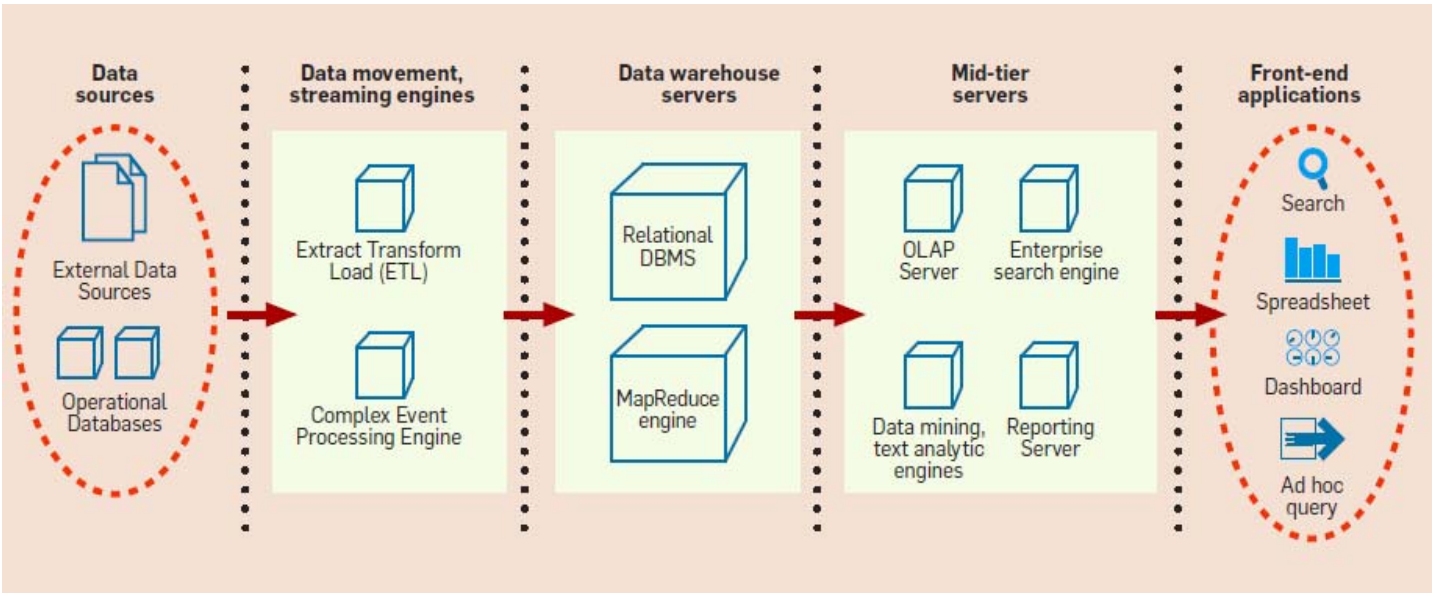
Large-Scale Data Analytics

Prof. Dr. E. Rahm
und Mitarbeiter

Seminar
WS 2012/13

Data Warehousing / OLAP



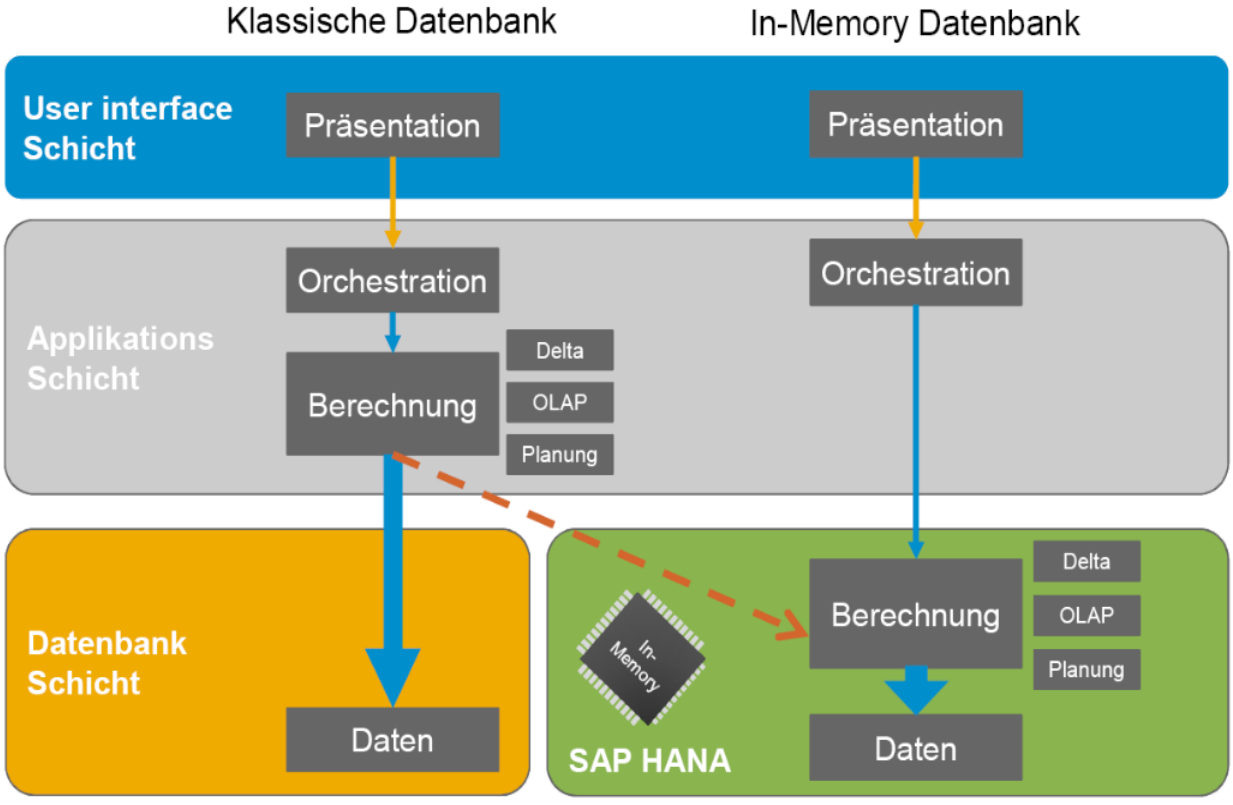
S. Chaudhuri et al, CACM, Aug. 2011

- Parallel DBS vs. Open Platforms for Big Data, e.g. HaDoop
- Near-Realtime OLAP

Data Warehouse Appliances

- Vorkonfigurierte, komplette Data Warehouse-Installation
  - Mehrere Server, Externspeicher, Software, etc.
  - Pricing nach Datenumfang (nicht Hardware)
- Pioniere mit Spezial-Hardware: Teradata, (IBM) Netezza
  - Weitere Systeme: SAP HANA, EMC Greenplum, MS SQL Server PDW (Parallel Data Warehouse), Oracle Big Data Appliance, HP Vertica, EXASOL
- Aktuelle Optimierungsansätze
  - Nutzung riesiger Hauptspeicher-Datenbanken
  - Nutzung von Flash-Speichern
  - Column-Store-Techniken
  - Parallelverarbeitung basierend auf PDBS-Cluster oder Map/Reduce
- Vorteile
  - Hohe Leistung / Skalierbarkeit
  - Hohe Verfügbarkeit (durch eingebaute Fehlerbehandlungsmechanismen)
  - Geringer Administrationsaufwand
  - Schnelle DWH-Realisierung/Nutzung

Beispiel: SAP HANA



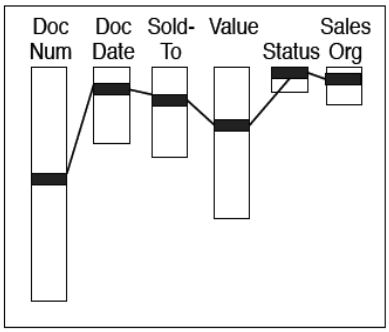
Quelle: SAP AG

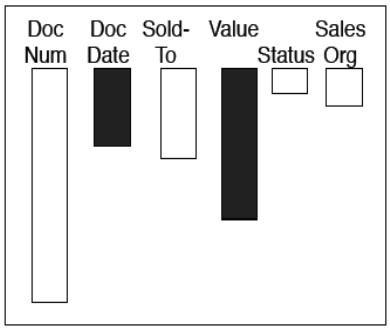
Anfragen – Column vs. Row Store *

SELECT *
FROM Sales Orders
WHERE Document Number = '95779216'

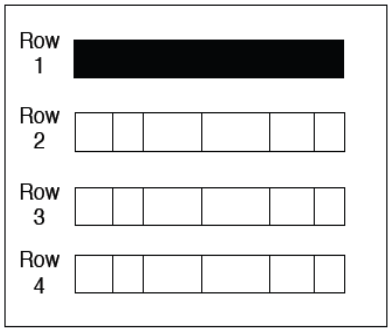
SELECT SUM(Order Value)
FROM Sales Orders
WHERE Document Date > 2009-01-20

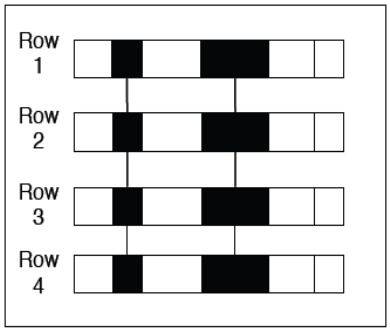
Column Store





Row Store

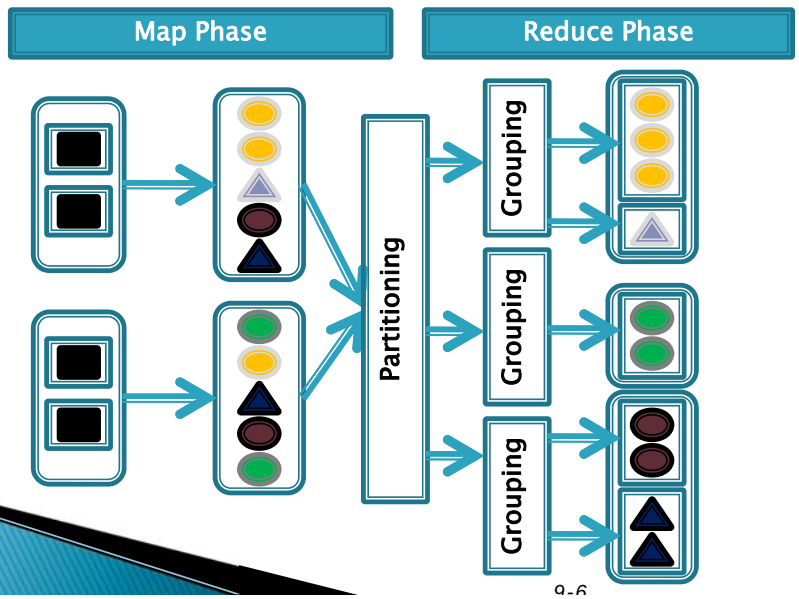




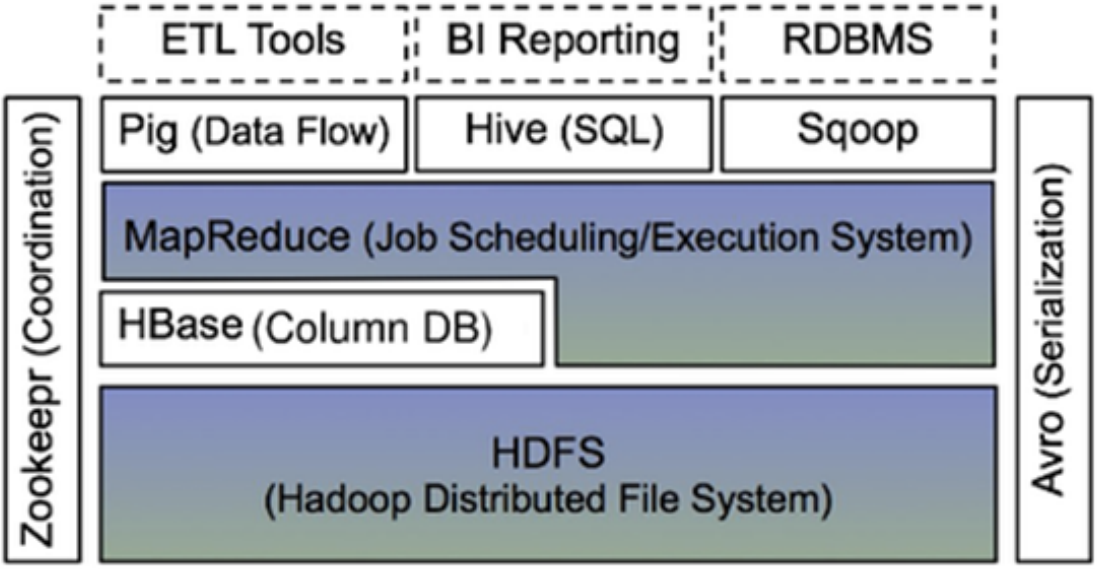
*** Quelle: Hasso Plattner: Enterprise Applications – OLTP and OLAP – Share One Database Architecture**

MapReduce

- Framework zur automatischen Parallelisierung vieler datenintensiver Anwendungen ("Web-scale" Big Data)
  - Datenparallelität
  - Key-basierte Umverteilung und Gruppierung der Daten



Hadoop-Plattform



Source: Cloudera.com

Big Data / NoSQL Datenbanksysteme

Key Value Stores / Tuple Stores

- Amazon Dynamo, Voldemort, Yahoo! Sherpa/PNUTS
- Membase, LevelDB …

Wide Column Store / Column Families

- Hadoop & Hbase
- Cassandra, Hypertable …

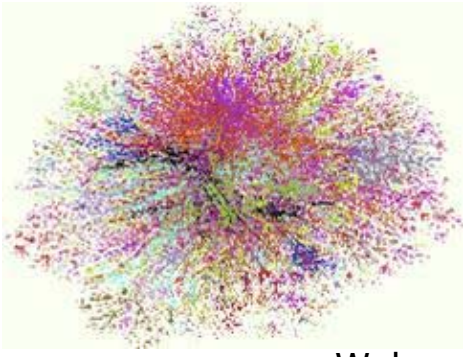
Document Stores

- CouchDB, MongoDB …

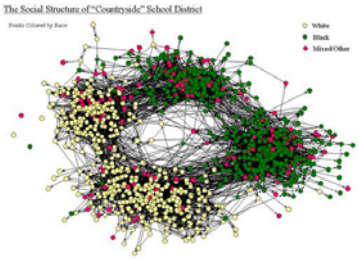
Graph Databases

- Neo4J …

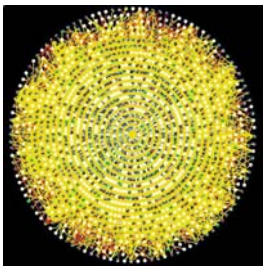
Large-Scale Graph Mining / Analysis



Web



Friendship Graph



Protein Interactions

- Strukturanalysen
- Recommendations …

Large-Scale Data Mining

- Leistungsprobleme / geringe Skalierbarkeit für viele Machine Learning-Verfahren
  - Kleine Datenmengen / Hauptspeicher
  - Nutzung ineffizienter Algorithmen als „Black Box“
- Verbesserung durch
  - Parallelisierung / Cloud-Infrastrukturen
  - Optimierte (DB-orientierte) Verfahren
  - Bessere Bibliotheken, z.B. *madlib.net*

SEMINAR

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)*
- Alte Studiengänge (Diplom, etc.)
  - Problemseminar

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 2013)
- aktive Teilnahme an allen Vortragsterminen
- Modul-Workload: 30h Präsenzzeit, 120 h Selbststudium

Seminar (3)

- Themenzuordnung
  - Koordinierungstreffen mit Betreuer bis spätestens 31.10.2012
  - ansonsten verfällt Seminaranmeldung
  - freiwilliger Rücktritt auch bis max. 31.10.2012
- Vortragstermine
  - 4x Montags, P701, ab 7. 1. 2013
  - 2-3 Doppelstunden: 9:15-, 11:15-, 13:15(-14:45)

Nr.	Thema	Termin	Betreuer	Studenten
1	Einführung Column Stores	7.1.	Hartung	Lyko, Wansing
2	Data Warehouse Appliances (IBM Netezza, SAP BI Software)	7.1.	Hartung	Recknagel, F. Jacobs, Seifert
3-4	In-Memory BI Architectures (SAP Hana, SanssouciDB, HyPer, QlikView)	14.1.	Wartner	Junghanns, Spangenberg, Westphal
5	Recommendations in sozialen Netzwerken (Facebook, Twitter, LinkedIn)	14./ 21.1.	Arnold	Chyhir, Merzdorf
6	Graph-Mining in sozialen Netzwerken (Zugriffsanalysen), Pegasus-System	21.1.	Arnold	Hartung, Webs
7	Cloud-basierte Datenanalyse (Hive, Stratosphere, Pig/PigLatin …)	21.1.	Kolb	Pfütze, Sehili
8	Data Mining und Machine Learning in der Cloud (madlib,…)	28.1.	Kolb	Sintschilin, Wermke
9-10	Large-Scale Data Analysis in speziellen Domänen (Life Sciences, Scientific Impact Analysis, Datenströme, …)	28.1.	Groß	M. Jacobs, Bugge, Christen, Schulze
	Nachrückerliste: Preßler, Hoffmann, Zühlsdorf, Zhang, Jie (Xu)			