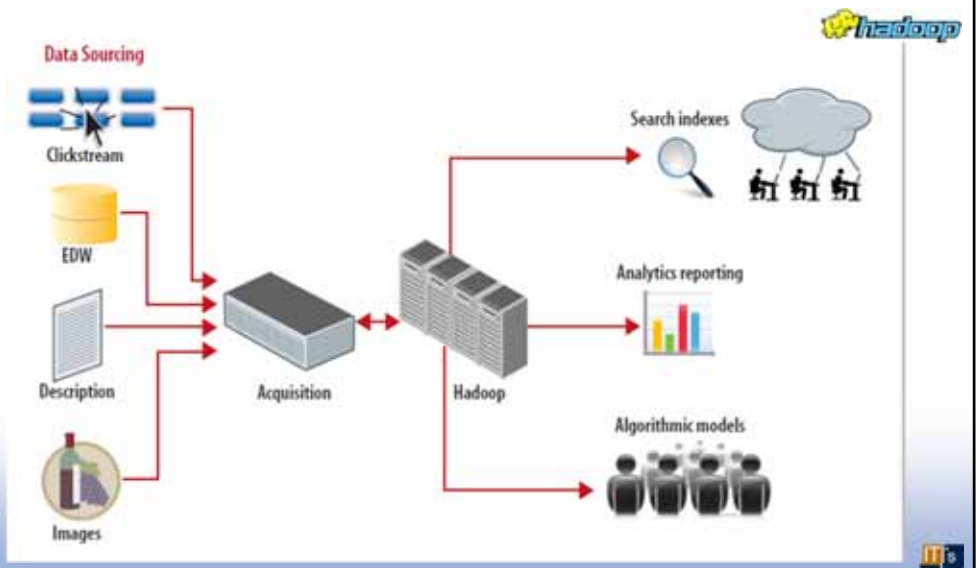
Wie mit Big Data arbeiten? (Latenzzeiten)

Native

SQL

MATERNA

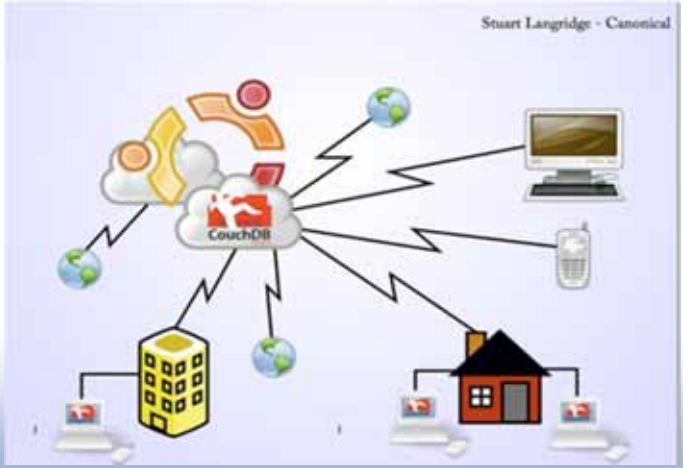
Clickstream-Analyse mit Hadoop bei eBay



MATERNA

CouchDB Ubuntu One: Cloud, PC, Smartphone

CouchDB auf über millionen PCs und in der Cloud mit Ubuntu 9.10 10.2009
EOL mit Ubuntu One 12.04 11.2011



MATERNA

Die Reiseroute

- Analog-Istan: Ein Blick zurück. Wie wichtig ist Physik?
- Die Herausforderungen: Schwimmen lernen im Datenstrom
- Dig-Italien: Besuch in einem unbekannten Land?
  - Leute: Neue Daten(-Banken) benötigt?
  - Sprache: Was ist NOSQL, MR?
  - Land: Welche Einsatzgebiete gibt es?
- Bewertung
- Ein Blick voraus

MATERNA

Zwischen Analog-Istan und Dig-Italien

MATERNA

Bewertung und Ausblick

- BigData: große Herausforderung, große Chance
- Die Hadoop-Plattform ist Defacto-Standard für Auswertung großer Datenmengen
- Standards (UnQL, Jaql, JSON Schema, HQL, BSON) fehlen
- MR, JSON-Datentyp werden in Datenbanken integriert
- Connectoren und Standardschnittstellen überbrücken Welten
- Professionalisierung der NOSQL-Produkte
- Marktkonsolidierung und Innovation birgt Risiken
- Fachkräftemangel und erste BestPracties-Erfahrungen
- "The Hadoop and MapReduce market will likely develop along the lines established by the development of the Linux ecosystem“ (IDC 2012, Dan Vesset)

MATERNA

Fazit

- BigData geht uns alle an (technisch, rechtlich)
- Neben der Verarbeitung wird Auswertung unstrukturierte Daten beschleunigt
- Starke Verwandtschaft mit Suche, BI, ETL, DW, DataMining, Batch, Streaming
- OpenSource, Connectoren und Appliance senken Einstiegshürden
- Für langfristigen Nutzen müssen Konzepte verstanden und für eigene Anforderungen angepasst werden

MATERNA

Big Data: Technologies and Techniques for Large-Scale Data

MATERNA

Datenwachstum ist nicht nur ein Datenbankproblem

MATERNA

Die Big Five

Schneller,Stärker,
Größer….

…haben keine
natürlichen
Feinde!

MATERNA

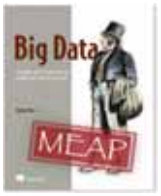
Ist BigData reif für die Praxis?

MATERNA

Es hängt davon ab, was man damit macht?

Weitere Informationen

http://nosql-database.org
http://www.emc.com/leadership/programs/digital-universe.htm



Vielen Dank für Ihre/Eure Aufmerksamkeit! Besuchen Sie unseren Stand

MATERNA GmbH
Dipl. Inform. Frank Pientka
Senior Software Architect
Business Division Applications

Telefon: +49 231 5599-8854
Telefax: +49 231 5599-272
Frank.Pientka@materna.de
http://xing.to/frank_pientka