

# Fujitsu BS2000/OSD Mainframe Summit 2013

## Mystifying Big Data

**Dr. Fritz Schinkel**  
Program Manager Solutions and Innovations

## ■ Das Big Data Phänomen

- Exponentielles Datenwachstum
- Definierende Eigenschaften für Big Data
- Einschätzungen zu Big Data

## ■ Big Data Beispiele

- Gewinnung von Marketing Informationen
- Persönliche Daten: Quantified Self bis Patients Like Me
- Crowd Sourcing: Wikipedia und Kaggle

## ■ Big Data Infrastruktur

- Technologien
- Referenzarchitektur für Big Data
- Big Data Angebot von Fujitsu

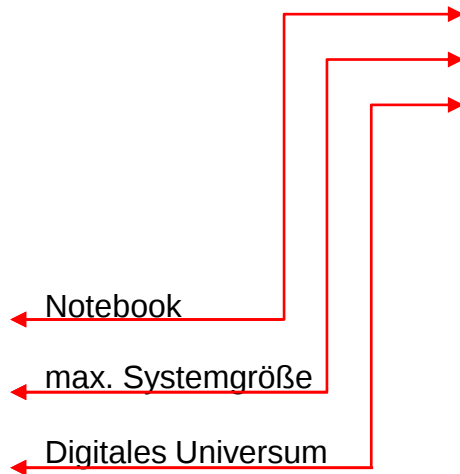
# Das Big Data Phänomen

# Exponentielles Datenwachstum

- 65% jährliches Wachstum / Verdopplung alle 18 Monate
- 94% der Daten stammen aus den letzten 5 Jahren
- 2.7 Zettabyte in 2012

| Dezimalpräfixe                 |                                                         |
|--------------------------------|---------------------------------------------------------|
| Name<br>(Symbol)               | Bedeutung <sup>[G 1]</sup>                              |
| Kilobyte (kB) <sup>[G 2]</sup> | $10^3$ Byte = 1.000 Byte                                |
| Megabyte (MB)                  | $10^6$ Byte = 1.000.000 Byte                            |
| Gigabyte (GB)                  | $10^9$ Byte = 1.000.000.000 Byte                        |
| Terabyte (TB)                  | $10^{12}$ Byte = 1.000.000.000.000 Byte                 |
| Petabyte (PB)                  | $10^{15}$ Byte = 1.000.000.000.000.000 Byte             |
| Exabyte (EB)                   | $10^{18}$ Byte = 1.000.000.000.000.000.000 Byte         |
| Zettabyte (ZB)                 | $10^{21}$ Byte = 1.000.000.000.000.000.000.000 Byte     |
| Yottabyte (YB)                 | $10^{24}$ Byte = 1.000.000.000.000.000.000.000.000 Byte |

Quelle: Wikipedia



| Zahl      | Deutsch           | Zeichen | Lesung     |
|-----------|-------------------|---------|------------|
| $10^1$    | Zehn              | 十       | jū         |
| $10^2$    | Hundert           | 百       | hyaku      |
| $10^3$    | Tausend           | 千       | sen        |
| $10^4$    | Zehntausend       | 万       | man        |
| $10^8$    | 100 Millionen     | 億       | oku        |
| $10^{12}$ | 1 Billion         | 兆       | chō        |
| $10^{16}$ | 10 Billiarden     | 京       | kei, kyō   |
| $10^{20}$ | 100 Trillionen    | 垓       | gai        |
| $10^{24}$ | 1 Quadrillion     | 秭       | shi, jo    |
| $10^{28}$ | 10 Quadrilliarden | 穰       | jō         |
| $10^{32}$ | 100 Quintillionen | 溝       | kō         |
| $10^{36}$ | 1 Sextillion      | 澗       | kan        |
| $10^{40}$ | 10 Sextilliarden  | 正       | sei        |
| $10^{44}$ | 100 Septillionen  | 載       | sai        |
| $10^{48}$ | 1 Oktillion       | 極       | goku       |
| $10^{52}$ | 10 Oktilliarden   | 恒河沙     | gōgasha    |
| $10^{56}$ | 100 Nonillionen   | 阿僧祇     | asōgi      |
| $10^{60}$ | 1 Dezillion       | 那由他     | nayuta     |
| $10^{64}$ | 10 Dezilliarden   | 不可思議    | fukashigi  |
| $10^{68}$ | 100 Undezillionen | 無量大数    | muryōtaisū |

# In Daten Informationen finden, um...

- ... Trends und Chancen vorherzusagen?
- ... die richtigen Entscheidungen zu treffen?
- ... Entscheidungen zu beschleunigen?
- ... automatische Reaktionen auszulösen?
- ... den Kosten auf den Grund zu gehen?
- ... überflüssige Aktivitäten zu vermeiden?
- ... Risiken einzuschätzen und zu minimieren?
- ... zu wissen, was wir nicht wissen?
- ... ..



**Können diese Träume Wirklichkeit werden?**

# Von Business Intelligence zu Big Data

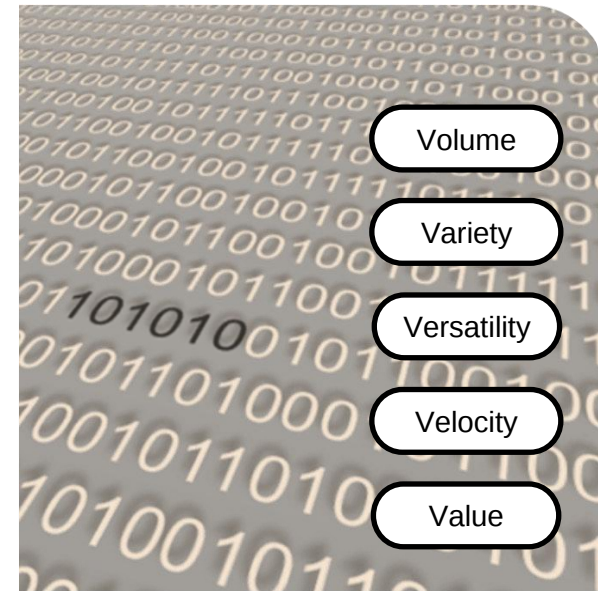
## BI bisher

- Interne Daten
- Strukturiert / relational
- Wenige Quellen
- GB und TB
- Berichtswesen
- Risikovermeidung
- Periodisch
- Batch
- Statisches Datenmodell
- Wenige direkte Benutzer
- Im eigenen Rechenzentrum

## Heutige Anforderungen

- Interne und externe Daten
- Un- / semi- / poly- / strukturiert
- Viele Quellen
- TB und PB
- Vorhersagen
- Chancen erkennen
- Ad-hoc
- Echtzeit (Analyse → Aktion)
- Versuch und Änderung
- Viele Benutzer, Mill. Events / sec
- Überall, von jedem Endgerät

## Big Data



**Erschwingliche Technologien zur schnellen Erfassung, Speicherung und Analyse**

# Einschätzungen zu Big Data

## The Problem With Big Data Is That Nobody Understands It

*The United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings.*

[McKinsey Study 2011]

Björn Bloching, Lars Luck, Thomas Ramge (2012):

## Data unser (In data we trust)

*Die Revolution der Kundendaten geht in die nächste Runde. ... Die Ära der Intuition ist vorbei, Daten sind der Kitt in der Kundenbeziehung*

**Klar:** Daten entstehen in großen Mengen

**Erkennbar:** Techniken zu Analyse

**Offen:** Was machen wir daraus?

Rudi Klausnitzer (2013):

## Das Ende des Zufalls

*Wir und unser Leben werden immer berechenbarer!*

*Wachsam gegenüber den Gefahren, aber offen für die riesigen Chancen.*

# Big Data: Herausforderungen und Chancen

**100,000**

Flights/day

**45 million**

Servers

**1 billion**

Users

**200 million**

Pictures/day

**1 billion**

PCs

**Billions  
of requests**

per day

Image Analysis

Social Complex Event Processing

Network Data Layer Management

Scan

Sensor Data

Points of Interest

Route  
search

Area Managements

Congestions Forecast

Personal Profiles

**Billion objects**

per hour

**1 billion**

Rides per day

**100s**

of millions  
Locations

**Billions of**

measurements per day

**600 million**

Smartphones

**1 billion**

Cars

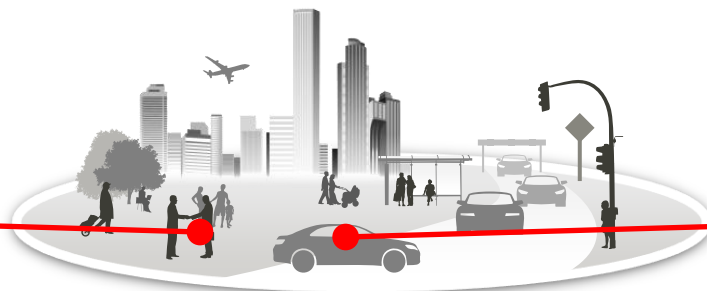


## Datenquellen

- Firmendaten, Historie
- Öffentliche Daten
- Internet-Nutzung
- Soziale Netzwerke
- Smartphone Nutzung
- Sensoren z.B. am Auto
- Quantified-self
- ...

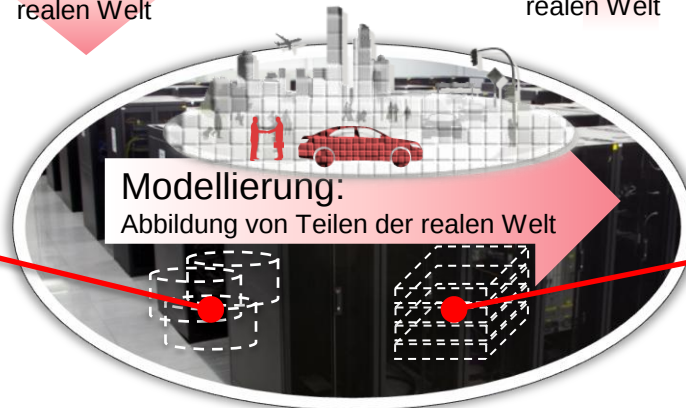
## Datenhaltung

- Private Datensammlungen
- Online / Nearline / Archiv
- Öffentliche Datendienste
- Kommerzielle Daten
- ...



**Sensorik:**  
Spur der  
realen Welt

**Feedback:**  
Aktionen in der  
realen Welt



**Modellierung:**  
Abbildung von Teilen der realen Welt

## Datennutzung

- Information
- Empfehlungen
- Marketing
- Produktoptimierung
- Entscheidungen
- Steuerung
- ...

## Datenanalyse

- Bereinigung
- Klassifikation
- Modellbildung
- Vorhersagen
- ...

# Schwangerschaftstest am Supermarktregal

## ■ Marketingziel: Junge Familie

- Familiengründung und Nachwuchs → Anschaffungen
- Wettlauf um wertvolle Marketinginformation

## ■ Wie kann ein Supermarkt eine Geburt voraussagen?

- Einkaufsbiographien durch Analyse einiger Millionen von Bons!
- Schwangerschaftsvorhersage basierend auf 25 Artikel
- Abschätzung des Geburtstermins auf wenige Tage

## ■ Wie ist die Information nutzbar?

- Die richtige Werbung für jede Phase der Schwangerschaft
- Beeinflussung des gesamten Konsumverhaltens der Zielgruppe

## ■ Gibt's das wirklich?

- US Supermarktkette Target
- Data Scientist Andrew Pole

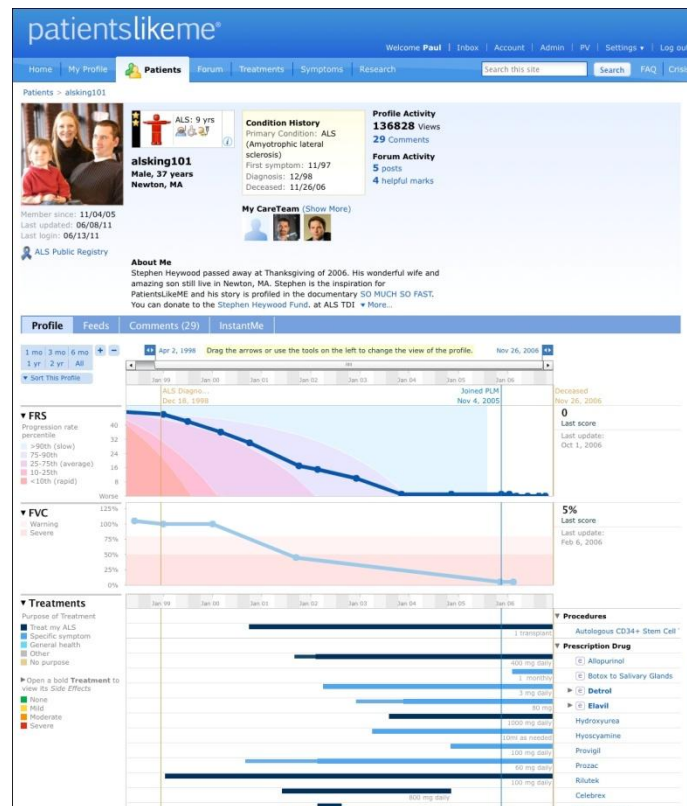


*"We'll be sending you coupons for things you want before you even know you want them."*

[Andrew Pole, NYT Magazine 2012]

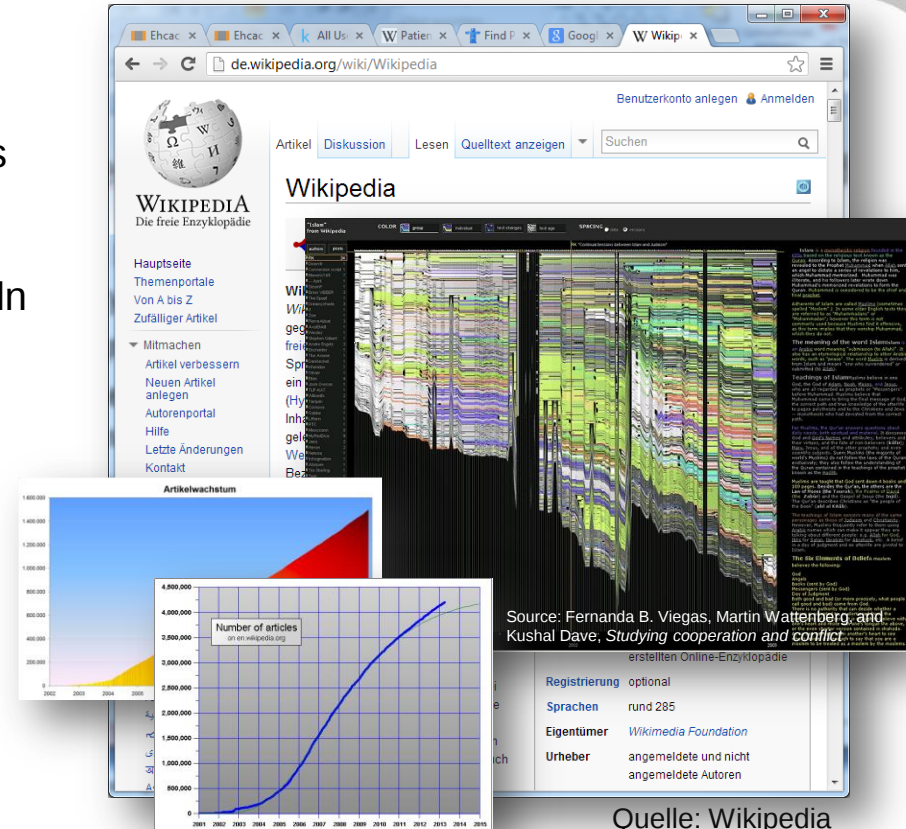
# Soziale Netzwerke: Patients Like Me

- Patienten Selbsthilfegruppen als Soziales Netzwerk im Web
- Gegründet 2004 von Jamie Heywood
- Anlass war die Amyotrophe Laterale Sklerose (ALS) Erkrankung seines Bruders Stephen Heywood
- Patienten geben Informationen über Krankheitsverlauf und Behandlungsergebnisse
- Profitieren sehr schnell von Erfahrungen anderer
- Trendvorhersagen für individuelle Therapieanpassungen und begleitend zu klinischen Studien
- Zur Zeit ca. 200.000 Mitglieder
- Profit-Organisation finanziert sich durch anonymisierte Datenweitergabe an Industrie, keine Werbung



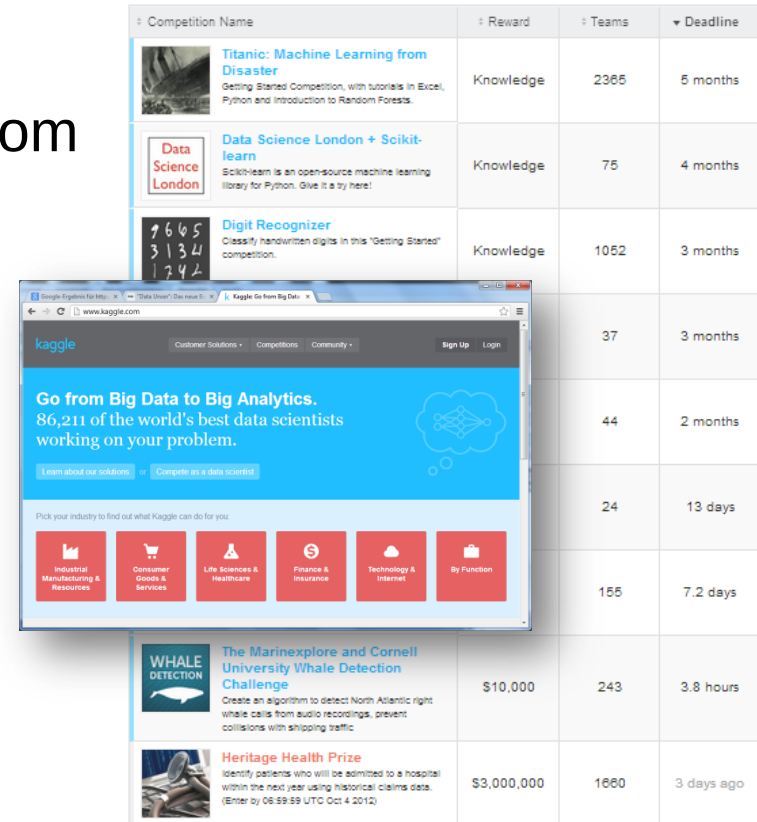
# Geteiltes und mitgeteiltes Wissen: Wikipedia

- Frei lizenzierte hochwertige Enzyklopädie
- Gegründet 2001 von Jimmy Wales
- Freier Zugang und geregelte Verwendbarkeit des Wissens
- Erstellung, Diskussion und Review der Artikel durch freie Autoren basierend auf wenigen Regeln
- Zur Zeit ca. 1,5 Millionen angemeldete Benutzer
- Aktuell 286 Sprachen, 4,2 Millionen englischsprachige und 1,5 Millionen deutschsprachige Artikel
- 25.000 - 60.000 Zugriffe pro Sekunde, > 400 Server
- Non Profit-Organisation, finanziert durch Spenden, keine Werbung



# Crowd Sourcing Data Science: kaggle

- Big Data Analytics als Wettbewerb
- Gegründet 2010 von Anthony Goldbloom
- Webportal verknüpft Aufgabensteller / Sponsoren mit Teilnehmern
- Zur Zeit über 85.000 registrierte Benutzer
- Bisläng 70 Wettbewerbe um Wissen, Jobs und Preisgeld
- Höchster bisläng ausgelobter Preis: Heritage Health Prize, 3.000.000 \$

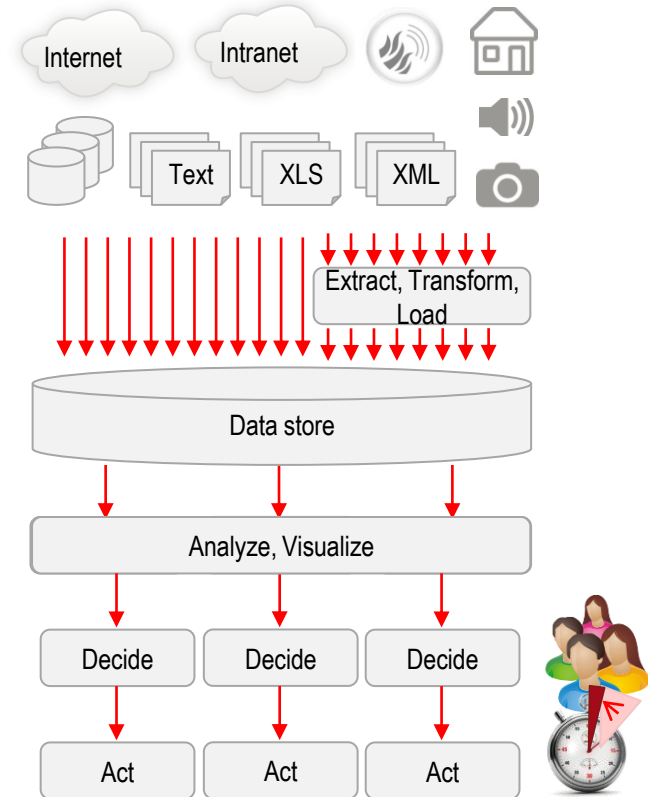
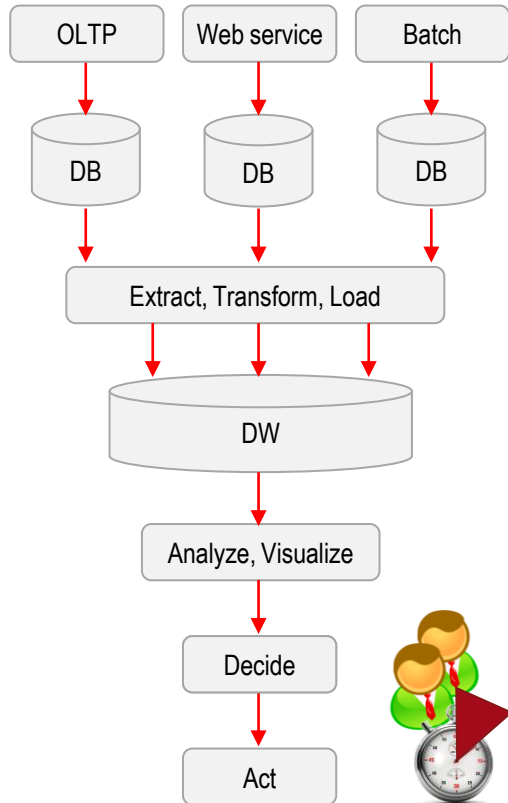


The image shows a screenshot of the Kaggle website. The main part of the image is a table listing various competitions. The table has columns for Competition Name, Reward, Teams, and Deadline. The competitions listed are 'Titanic: Machine Learning from Disaster', 'Data Science London + Scikit-learn', 'Digit Recognizer', 'Whale Detection', and 'Heritage Health Prize'. A smaller inset image shows a detailed view of the 'Titanic' competition page, which includes a description of the challenge and a 'Get Started' button.

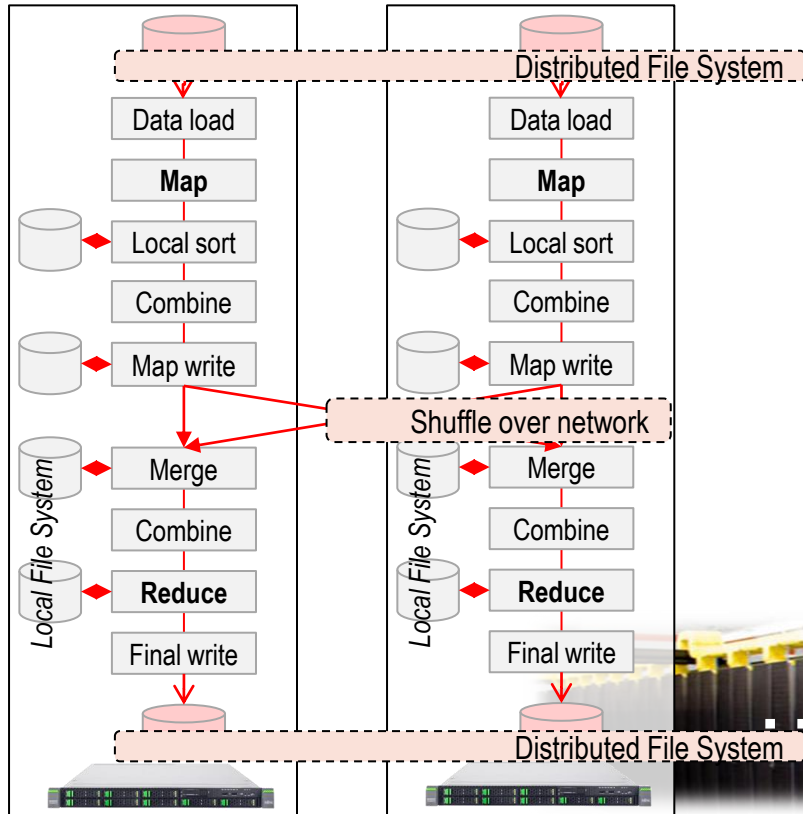
| Competition Name                                                                                                                                                                                                                      | Reward      | Teams | Deadline   |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|-------|------------|
| <b>Titanic: Machine Learning from Disaster</b><br>Getting Started Competition, with tutorials in Excel, Python and introduction to Random Forests.                                                                                    | Knowledge   | 2365  | 5 months   |
| <b>Data Science London + Scikit-learn</b><br>Scikit-learn is an open-source machine learning library for Python. Give it a try here!                                                                                                  | Knowledge   | 75    | 4 months   |
| <b>Digit Recognizer</b><br>Classify handwritten digits in this 'Getting Started' competition.                                                                                                                                         | Knowledge   | 1052  | 3 months   |
|                                                                                                                                                                                                                                       |             | 37    | 3 months   |
|                                                                                                                                                                                                                                       |             | 44    | 2 months   |
|                                                                                                                                                                                                                                       |             | 24    | 13 days    |
|                                                                                                                                                                                                                                       |             | 155   | 7.2 days   |
| <b>WHALE DETECTION</b><br><b>The Marinexplore and Cornell University Whale Detection Challenge</b><br>Create an algorithm to detect North Atlantic right whale calls from audio recordings, prevent collisions with shipping traffic. | \$10,000    | 243   | 3.8 hours  |
| <b>Heritage Health Prize</b><br>Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)                                                            | \$3,000,000 | 1660  | 3 days ago |

# Big Data Infrastruktur

# Vom Data Warehouse zu BigData



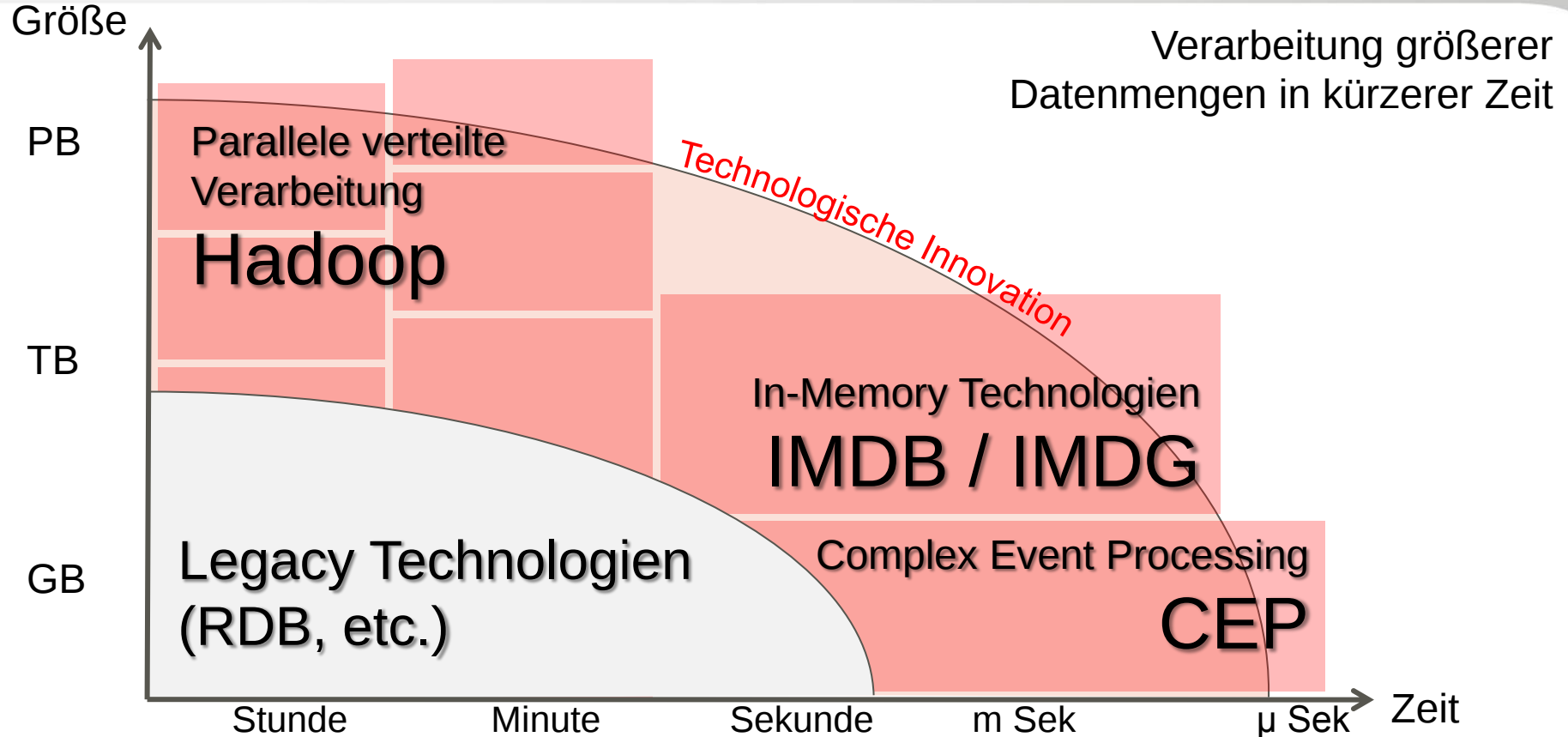
# Am Anfang war MapReduce



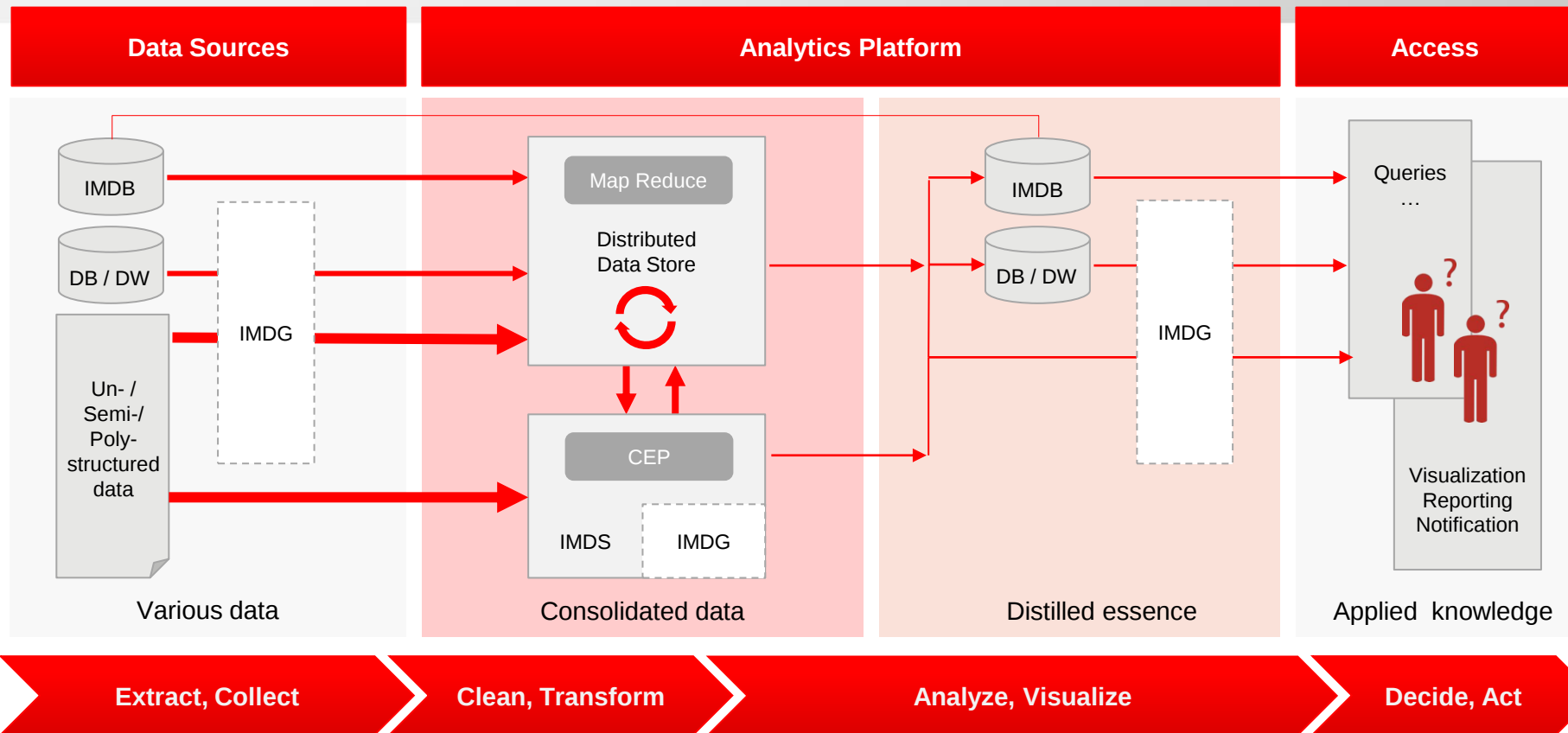
- MapReduce entstand als “Shared-Nothing” Analyse Framework bei Google.
- Aus Google Veröffentlichungen entstand MapReduce als Kern für Apache Hadoop.
- Hadoop ist bei Facebook und Yahoo! im Einsatz.
- Hadoop findet Eingang in Industrielösungen.



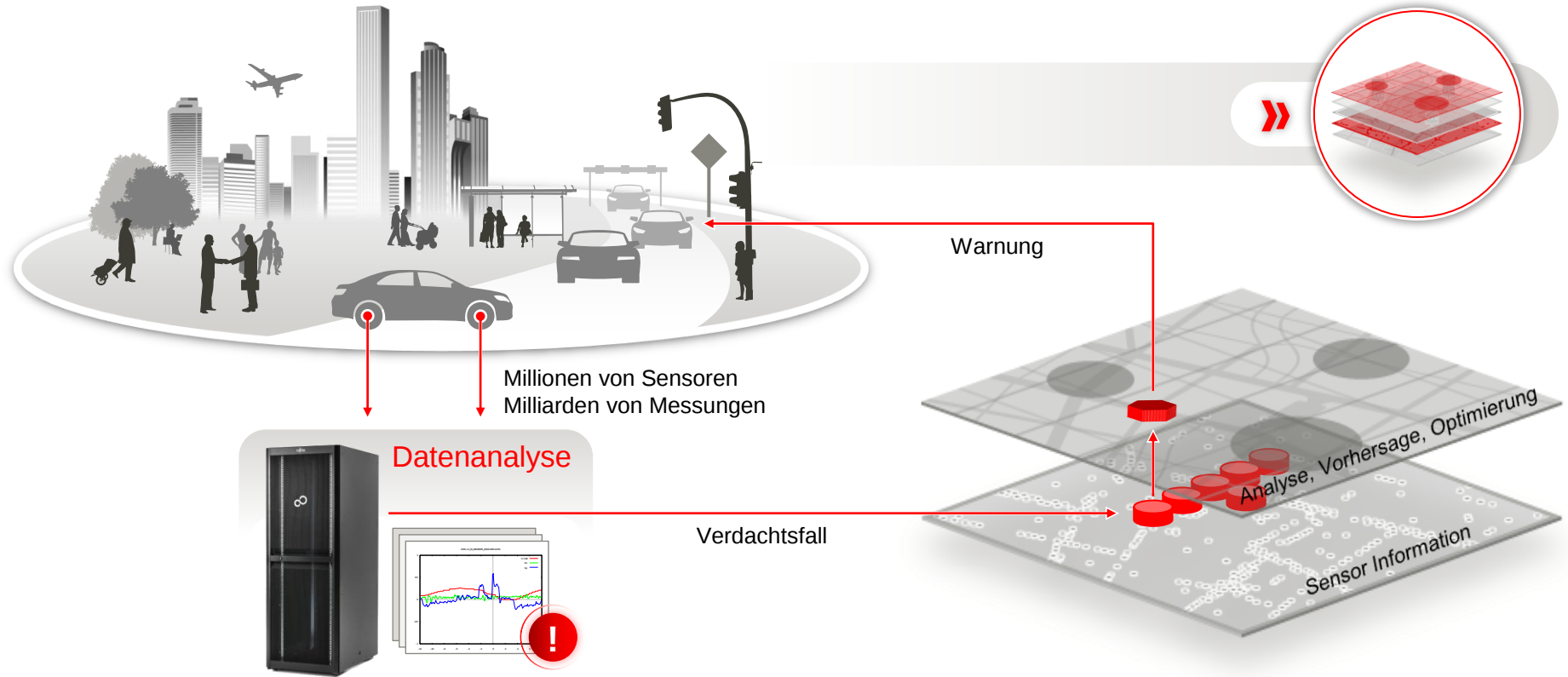
# Technologien für Big Data



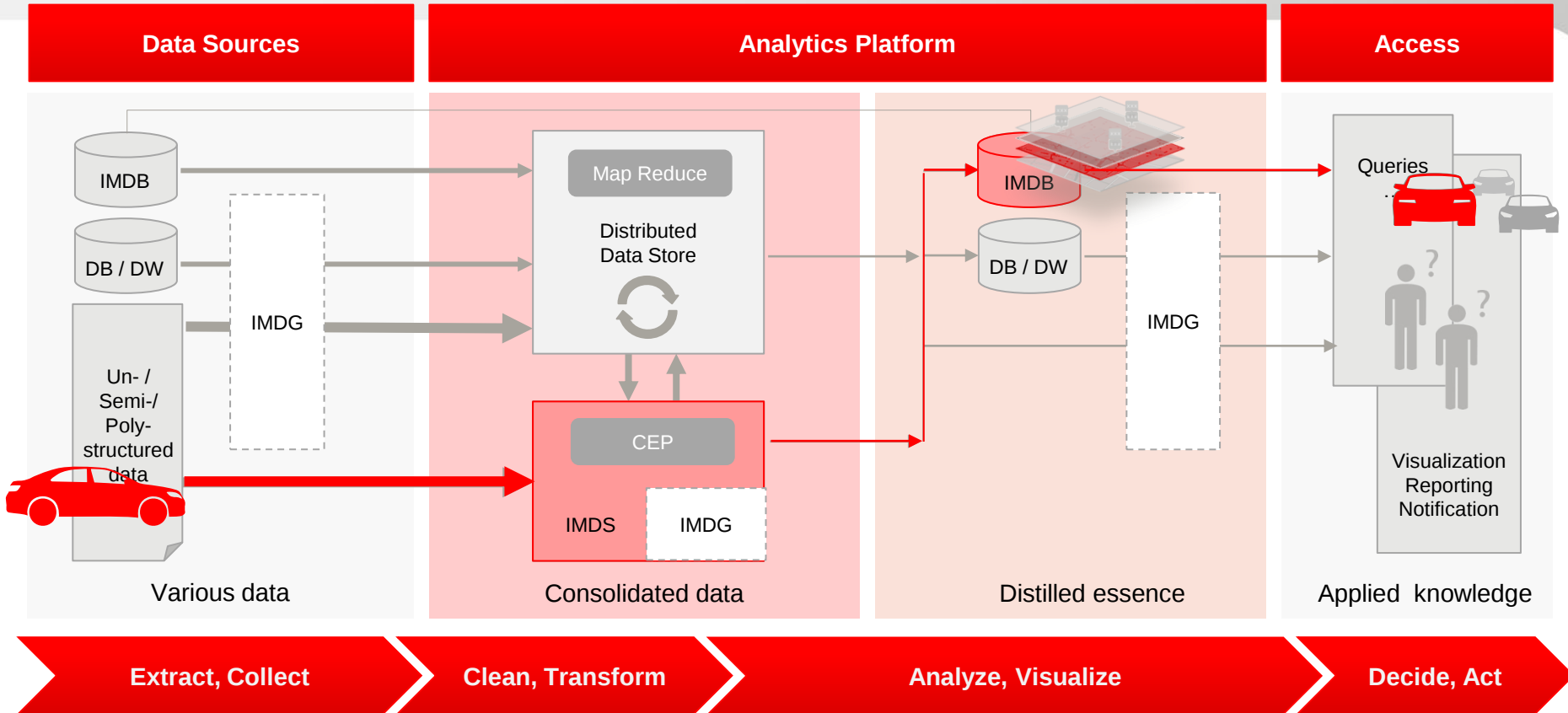
# Referenzarchitektur für Big Data



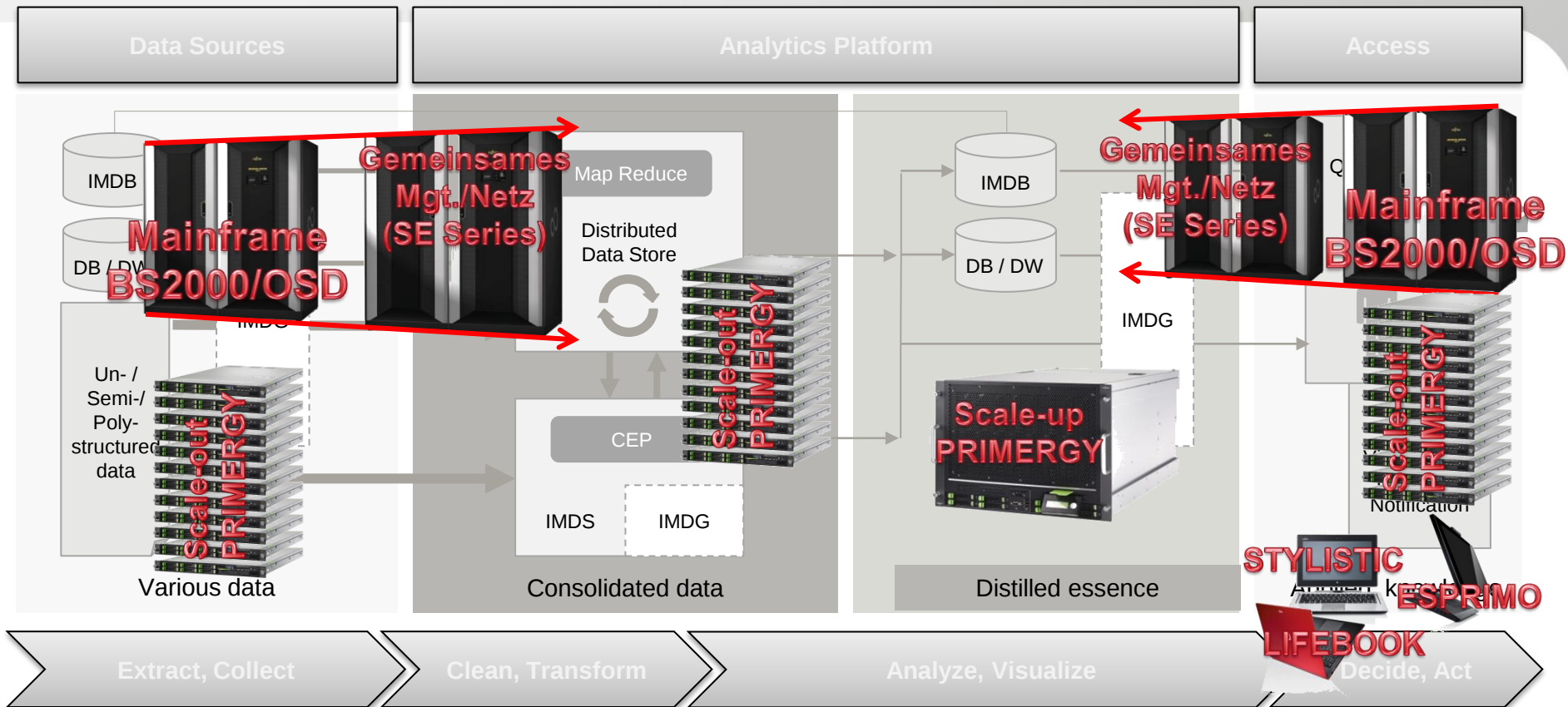
# Beispiel: Warnung vor Ölspur



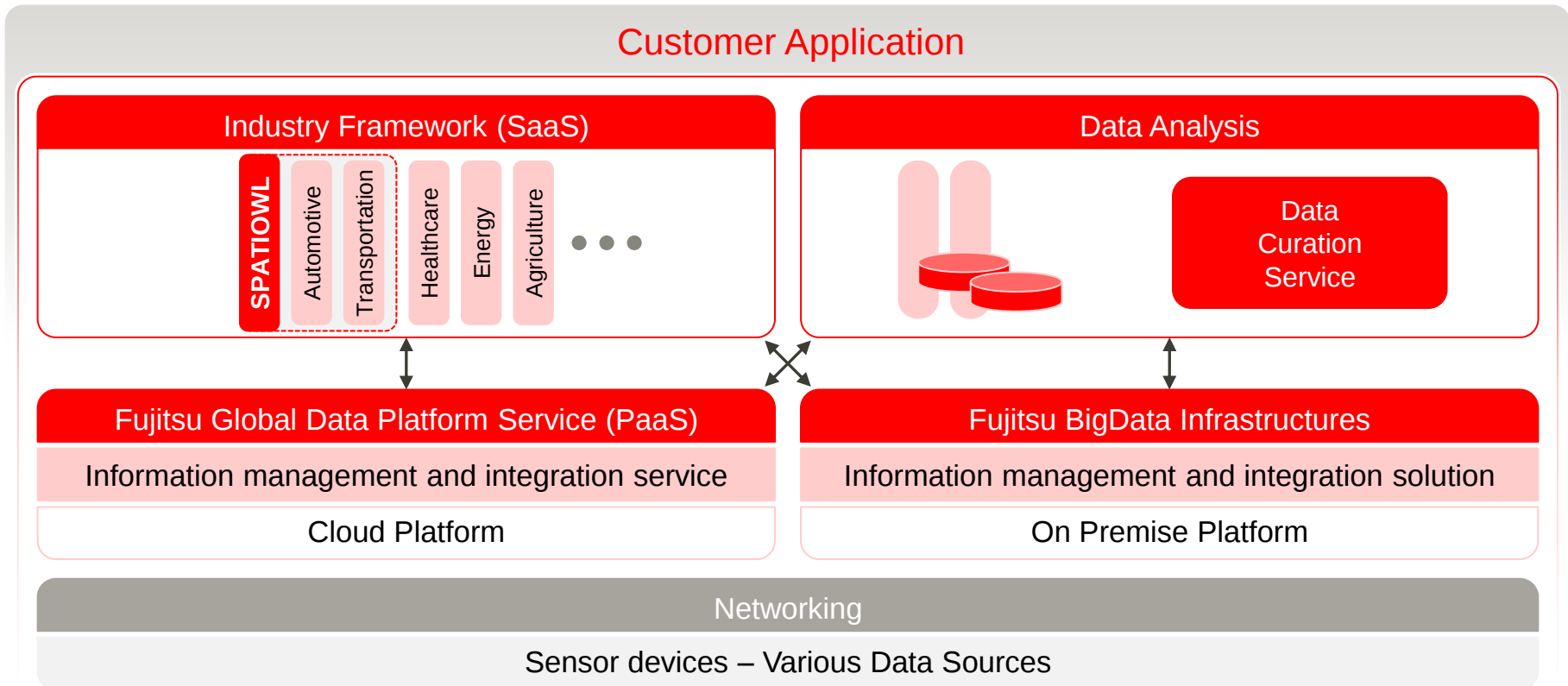
# Von der Messung zur Meldung



# Infrastruktur für Big Data



# Big Data Angebot von Fujitsu im Überblick



- Wachstum und Wachstumsgeschwindigkeit der Daten ist Realität
- Daten enthalten versteckte Informationen und Wissen
- Daten sind Grundlage neuer Geschäftsmodelle
- Neue technische Möglichkeiten der Datenhaltung und -verarbeitung
- Mainframes bleiben die Plattform für hohen Datendurchsatz
  
- Noch nicht alles klar,  
noch eher “mystifying” als “demystifying” und ...

# Visualisierung von Big Data

## ■ Google books

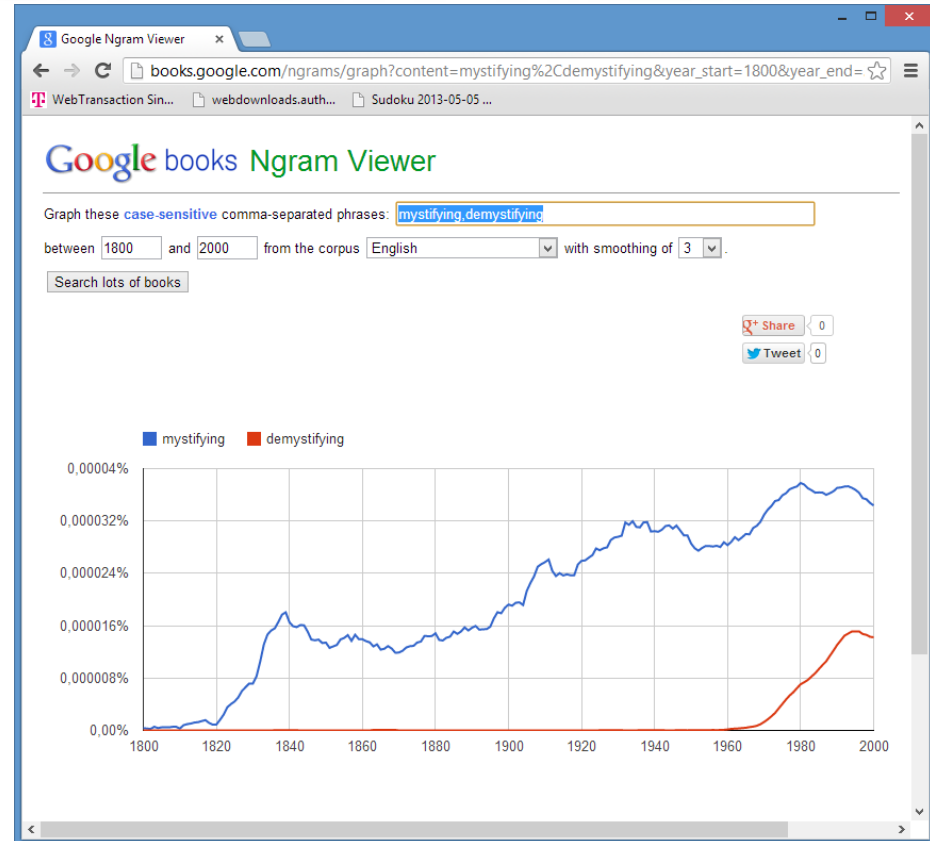
- Digitalisierung analoger Buchbestände
- 15 Millionen Bücher bis 2015

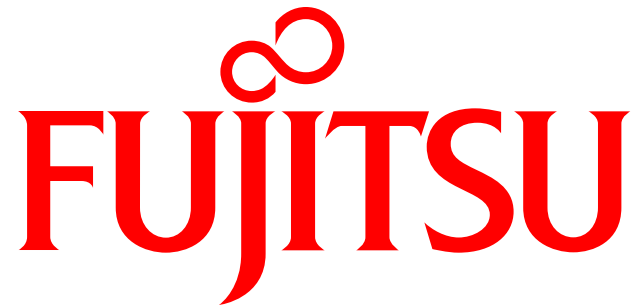
## ■ Auswertung von n-Grammen

- n-Gramm hier Phrase von n Wörtern
- Relative Häufigkeit über die Zeit
- Forschungsarbeit von Erez Lieberman Aiden und Jean-Baptiste Michel

## ■ Verfügbar als Web-Service

- <http://books.google.com/ngrams>





shaping tomorrow with you