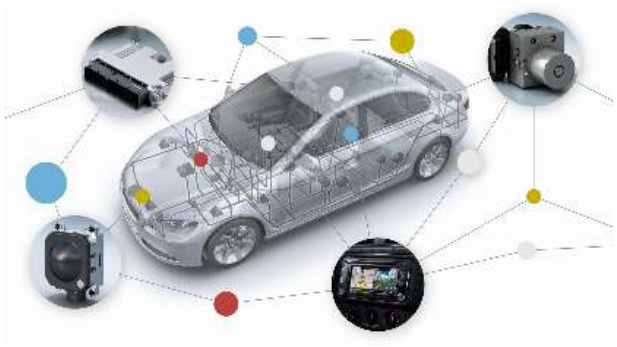
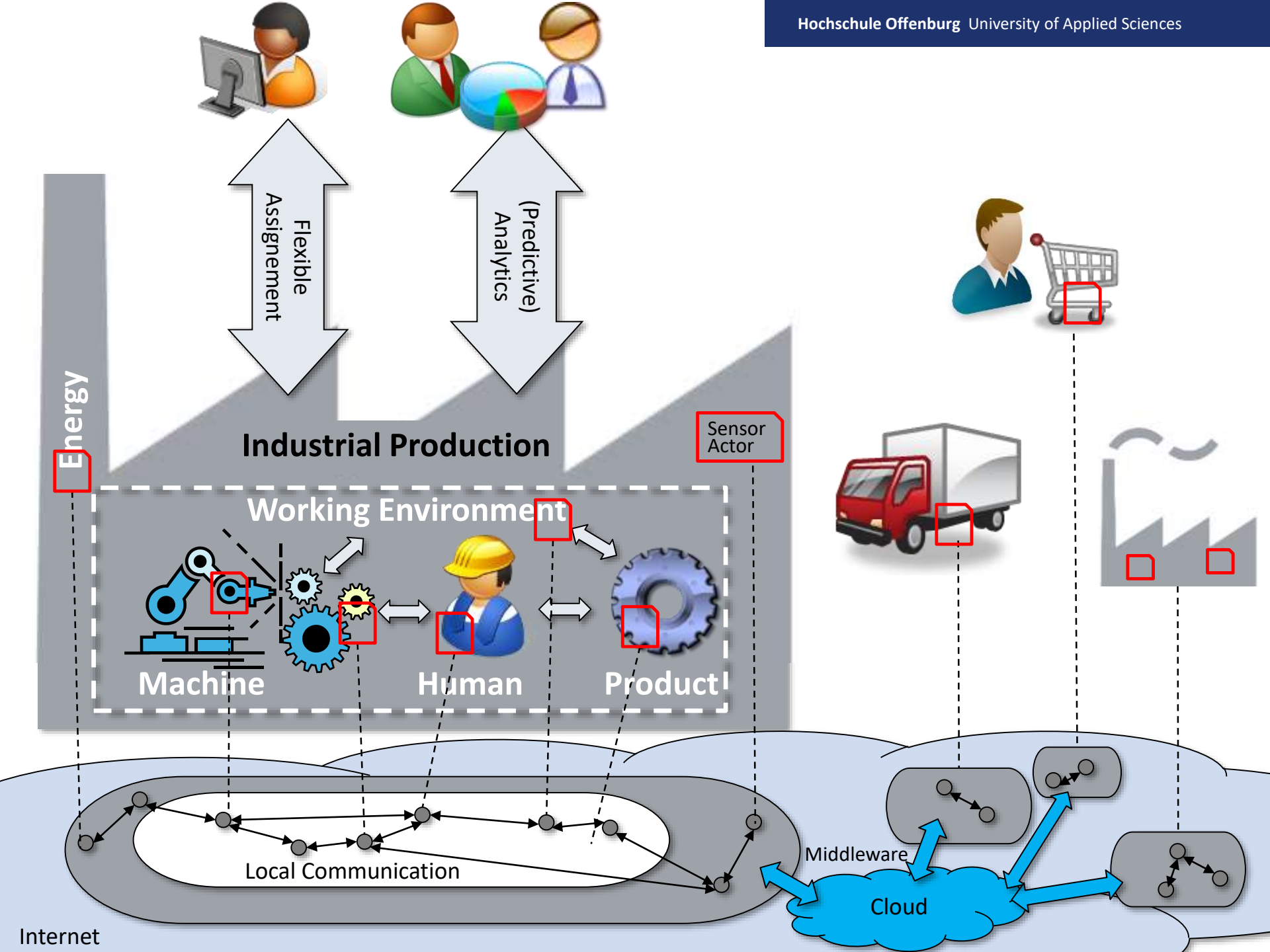
Einführung
Big Data Analytics

Prof. Dr. Stephan Trahasch
Hochschule Offenburg

Digitalisierung erfasst immer mehr Bereiche unseres Lebens und der Wirtschaft

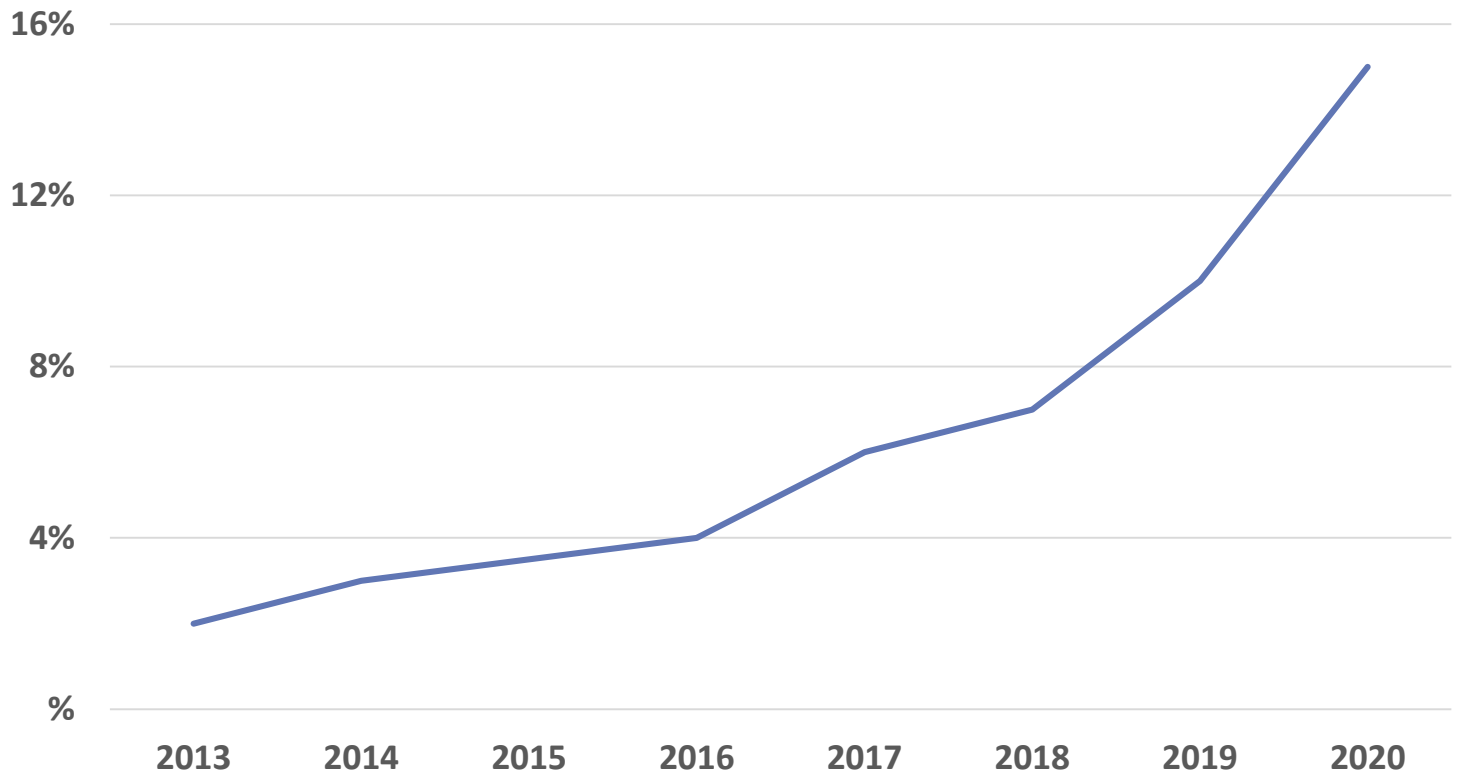


Wahl des Papstes


Flexible Assignement
(Predictive) Analytics
Energy
Industrial Production
Sensor Actor
Working Environment
Machine
Human
Product
Local Communication
Middleware
Cloud
Internet

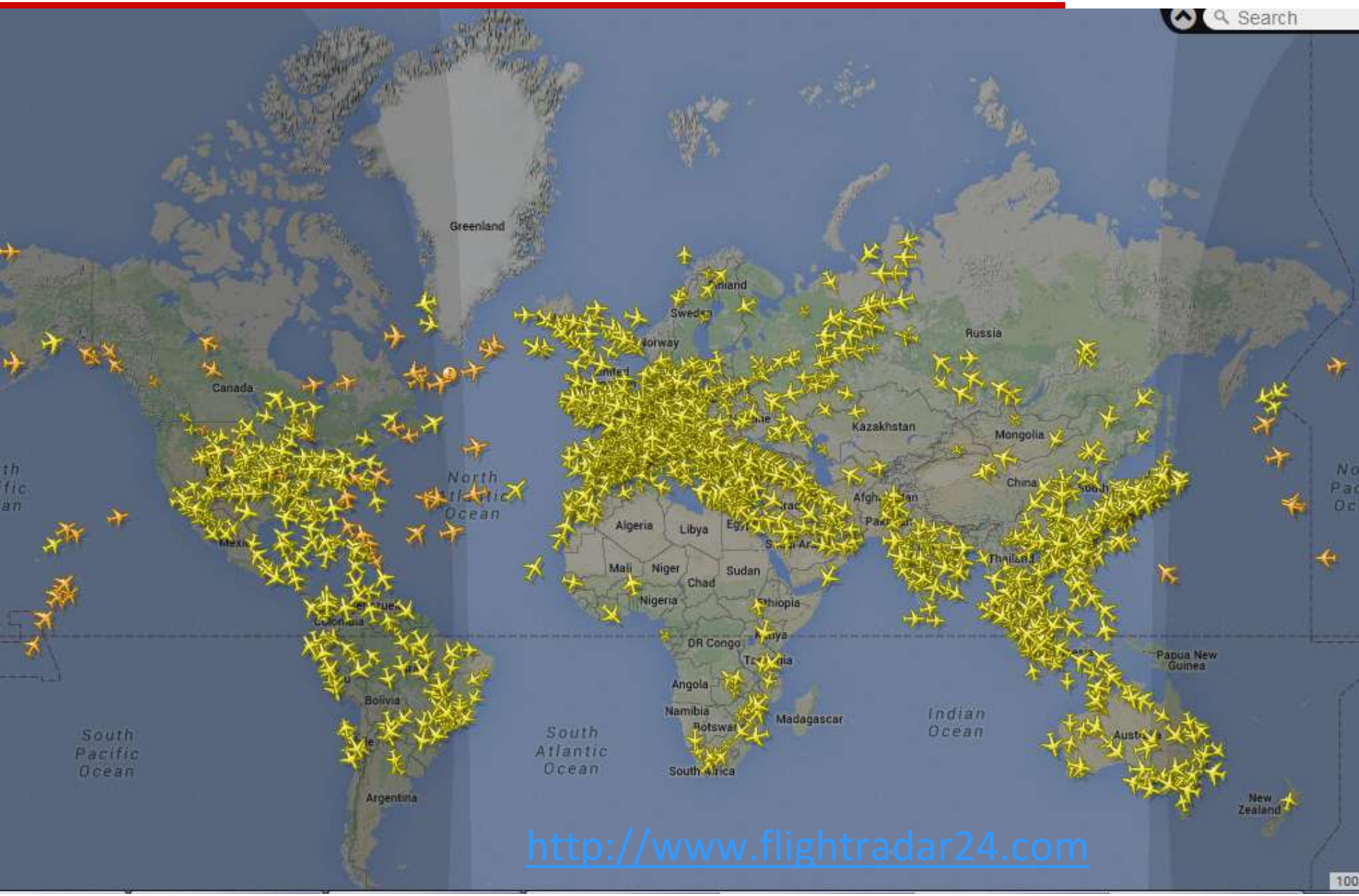
Internet der Dinge wird größeren Anteil am digitalen Universum in Deutschland haben

Anteil von Internet of Things / Embedded Systems Geräten am Datenverkehr in % in Deutschland



Quelle: IDC, 2014

Data never sleeps



Engineering

Source: Wikipedia (Three Gorges Dam)

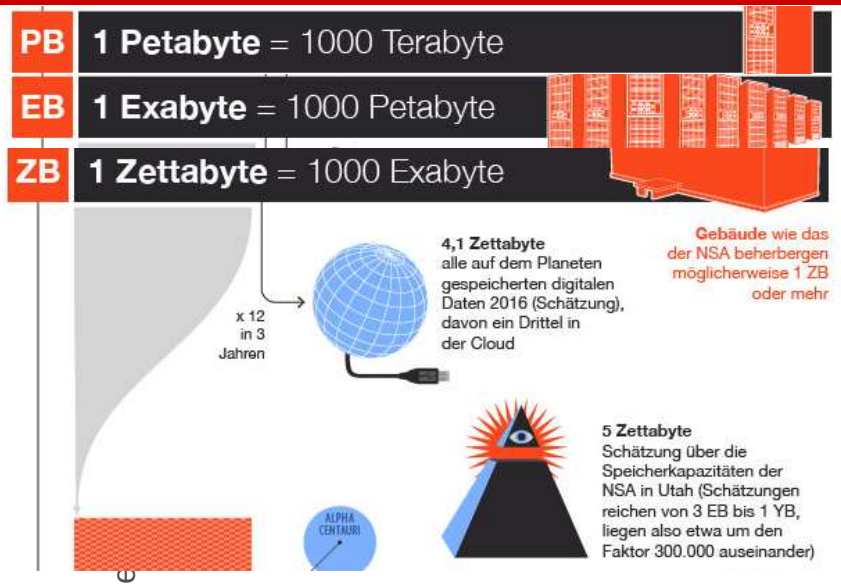
Maximilien Brice, © CERN

Maximilien Brice, © CERN

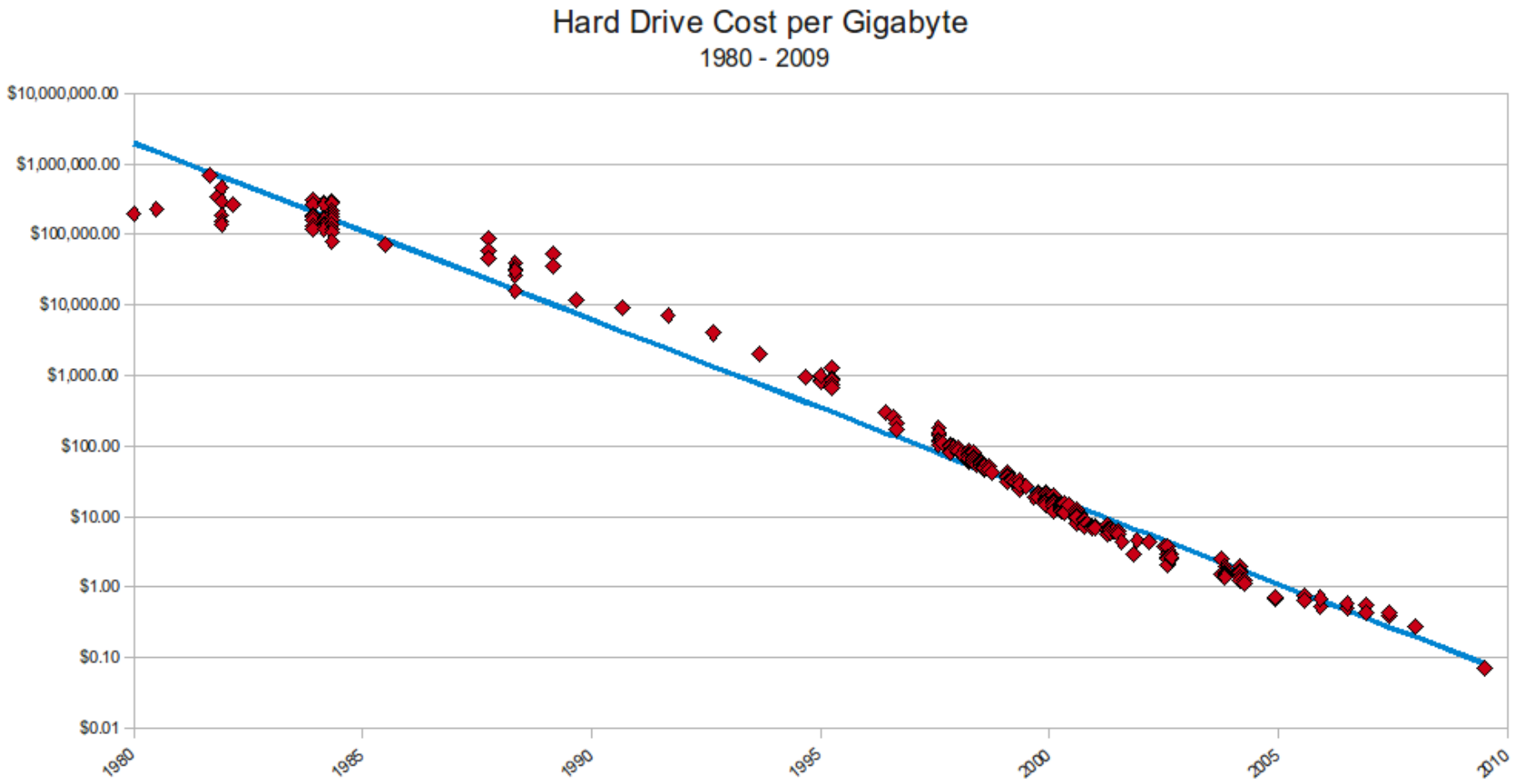
Transport

https://commons.wikimedia.org/wiki/File:Airbus_A350-900_XWB_Airbus_Industries_(AIB)_MSN_001_-_F-WXWB_(9D87432464).jpg

Prognose zum Volumen der jährlich generierten digitalen Datenmenge weltweit



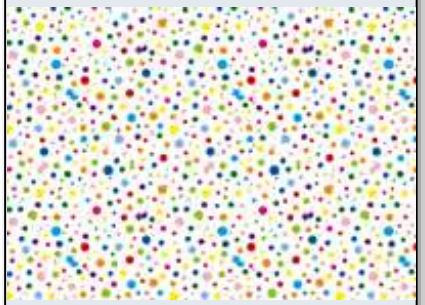
Quelle: IDC, http://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen

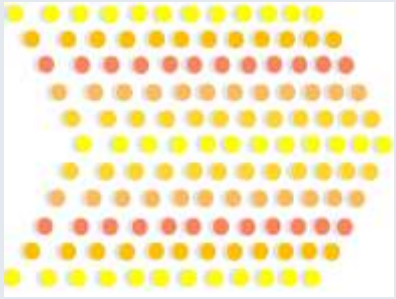


http://ns1758.ca/winch/winchest.html

Charakterisierung von Big Data durch drei V's:

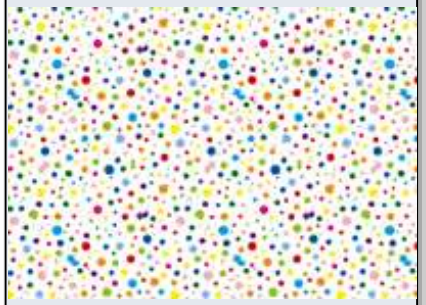
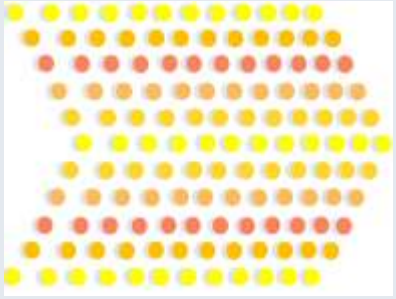
Volume	Velocity	Variety
Data at scale Terabyte to Petabytes	Data in motion Streaming Data	Data in many forms Structured, semi-structured, unstructured





Diese Definition der Eigenschaften von Big Data erfolgt durch Gartner 2011 [1]. Das darin verwendete 3-V-Modell geht auf einen Forschungsbericht das Analysten Doug Laney zurück, der die Herausforderungen des Datenwachstums als dreidimensional bezeichnet hat [2].

Manchmal auch 4 Vs oder auch mehr Vs…

Volume	Velocity	Variety	Veracity
Data at scale Terabyte to Petabytes	Data in motion Streaming Data	Data in many forms Structured, semi-structured, unstructured	Data uncertainity Unzuverlässigkeit und Unschärfe
			

Lösungsansätze

Storage

- **Apache Hadoop File System**
- Cloud

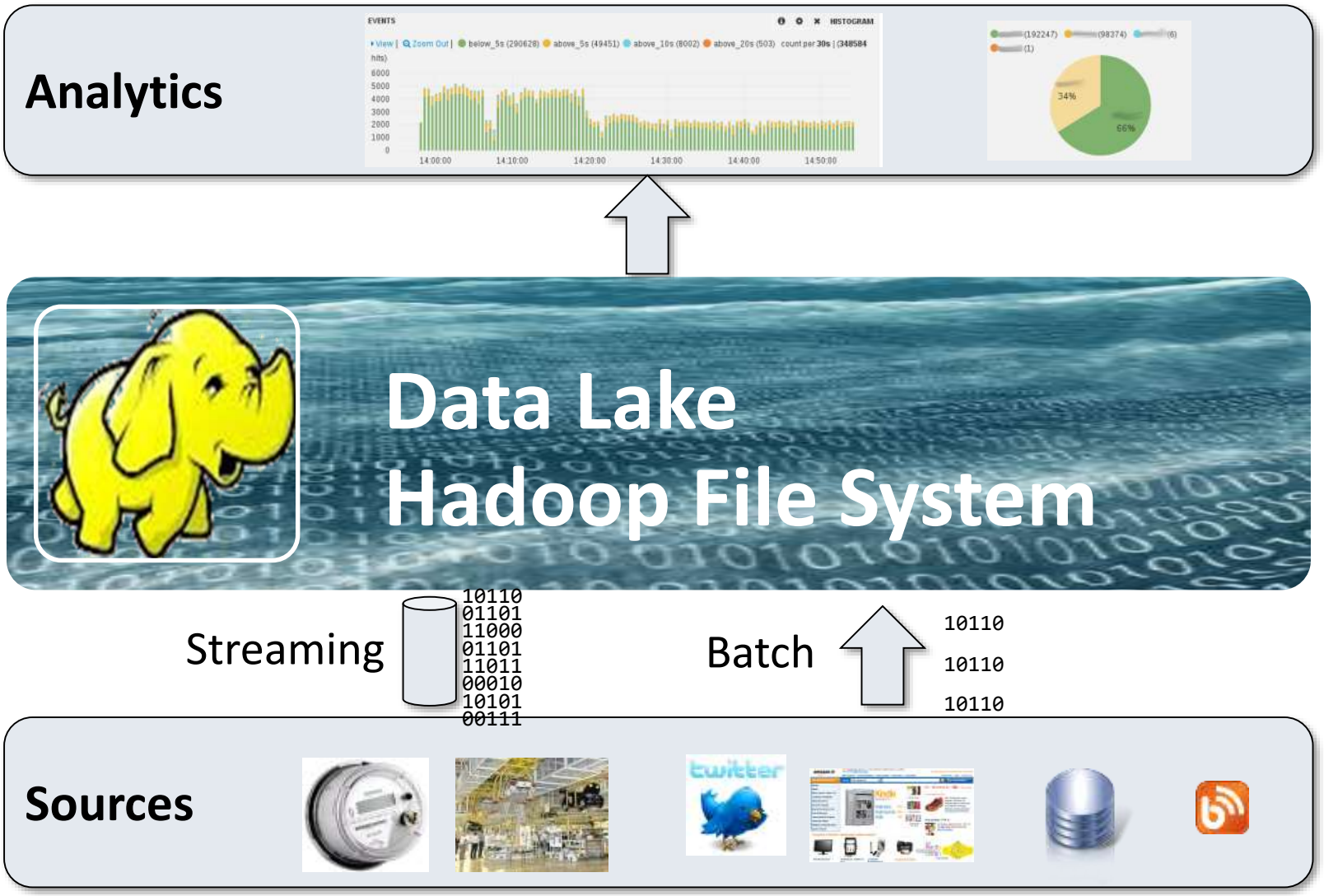
Performance

- Scale-out
- **In-Memory-Datenbanken Spark**
- GPUs

Algorithms

- **MapReduce**
- Streaming
- Data Mining

Lösungsansätze für Volume: Speicherung in einem skalierbaren Data Lake



Skalierung am Beispiel eines Transportunternehmens

http://www.acclaimedmovers.com/blog/wp-content/uploads/2013/02/Movers-Bronx-NY.jpg

Skalierung am Beispiel eines Transportunternehmens

Skalierung am Beispiel eines Transportunternehmens

http://www.unsafepictures.com/Overloading.htm

Skalierung am Beispiel eines Transportunternehmens

Skalierung am Beispiel eines Transportunternehmens

Skalierung am Beispiel eines Transportunternehmens

$5.000.000 Anschaffung

Spezial-Know-how für Betrieb und Wartung

$46.000 pro Reifen

Hoher Schaden bei Ausfall

Beispiel von P. Thoma, inovex

Skalierung am Beispiel eines Transportunternehmens

40.000€ Anschaffung pro Fahrzeug

skaliert in beide Richtungen

modernisierbar

keine Spezialkenntnisse erforderlich

Ausfall einzelner Fahrzeuge kompensierbar

Scale up vs. Scale out

Scale up: big single node system

Source: IBM

Move data to the server

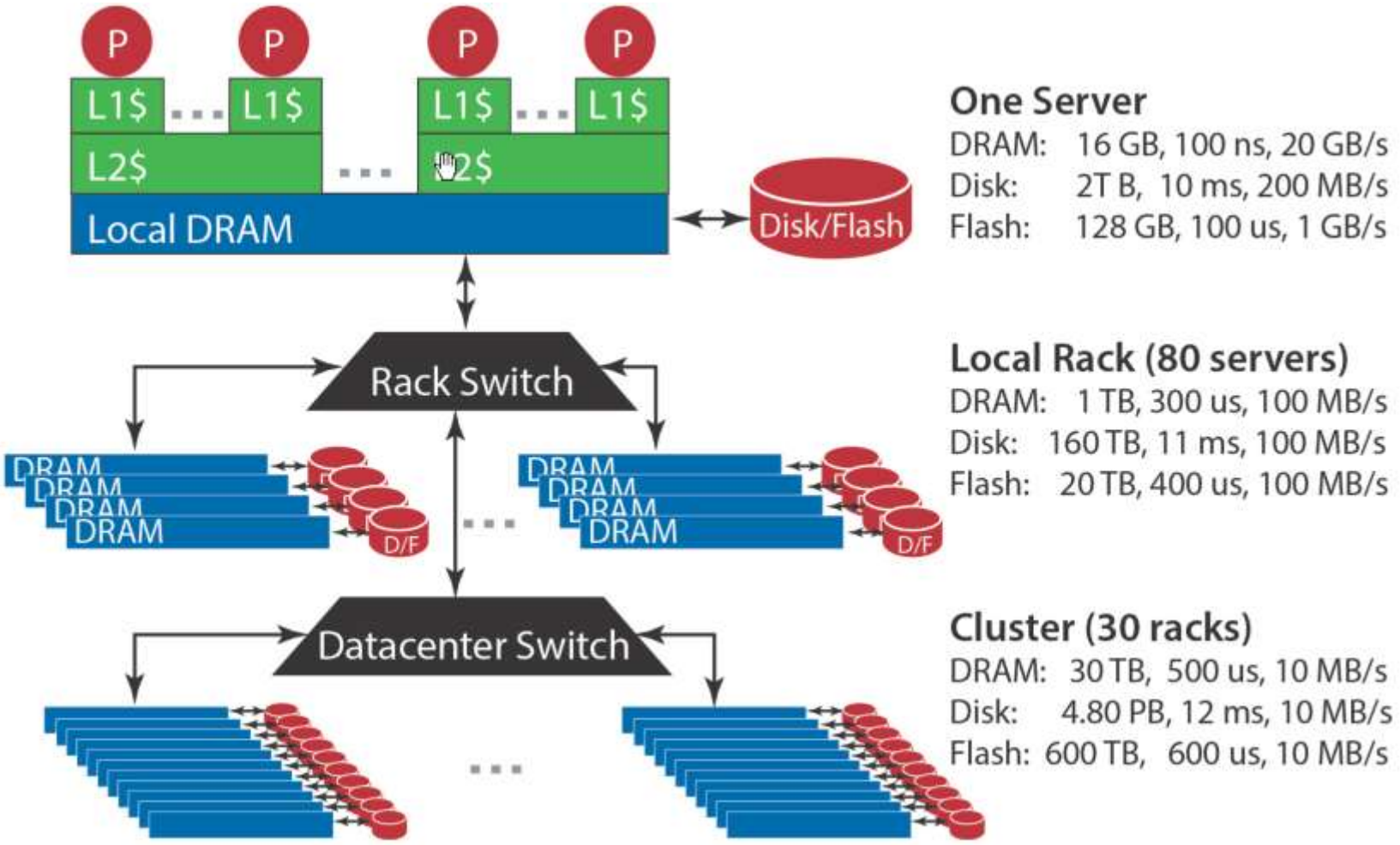
Scale out: many “small” nodes

Source: Google

Move algorithm to data

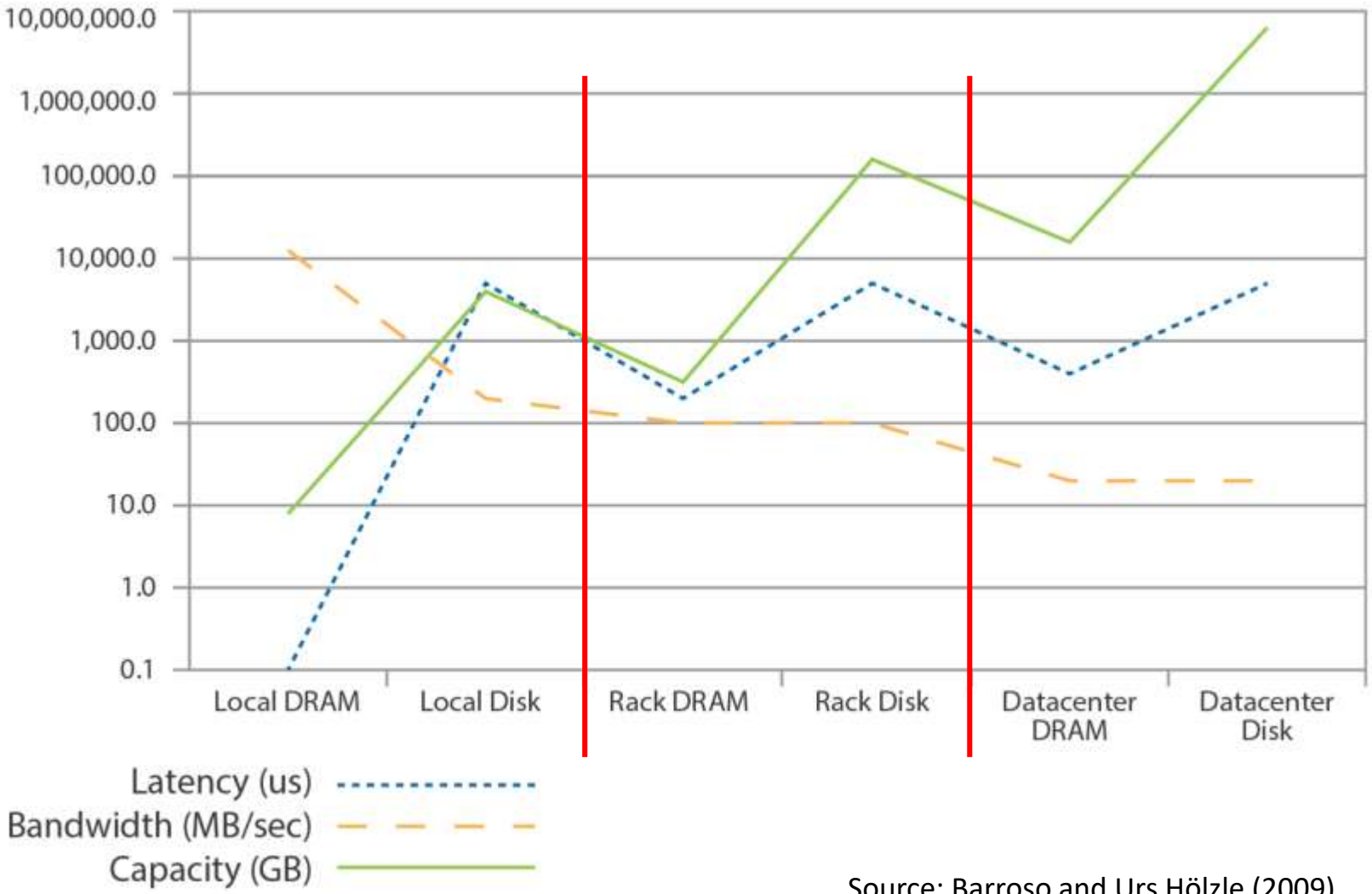
Source: Google

Storage Hierarchy



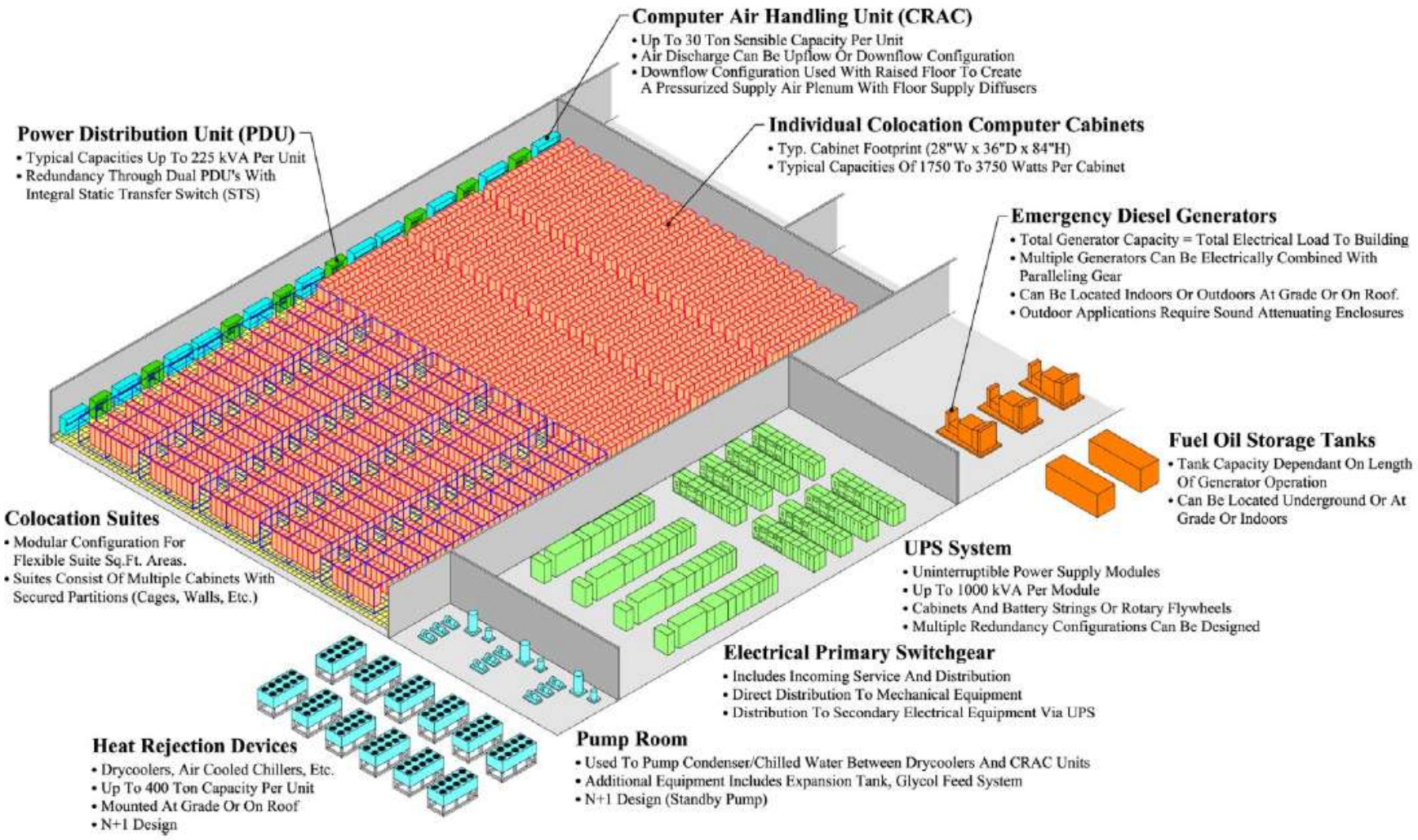
Source: Barroso and Urs Hölzle (2009)

Storage Hierarchy



Source: Barroso and Urs Hölzle (2009)

Anatomy of a Datacenter



Source: Barroso and Urs Hölzle (2009)

Big Data Problem

Petabytes an Daten, die auf mehrere Computer verteilt, verarbeitet und analysiert werden müssen.

Daten sind zu groß, um sie auf einer einzigen Maschine sequenziell zu bearbeiten

- Wie oft kommen welche Wörter in Textdateien (z.B. Suchlogs) vor?
- Funktionalität (Wörter zählen) trivial implementierbar

Aber wie verteilen?

Hadoop Distributed File System = Speicherung

MapReduce = Batch-Verarbeitung

Verschiebe Programme zu den Daten

Daten werden in einem verteilten Dateisystem auf „Commodity Hardware“ gespeichert .

- GFS (Google File System) für Googles MapReduce
- HDFS (Hadoop Distributed File System) für Hadoop

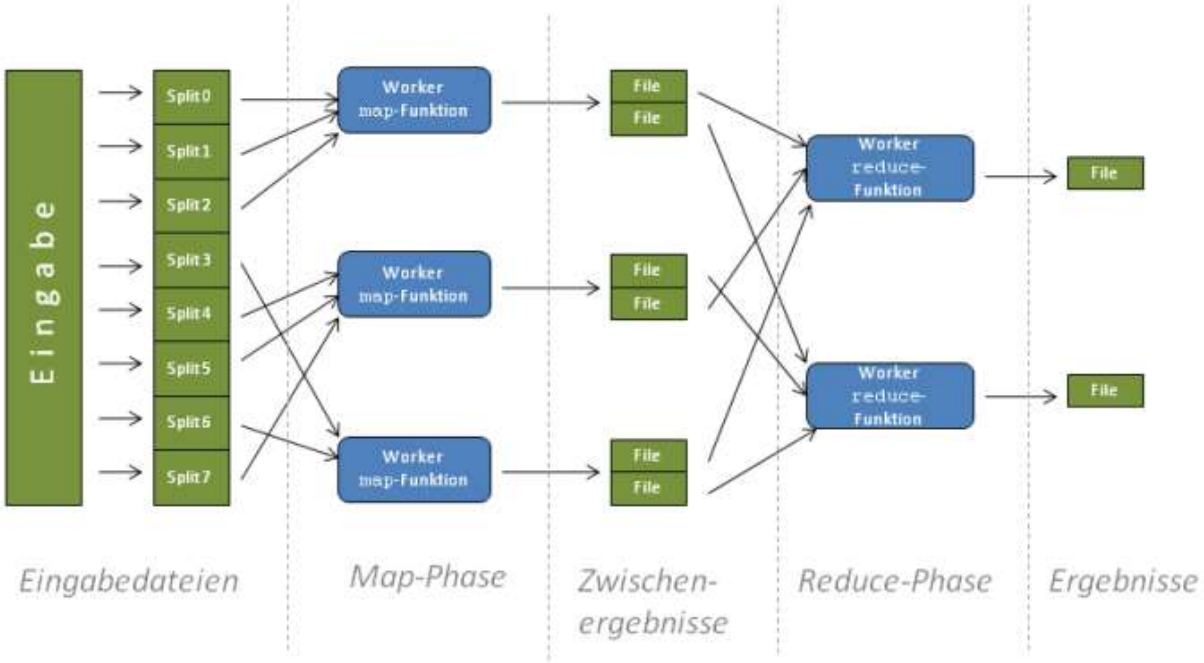
Verschiebe nicht die Daten zu den Programmen, sondern die Programme zu den Daten!

- Daten sind auf den lokalen Festplatten der einzelnen Knoten im Cluster gespeichert
- Starten der Workers, an welchen Daten lokal vorliegen

Idee von Map und Reduce

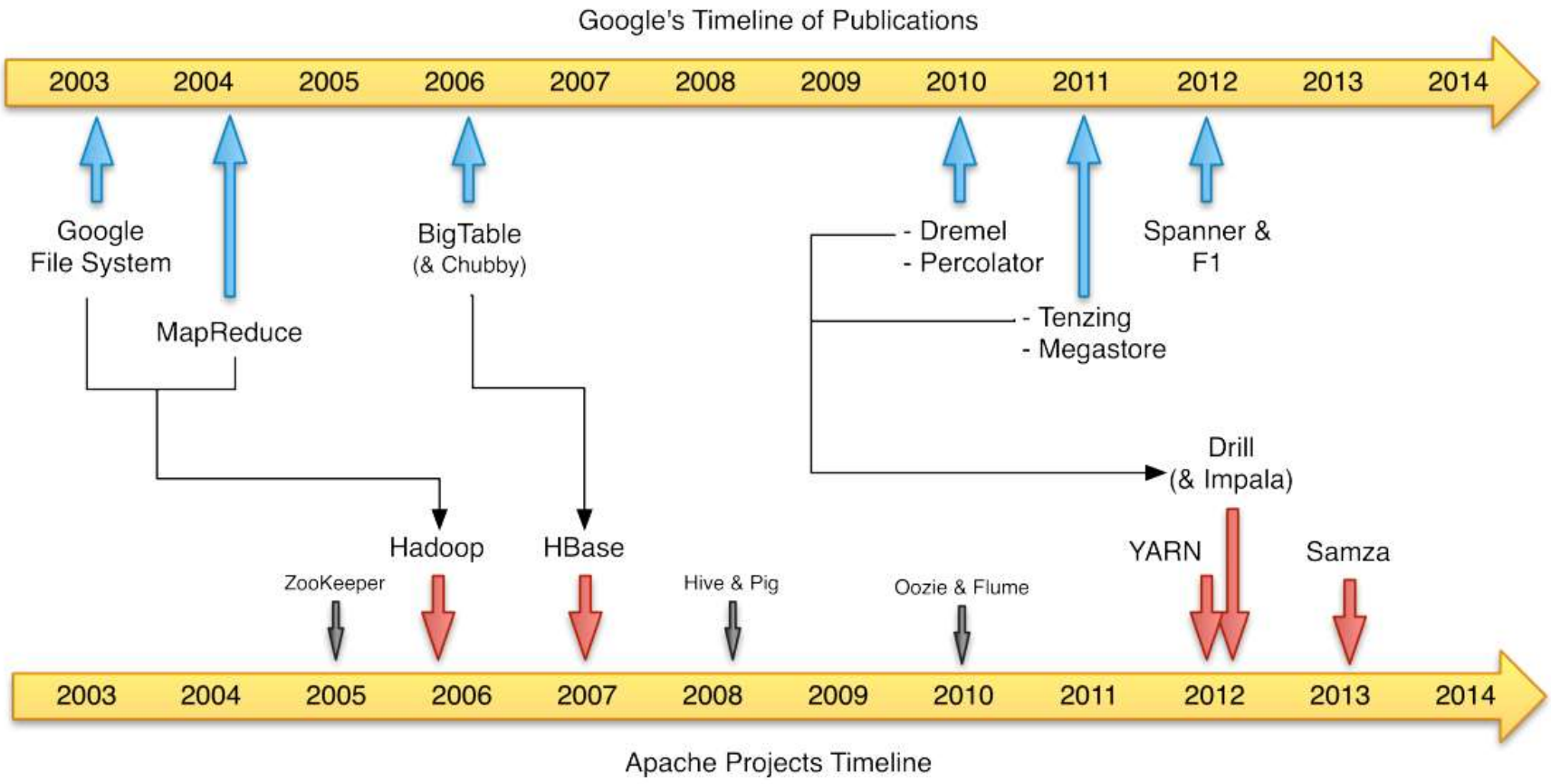
Map
Generiert aus Eingabedaten eine Sammlung von Zwischenergebnissen

Reduce
Generiert daraus aggregierte Ergebnisse



http://www.heise.de/developer/artikel/Programmiermodell-und-Framework-964823.htm

Viele Ideen stammen von Google



Idee stammte von Google, erste Hadoop Implementierung von Doug Cutting

Problem: Ganzes Web downloaden und verarbeiten

20+ Milliarden x 20 KB = 400+ terabytes

Lesen von einer Festplatte = 3 Monate

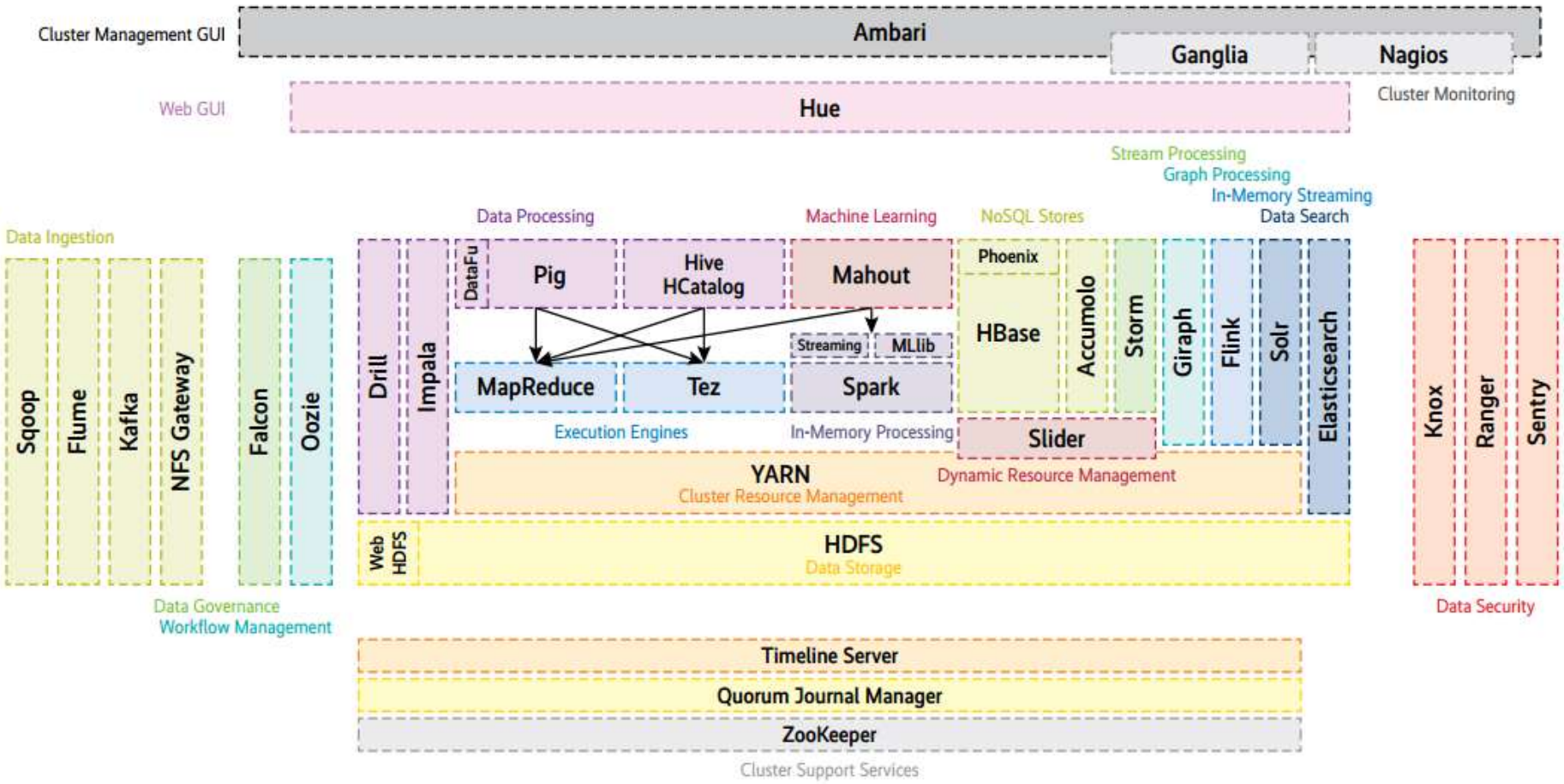
Doug Cutting

Anwender


Adobe
IBM
YAHOO!
Microsoft
Apple
LinkedIn
The New York Times
twitter
facebook
Chevron
QUALCOMM
ebay
SAMSUNG
Expedia

Heute Hadoop-Zoo:
Alles ist da, was man braucht. Und alles Open Source.

Viele Tools. Welche (Versionen) passen zusammen?



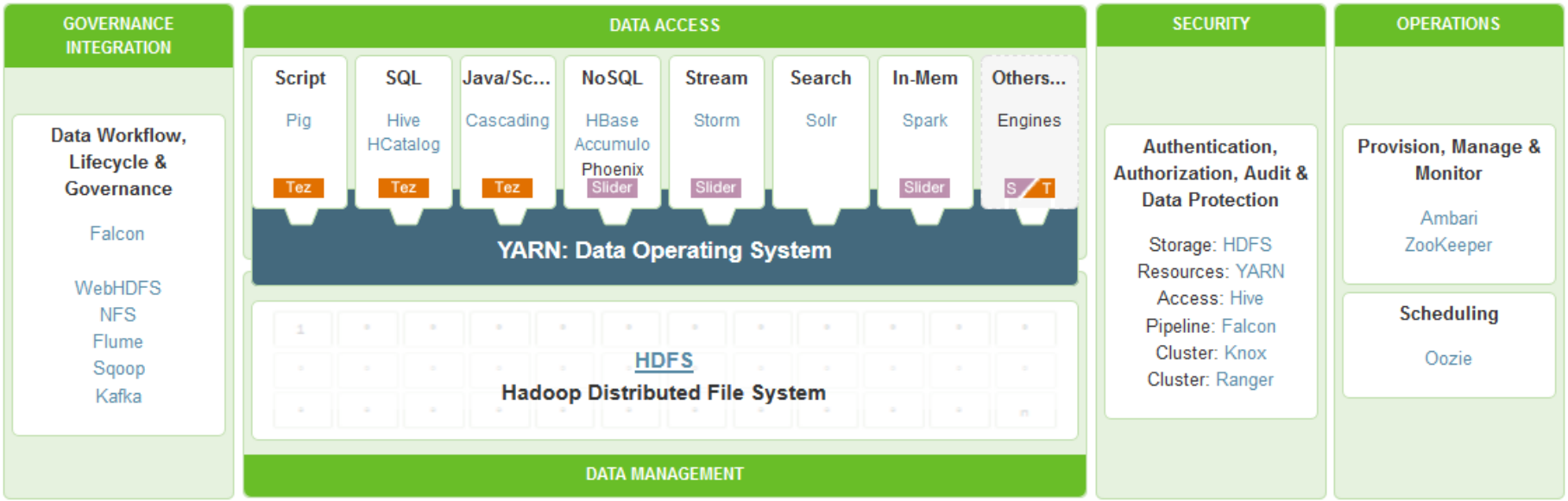
Hadoop Distributionen

- Quelloffen und Apache lizensiert
- Viele Entwickler von namhaften Firmen
  - Cloudera, Hortonworks, Google, Yahoo!, Facebook, etc.
- Viele zusammengehörige Projekte, Anwendungen, Werkzeuge, etc.

Es gibt heute mehrere Möglichkeiten:

- Apache Hadoop
- Quelloffene Distribution von HortonWorks, Cloudera
- Kommerzielle Variante von IBM, MapR etc.

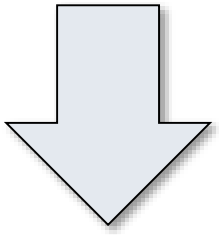
Beispiel: HortonWorks



Gliederung

- Was ist Big Data?
- Potentiale von Big Data
- Herausforderungen von Big Data

Bekannte Szenarien im Web



Google

facebook

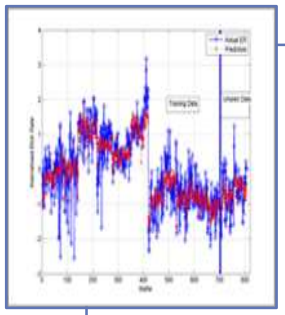
YAHOO!

twitter

PayPal

Potentiale von (Big) Data in produzierenden Unternehmen

Kann man anhand von Sensordaten einer Maschine ein typisches Muster im zeitlichen Verlauf erkennen, um Ausfälle vorhersagen zu können?



Kann man aufgrund von Stichprobenmessungen, Materialzusammensetzung etc. Qualitäts-eigenschaften für Produkte vorhersagen?

Simulation von Produktionsprozessen und Umstellung der Produktion per Knopfdruck. Auch Adaptive Produktionskonzepte.

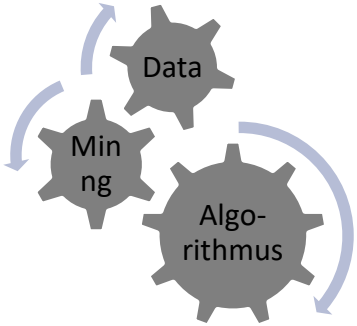
Algorithmen können Muster in Daten erkennen

Data Mining ist der "… non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data …" (Fayyad et al. 1996)

Extraktion von

- **impliziten,**
- **bislang unbekannten,**
- **potenziell nützlichen**

Mustern aus Daten.



Big Data goes Smart Data

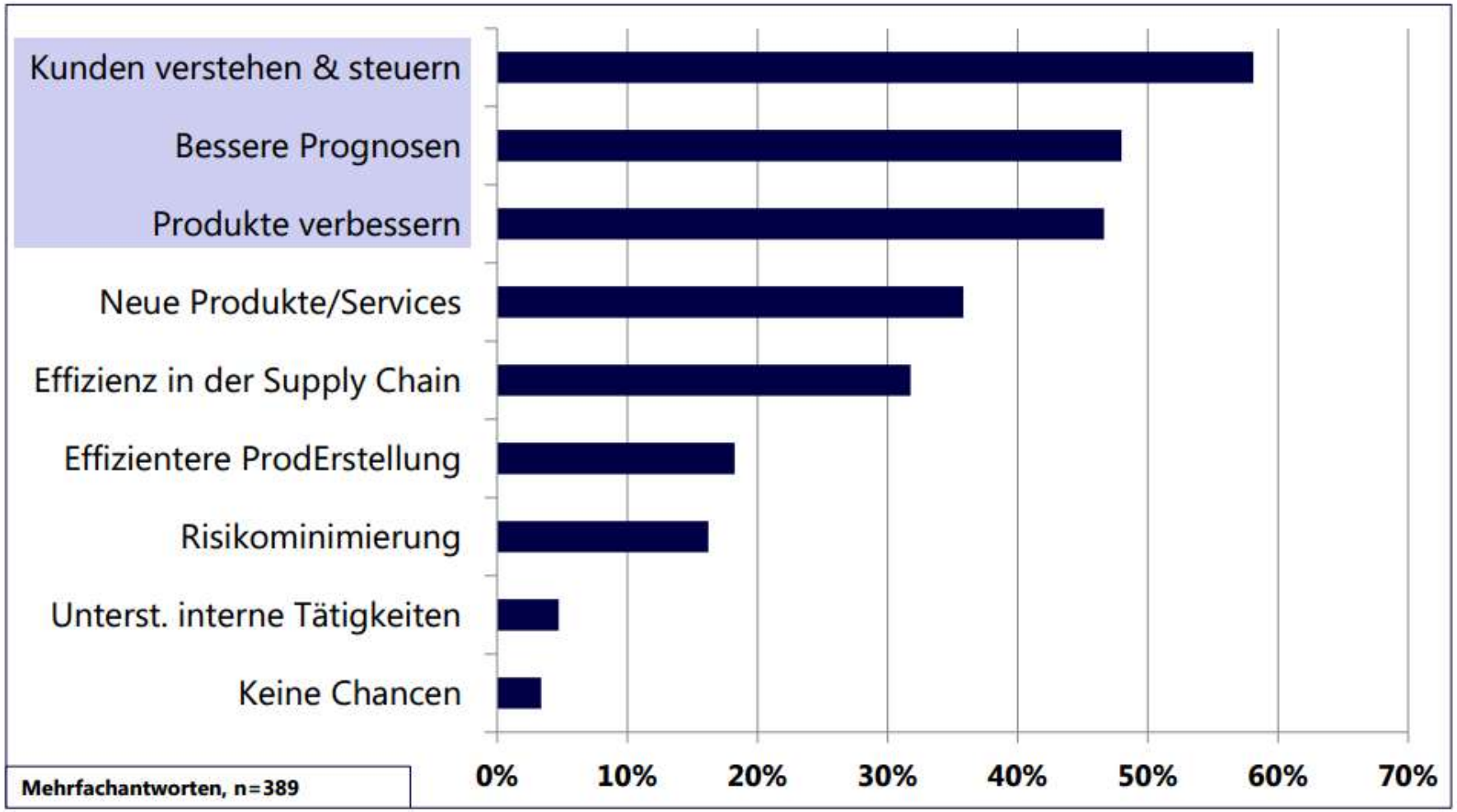
Größe allein ist nicht entscheidend!

Mehrwert entsteht durch intelligente Kombination von

1. **Daten (Volume, Velocity, Variety, Veracity)**
2. **Data Mining und**
3. **Dienstleistungen**

zu neuen Gesamtsystemen.

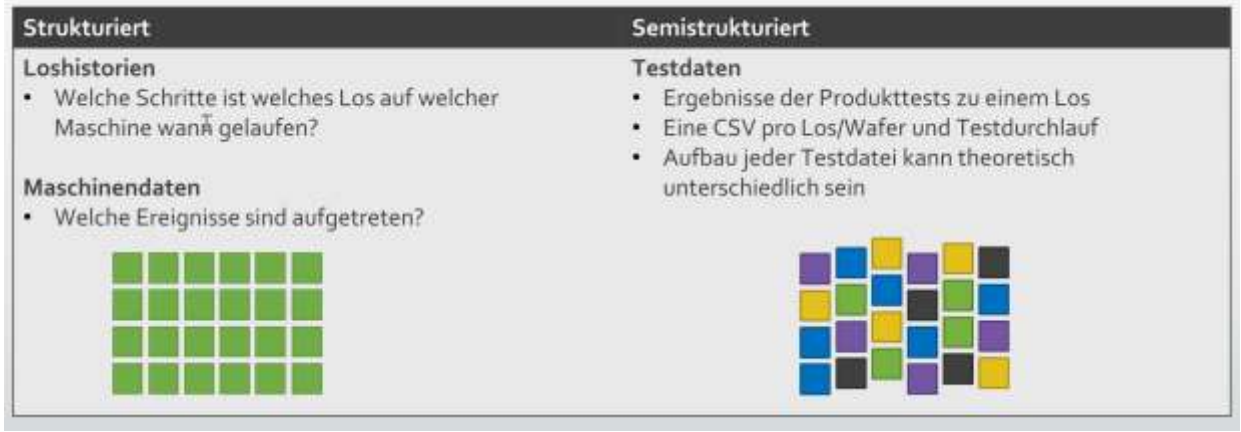
Welche Big Data Anwendungsszenarien sind für Ihr Unternehmen wirtschaftlich am sinnvollsten?



Quelle. MHP Studie „BIG DATA Future – Chancen und Herausforderungen für die deutsche Industrie“, 2014

Projektbeispiel von Seamless Analytics

- Mit Sensoren bestückte Maschinen produzieren große Datenmengen
- Loshistorien, Ereignisse der Maschinen, Ergebnisse von Qualitätsprüfungen
- Problemstellung
  - Daten werden teils automatisiert, teils manuell ermittelt
  - Hohe zeitliche Belastung der Mitarbeiter
  - Zeitnahe Auswertungen nur eingeschränkt möglich
  - Bisheriges System kommt an Kapazitätsgrenzen
  - Oracle Datenbank mit 13 TB Testdaten

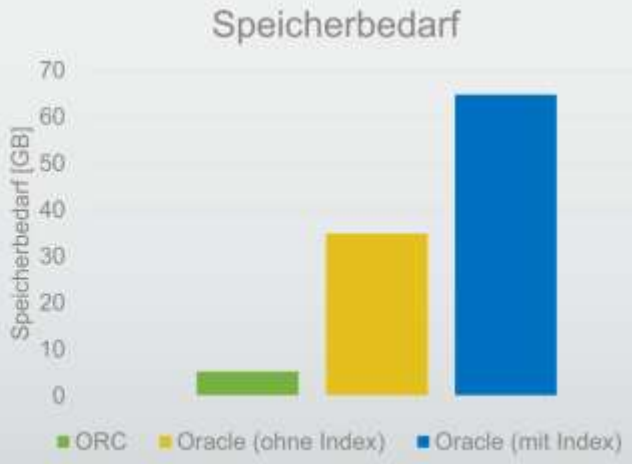


Projektbeispiel von Seamless Analytics

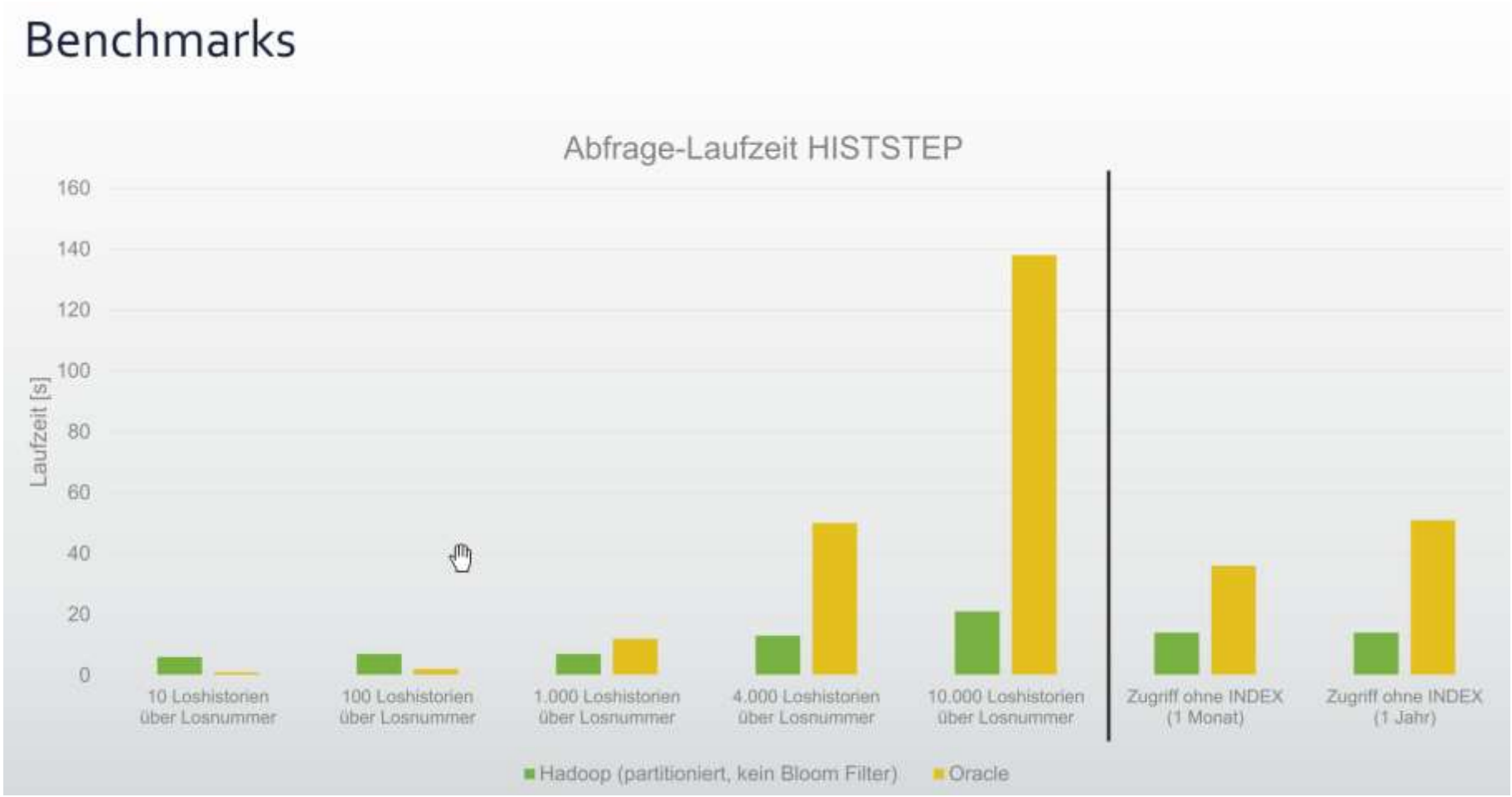
Benchmarks

Tabelle HISTSTEP

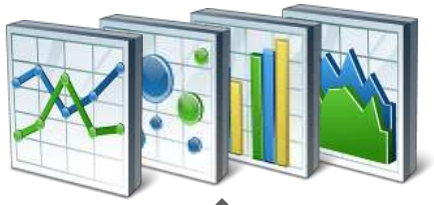
- Enthält die durchlaufenen Arbeitsschritte von Losen
- ~275 Mio. Zeilen
- 17 Spalten
  - STRING, VARCHAR(1-20)
- Speicherbedarf
  - Oracle: 34,9 GB (+29,9 GB Index)
  - ORC: 5,2 GB (partitioniert, ohne Bloom Filter Columns)



Projektbeispiel von Seamless Analytics



Projektbeispiel: Web.Intelligence bei 1&1



Die spaltenorientierte Datenbank stieß an ihre Grenzen:

Web.
Intelligence
BI Platform

- Verarbeitungsgeschwindigkeit nicht mehr ausreichend
- Aufrüstung teuer
- Begrenzte Ressourcen

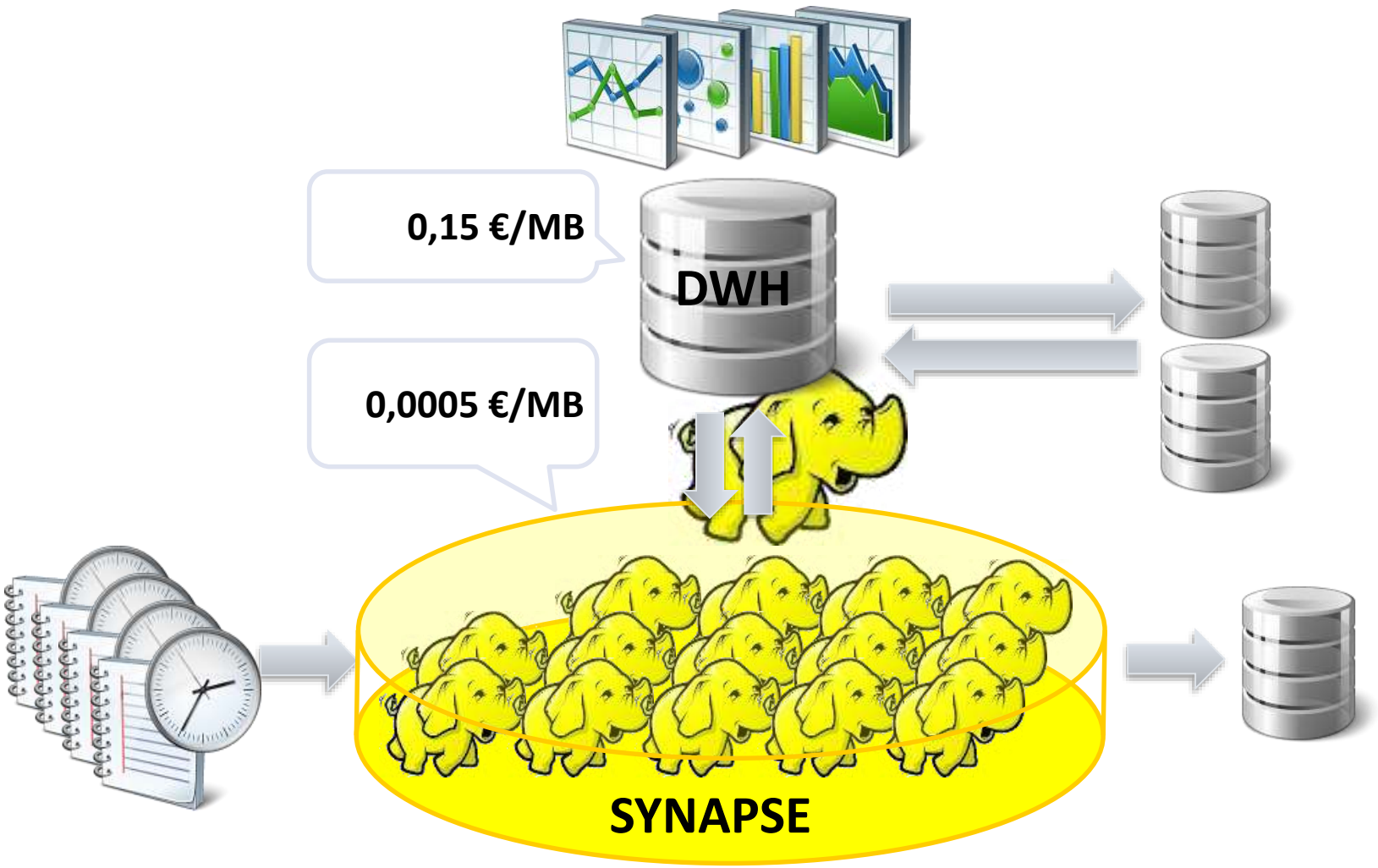
WEB.DE

1&1

GMX

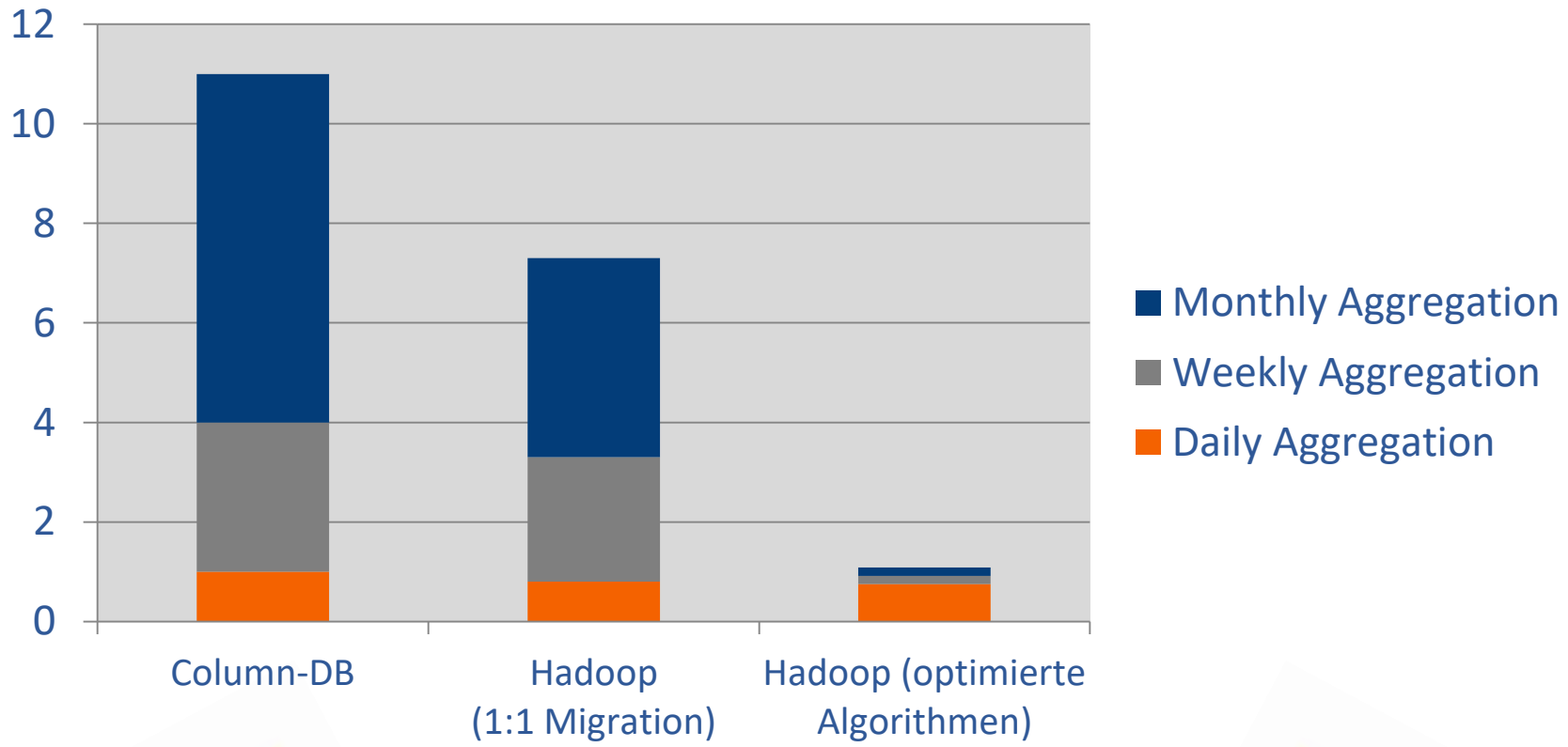
Web-Analytics: 240 Files/d
200 GB/d * 90d = 18 TB
Log-Analytics: 15.000 Files/d
2.000 GB/d * 30d = 60 TB

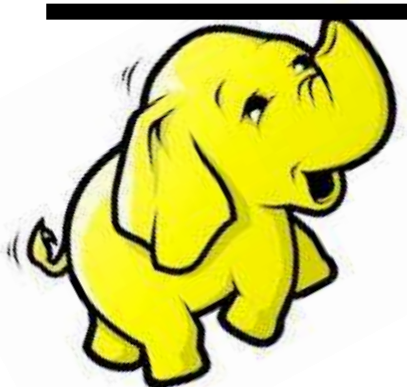
Hadoop-Cluster ermöglicht kostengünstige und skalierbare Speicherung



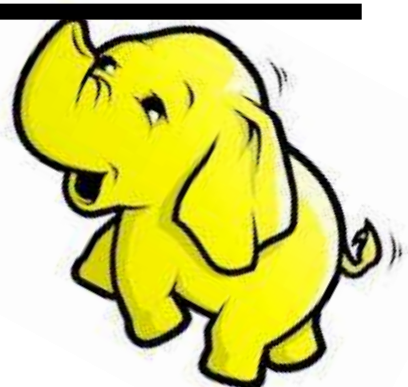
* **SYN**ergetic **A**nalytical **P**rocessing and **S**torage **E**ngine

Performanceverbesserung um bis zu Faktor 40 erreicht





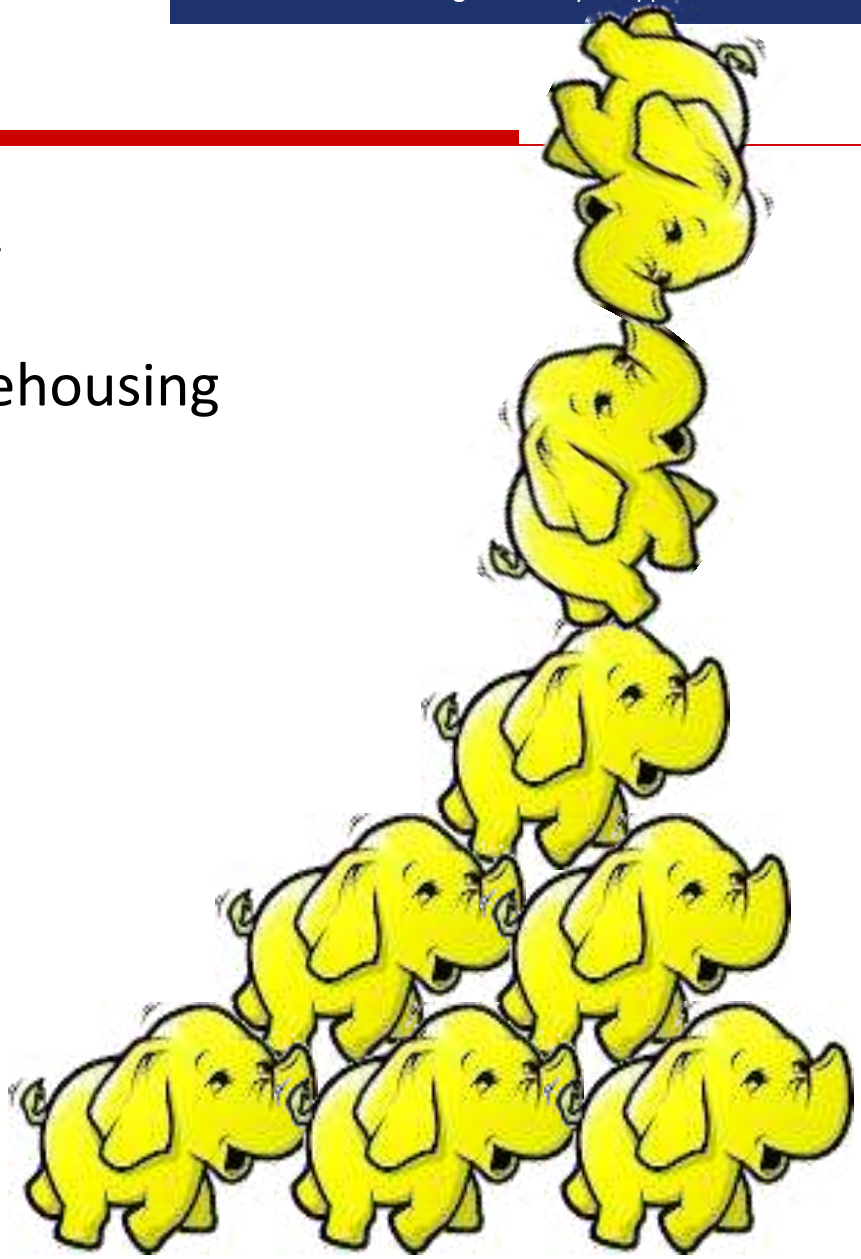
Schlüssel liegt in der Anwendung des MapReduce-Paradigmas



Das Fazit der 1&1: Hadoop beeindruckt

Massendatenverarbeitung bei 1&1
ist für Web- und Media-Analytics,
Logfile-Verarbeitung und Datawarehousing
mit Hadoop messbar

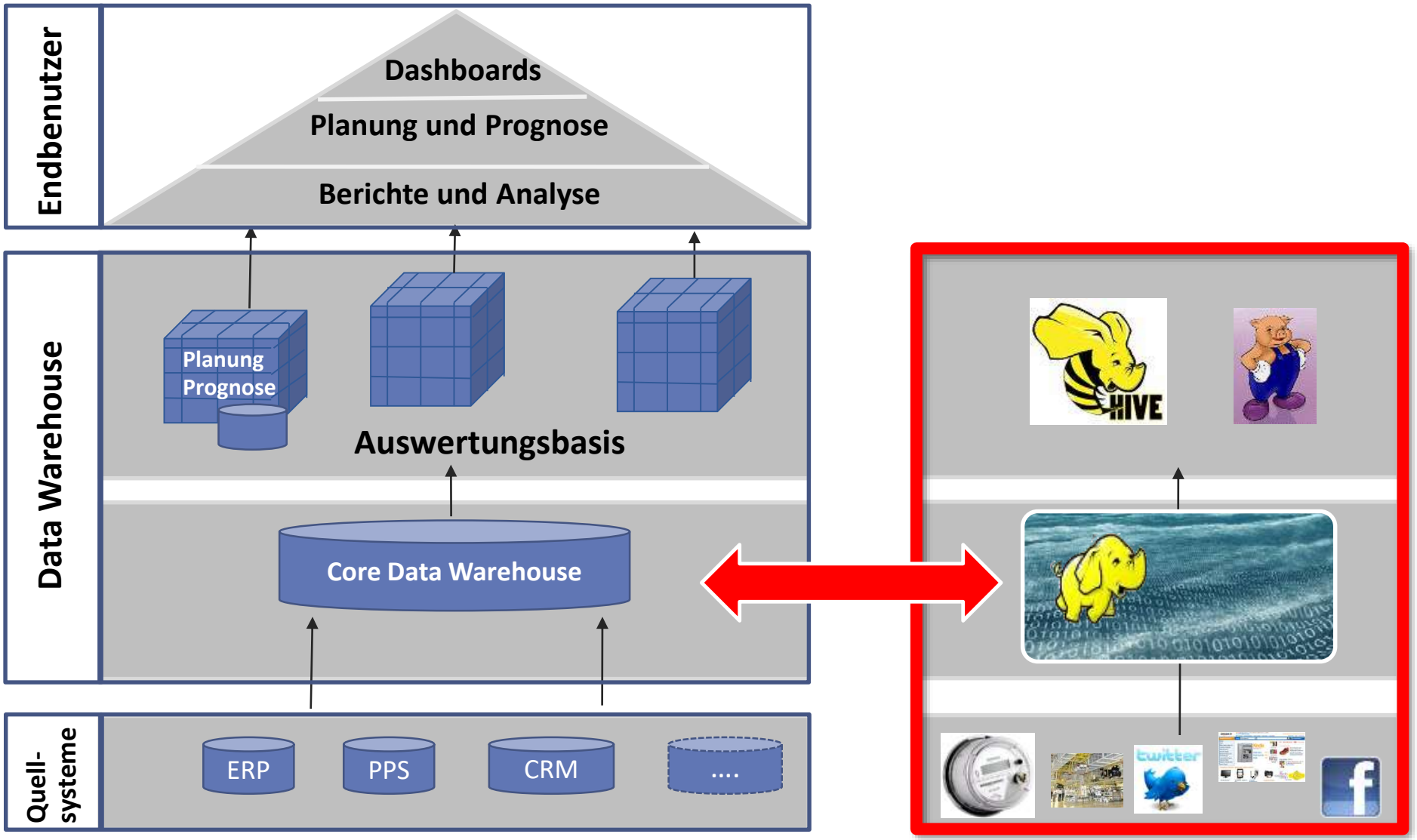
- performanter,
- kostengünstiger,
- skalierbarer,
- flexibler,
- und zukunftsfähiger.



Gliederung

- Was ist Big Data?
- Potentiale von Big Data
- Herausforderungen von Big Data

Zusammenspiel Business Intelligence und Big Data?



Herausforderungen für Big Data

- Big Data ist nicht ein Thema der Größe.
  Herausforderungen sind Heterogenität und Datenqualität.
- Big Data Projekte sind erfolgreich, wenn Analyse an Unternehmensprozesse und Produkte gekoppelt ist.

- Echtzeitverarbeitung von großen Datenströmen
- Neue Sicherheitskonzepte für Anbindung Shop-Floor-Systeme und Embedded Systems

- Datenschutz und ethische Fragestellungen müssen technisch, algorithmisch, rechtlich und gesellschaftlich betrachtet werden.
- Know-How in Data Analytics und Unternehmensprozessen

„Wir sind ein Softwareunternehmen genauso, wie wir ein Hardwareunternehmen sind.“

Tesla-Firmenchef Elon Musk