Big-Data-Technologien
- Überblick -

Prof. Dr. Jens Albrecht

BIG DATA LAB

TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

ADIDAS SMART BALL

24.06.2014, 08:25 Uhr | 0

Elektronischer Fußball und Smartphone-App helfen beim Training

Pünktlich zur WM präsentiert Adidas einen elektronischen Fußball: Er misst mit Sensoren Schussgeschwindigkeit, Flugbahn und Spin. Auf einer Smartphone-App kann der Hobbyspieler dann schwarz auf weiß sehen, wie weit er von der Schusskraft eines Ronaldo entfernt ist.

Quelle: http://www.ingenieur.de/Panorama/Fussball-WM-in-Brasilien/Elektronischer-Fussball-Smartphone-App-helfen-Training

Big-Data-Anwendungen im Unternehmen

- **Marketing / Vertrieb / Kundenservice**
  - Was interessiert den Kunden?
  - Wie reagiert der Kunde?

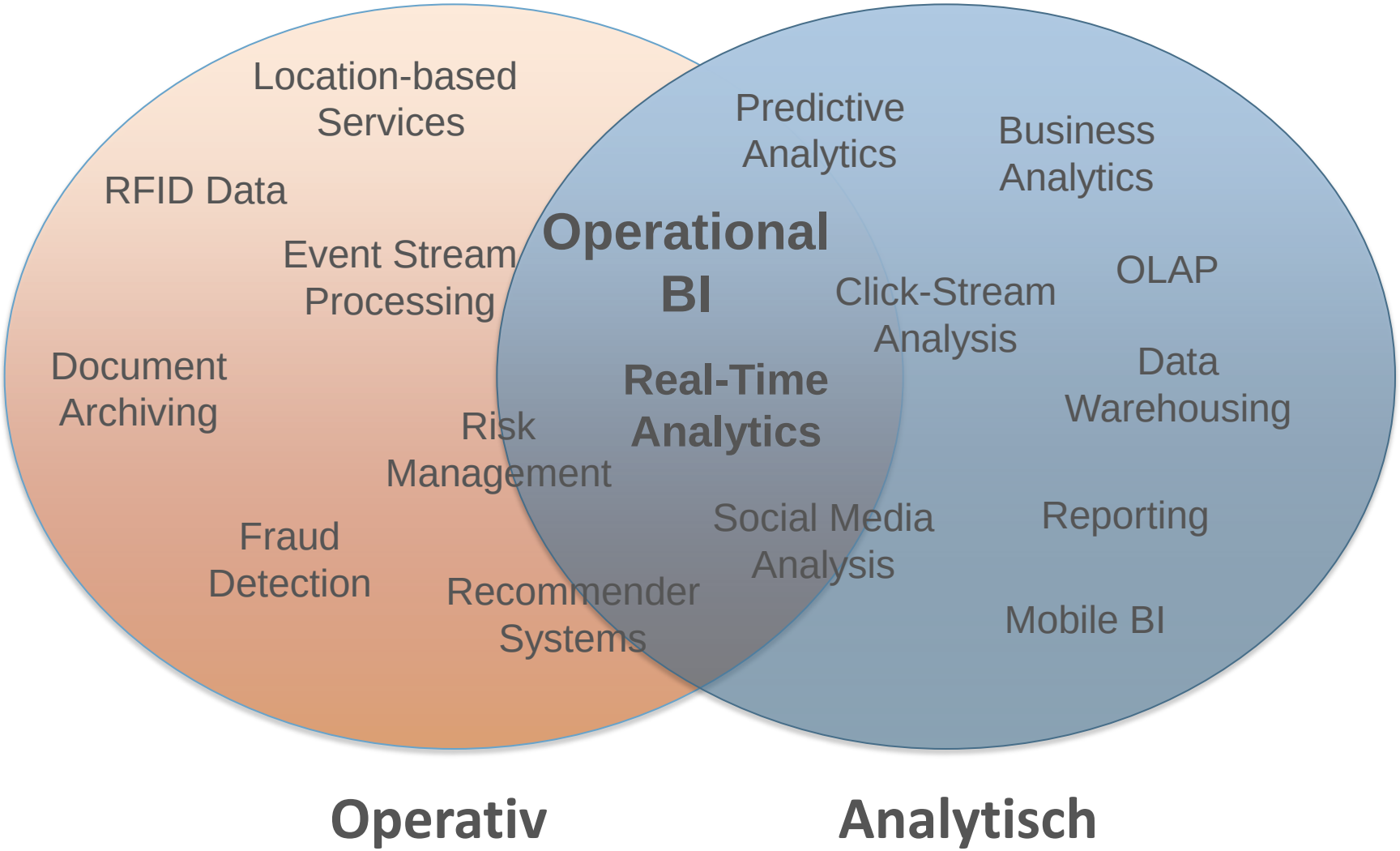
- **Produktion**
  - Qualitätsanalysen
  - Prozessoptimierung
  - Diagnose
  - Vorausschauende Instandhaltung

- **Logistik**
  - Optimierung von Beständen und Warenströmen

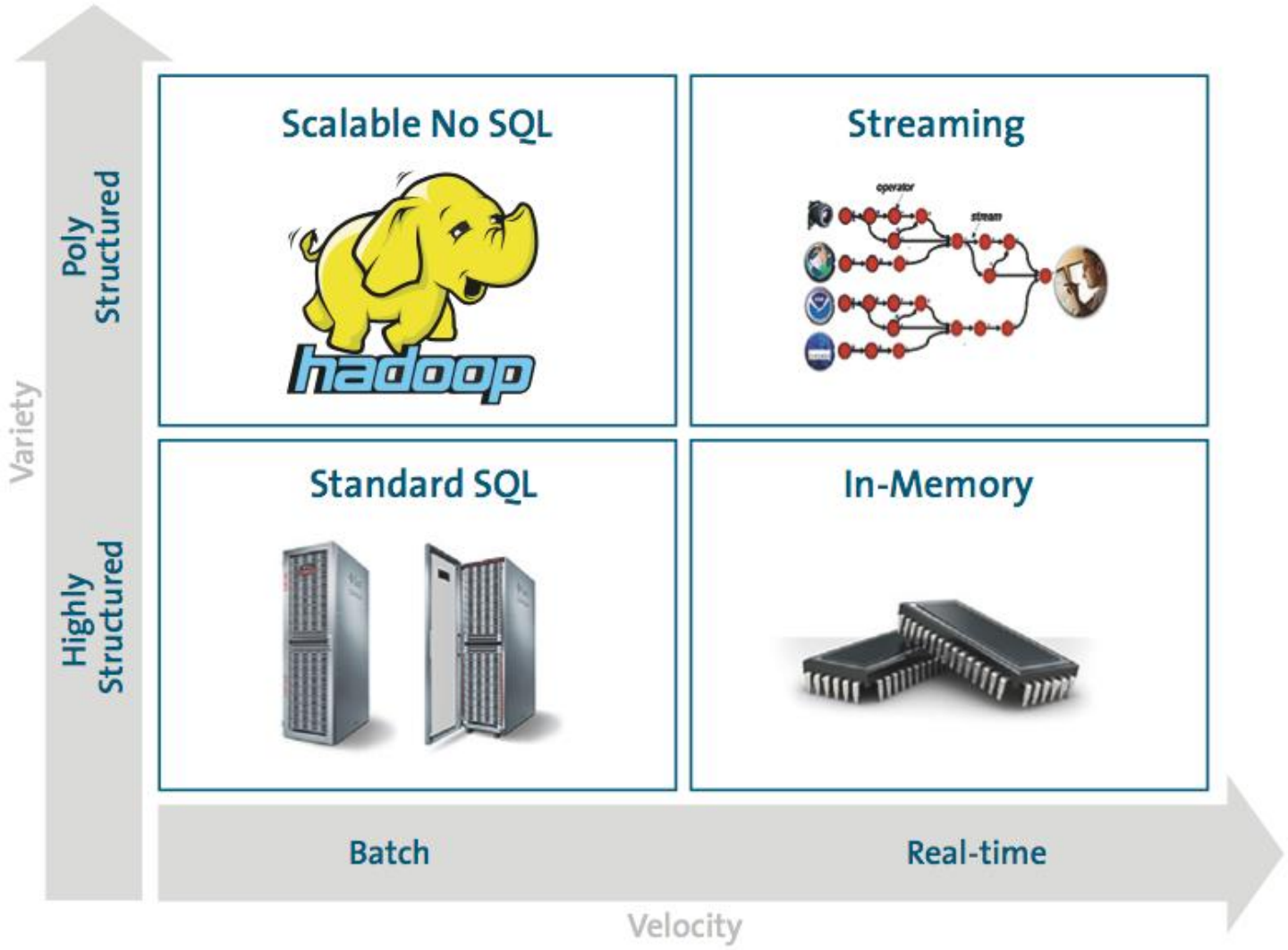
- **Controlling**
  - Analyse von Detaildaten
  - Vorhersagen
  - Risikoanalyse und Betrugserkennung

- **IT**
  - Problemsuche
  - Performance-Analyse
  - Sicherheit

Charakter von Big-Data-Anwendungen



Big-Data-Technologien im Überblick



Quelle: Big Data Technologien, Bitkom Leitfaden, 2014 (http://www.bitkom.org/de/publikationen/38337_78776.aspx)

In-Memory RDBMS

Technologie

- Spalten-orientiert
- Komprimiert
- CPU-Pipeline-optimiert

Stärken

- Vertrautes Datenmodell
- **Sehr kurze Antwortzeiten**

Limitierungen

- Gespeichertes Datenvolumen treibt Hauptspeicherbedarf
- Kosten

NoSQL-Datenbanken

Technologie

- Key-Value-Store
- Document-Store
- Wide-Column-Store
- Graph-DB

Stärken

- **Flexibles Datenmodell**
- Skalierbar (insbes. Velocity)
- Kostengünstig

Limitierungen

- **Nur einfachste Anfragen**
- **Keine Transaktionssicherheit**

JSON – JavaScript Object Notation

```
{
    "ProductId": "69451",
    "Type": "Smartphone",
    "Name": "S4",
    "Brand": "Samsung",
    "Features" : {
        "Weight": 499,
        "Colors": [ "black", "blue" ],
        "ScreenSize": 12.7,
        "CameraResolution": "13 Megapixel"
    }
}
```


```
{
    "ProductId": "78462",
    "Type": "Shoe",
    "Name": "Timberland Classic",
    "Brand": "Timberland",
    "Weight": 499,
    "Features" : {
        "Weight": 1400,
        "Sizes": [ 41, 42, 43, 44, 45 ],
        "Material": "Leather"
    }
}
```

Hadoop

Technologie

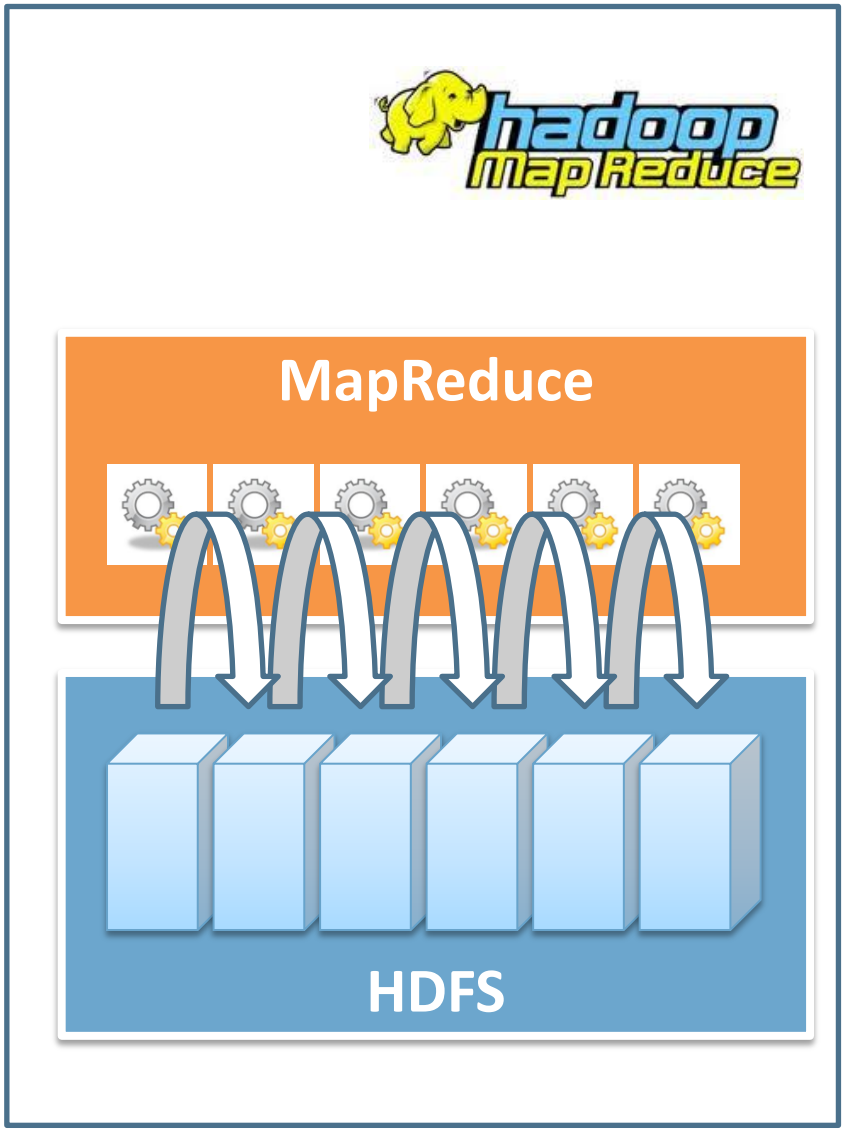
- HDFS
- Map-Reduce
- Hive, Pig, u.a. Tools
- Neu: Spark (In-Memory)

Stärken

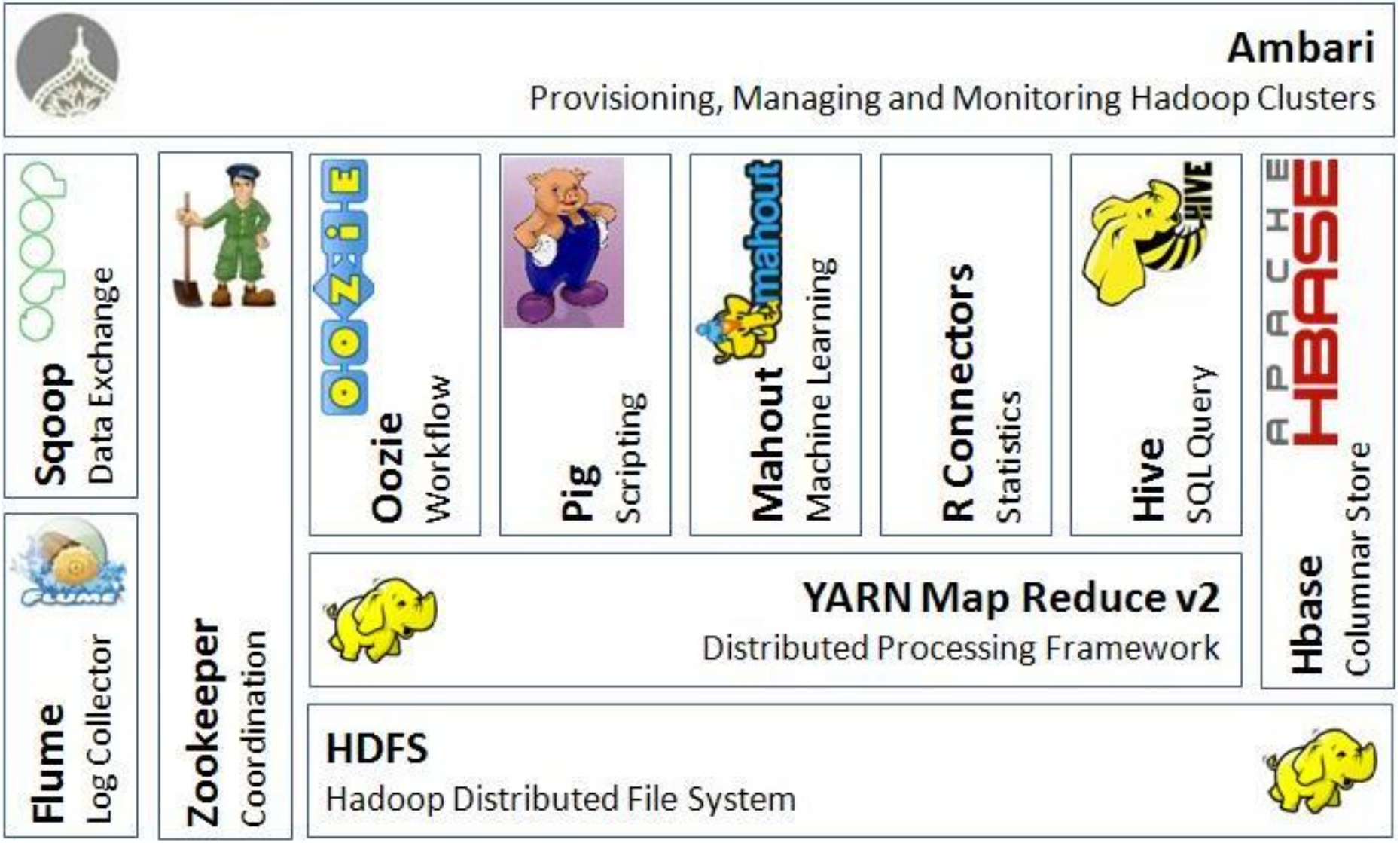
- Skalierbar (insbes. Volume)
- **Schema-on-Read**
- Kostengünstig
- Schnittstellen zu fast allen RDBMS verfügbar

Limitierungen

- Optimiert für Batch-Verabeitung



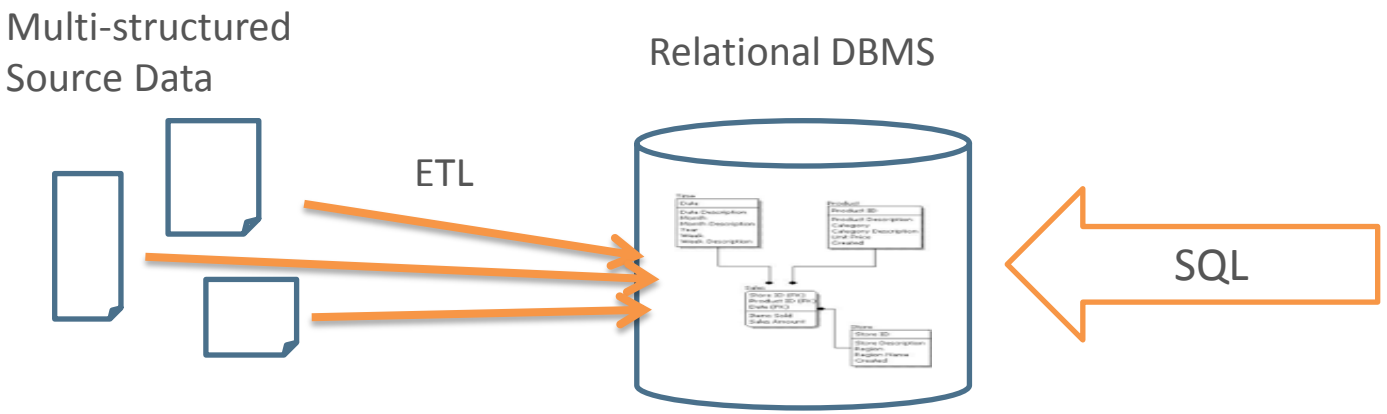
Hadoop Ökosystem



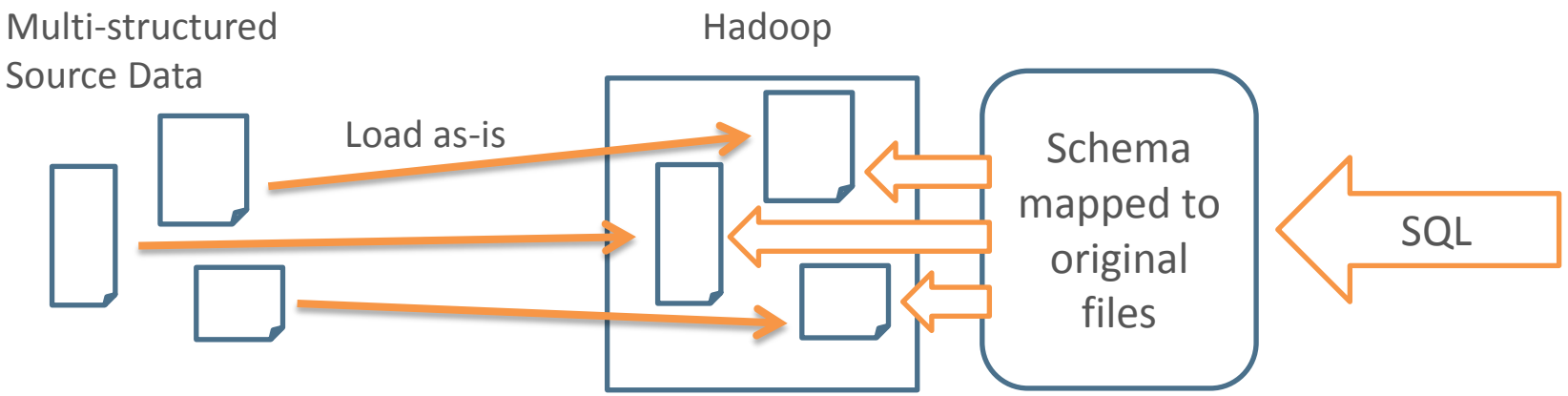
Quelle: http://techblog.baghel.com/index.php?itemid=132

Schema-on-Write vs. Schema-on-Read

- **Relational Database: Schema-on-Write**

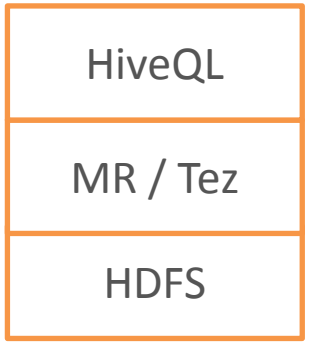


- **Big Data Processing: Schema-on-Read**



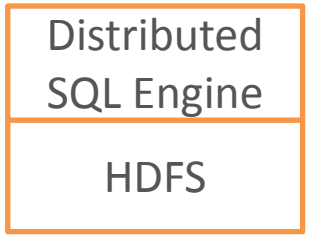
Hadoop-SQL Integration

Hive (Native Hadoop)



Hortonworks Stinger

Pure Hadoop SQL Engines



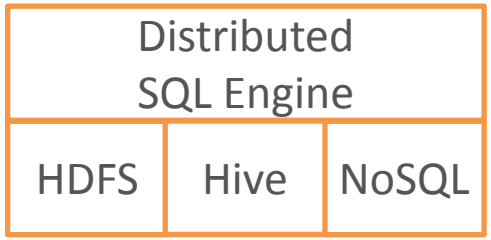
cloudera IMPALA

IBM Big Insights

EMC² PIVOTAL HD HAWQ

Spark

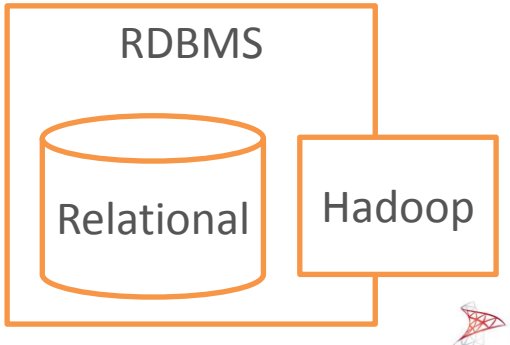
Format-agnostic SQL Engines



APACHE DRILL MAPR

presto facebook

RDBMS with Hadoop Access



Microsoft SQL Server

ORACLE

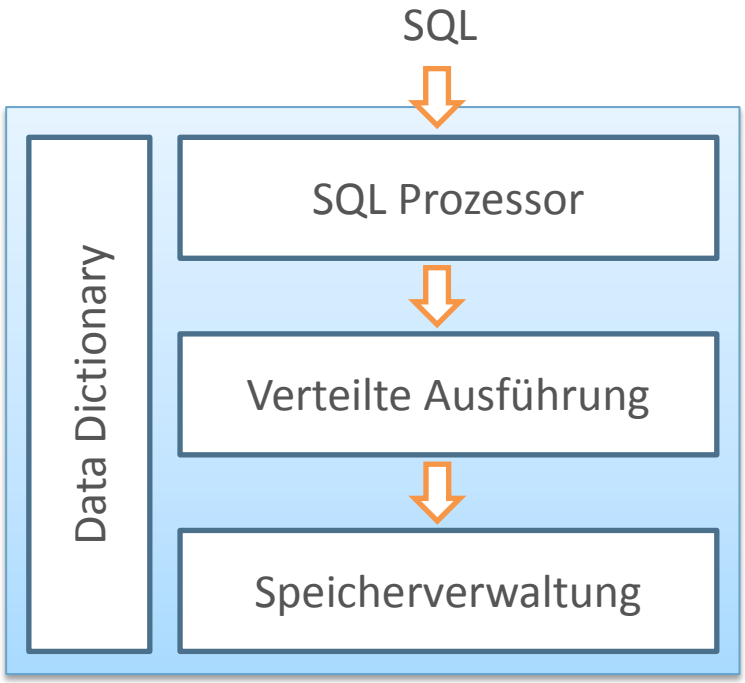
Beispiel: Apache Drill - SQL für heterogene Daten

```
select USERS.name, USERS.emails.work
from
  dfs.logs.`/data/logs`  LOGS,
  dfs.users.`/profiles.json`  USERS,
where
   LOGS.uid = USERS.uid and
   errorLevel > 5
order by count(*);
```

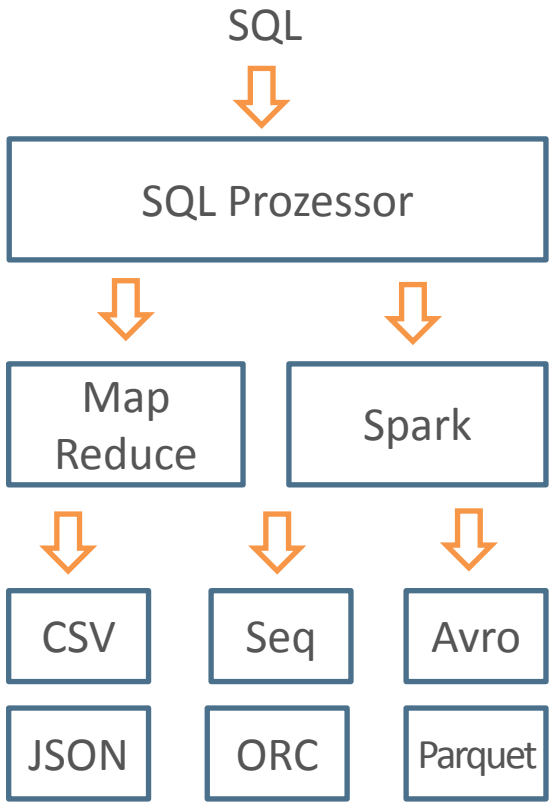
- **Formate**
  - JSON, CSV
  - ORC, Parquet
  - HBase, Hive

Datenbanken als Lego-Baukasten?

Klassisches Monolitisches System



Hadoop Baukasten

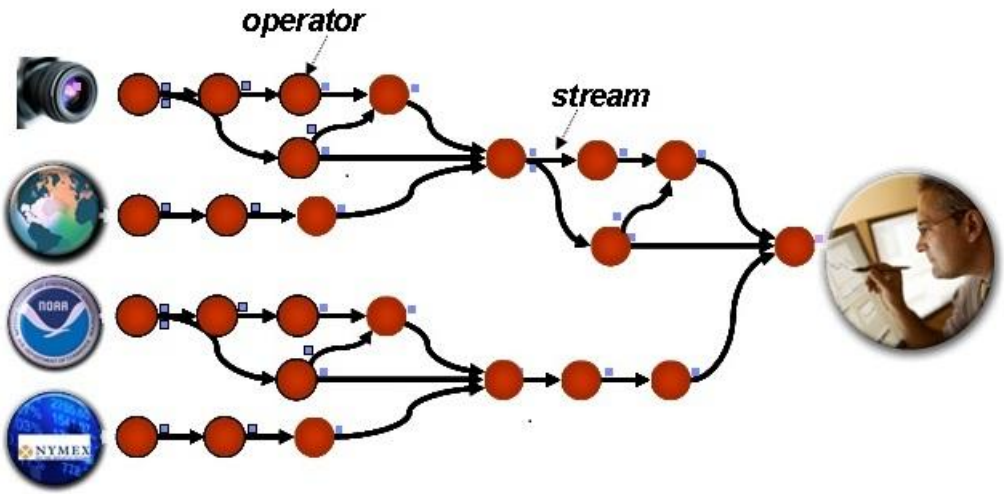


- Generische Ausführungs-Engine
- Metadaten-Sharing über Hive Repository oder selbstbeschreibende Dateiformate
- Operatoren-Push-Down durch intelligente Dateien

Data Stream Processing on Hadoop

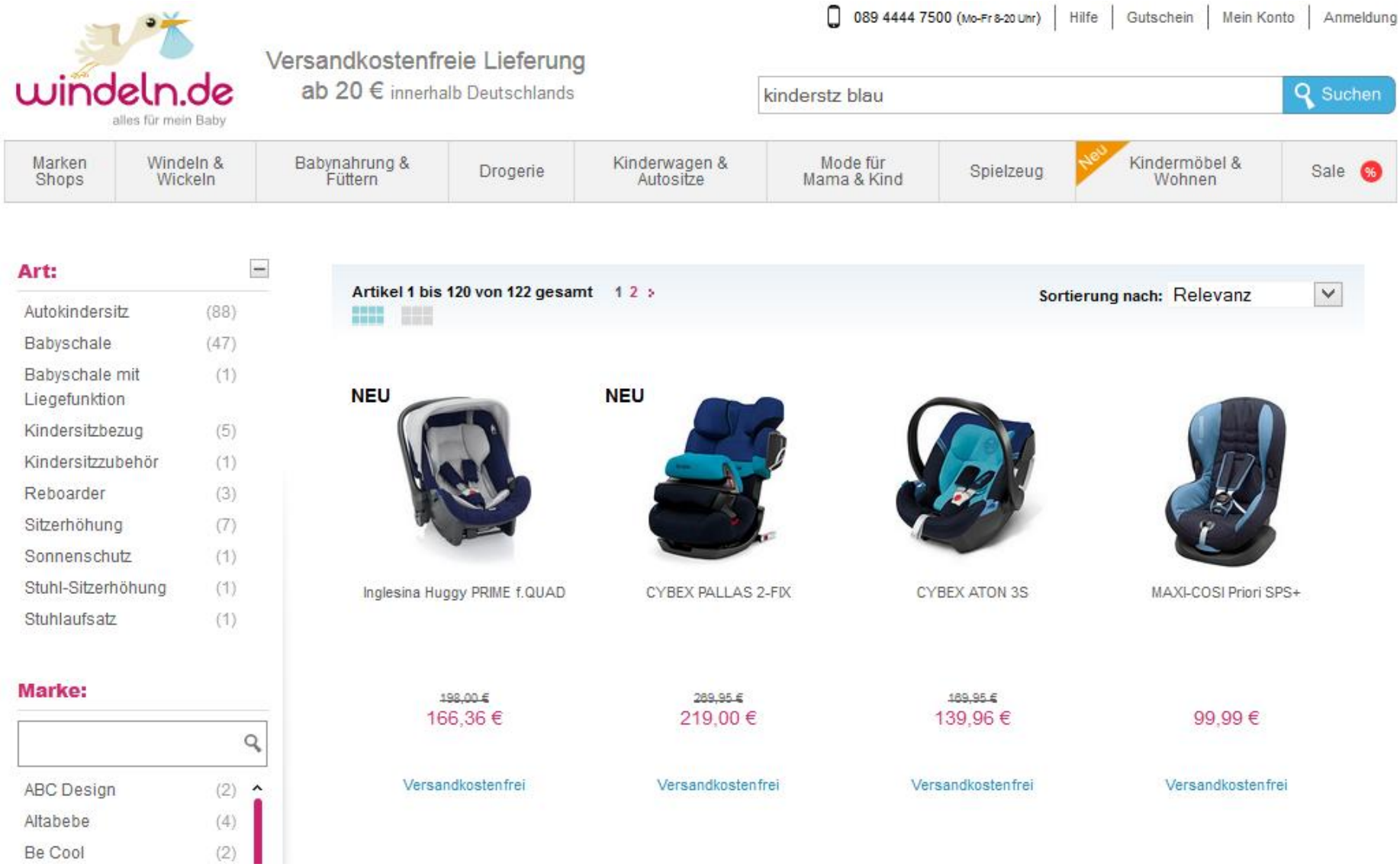
- **Datenstromverarbeitung: "Big Data in Motion"**
  - Kontinierliche Verarbeitung von Events
  - Parsen, Anreichern, Filtern, Aggregieren, Joinen
  - Auswertungen beziehen sich auf (kurze) Zeitfenster

- **Hadoop-Erweiterungen: Skalierbar, ausfallsicher**
  - Kafka (Linked-In): Message-Queue
  - Storm (Twitter): Framework zur Datenstrom-Verarbeitung



http://researcher.watson.ibm.com/researcher/view_group.php?id=2531

Suche mit Solr/ElasticSearch



Schritte in Richtung Big Data

Aspekt	Schritt in Richtung Big Data: Prüfung des Einsatzes von...
Datenintegration	... linguistischer Datenverarbeitung zur Auswertung von Textdokumenten.
Verarbeitungsgeschwindigkeit und Skalierbarkeit	... In-Memory-Technologien oder dedizierten Appliances für transaktionale Systeme
Analyse und Speicherung großer Datenmengen	... Hadoop-Systemen zusammen mit bestehenden Data-Warehouse-, BI- und ETL-Systemen
Entscheidungsfindung	... CEP-Technologien zur Verarbeitung bestehender Datenströme
Investitionskosten	... standardisierter, eventuell quell-offener Software und von bestehenden, direkt abrufbaren Cloud-Lösungen
Entwicklungs- und Analysezyklen	... explorativen Analyseansätzen und agilen Projektmanagement-Methoden für die Weiterentwicklung bestehender Systeme

Quelle: BITKOM 2012

Zusammenfassung

- **Big Data Technologien erweitern die IT**
  - Schneller, höher, weiter
  - **Vor allem: Operativer**

- **"Klassische" BI-Probleme sollten erst gelöst werden**
  - Performance, Anbindung von Datenquellen
  - Data Governance

☛ **Der Zug kommt erst noch!**
  - Klären, wohin die Reise gehen soll
  - Big Data Strategie hilft, Weichen zu stellen

Fragen und Antworten

- **Kontakt**
  - jens.albrecht@th-nuernberg.de
  - Big Data Lab e.V. (http://bigdata-lab.de)
  - XING, Linked-In

☛ **Beratung, Training, Hochschul-Kooperation**