# Big Data – Big Mess?

## Ein Versuch einer Positionierung…

Autor: Daniel Liebhart (Peter Welkenbach)

Datum: 10. Oktober 2012

Ort: DBTA Workshop on Big Data, Cloud Data Management and NoSQL

# Agenda

- Some Definitions
- Big Data – Big Business?
- Some technical Remarks
- Business Cases
- Ready to go?

# Definitions: Big Data

- **Big Data:** Volume of Data (Terabyte – Petabyte Range)
- **Fast Data:** Near Real-Time Access
- **All Data:** Structured – Semi Structured - Unstructured
- **IDC:** „The bringing together of a vast amount of data from public and private sources, combined with the intuition of business and thought leaders and the speed and affordability of today's computers, is what Big Data is all about“
- **Gartner:** “The term "big data" puts an inordinate focus on the issue of information volume (in every aspect from storage through transform/transport to analysis). Big data is also heavily weighted toward current issues and can lead to short-sighted decisions that will hamper the enterprise's information architecture as IT leaders try to expand and change it to meet changing business needs.”

# Definitions: NoSQL - NewSQL

- **NoSQL:** Means Not only SQL
- Consists of different Non-Relational Storage & Data Access Technologies like GraphDB, Big Tables, Key Value, Document, Grid and Distributed Cache
- **Examples:** Neo4J (GraphDB), HBase (Big Tables), Oracle NoSQL (Key Value), MongoDB (Document), Oracle Coherence (Grid), memcached (Distributed Cache)
- **NewSQL:** New Breed of RDBMS
- Different that existing RDDMS Technolgies, optimized for fast & distributed access –
- „Scale Out Architectures“ – usually In-Memory Technologies
- **Examples:** Oracle TimesTen, VoltDB

# Big Data – Big Business - Drivers?

- **Future Issues drives Big Data:**
- **Private / Business:** No Separation any more (60% employees wishes to work at home)
- **Tablets / Mobile:** More and More (5 billion mobile phones today)
- **Social Networks:** Enterprise Agents
- **Social Networks:** Geo-Information

- **Analysts Expectations for 2015:** 25% more productivity – 60% more IT-spending / employee

- **Data Storage Issue:** Data Analysis on the base of replication (DWH) is not anymore possible (the amount of data is too big)

# Big Data – Big Business?

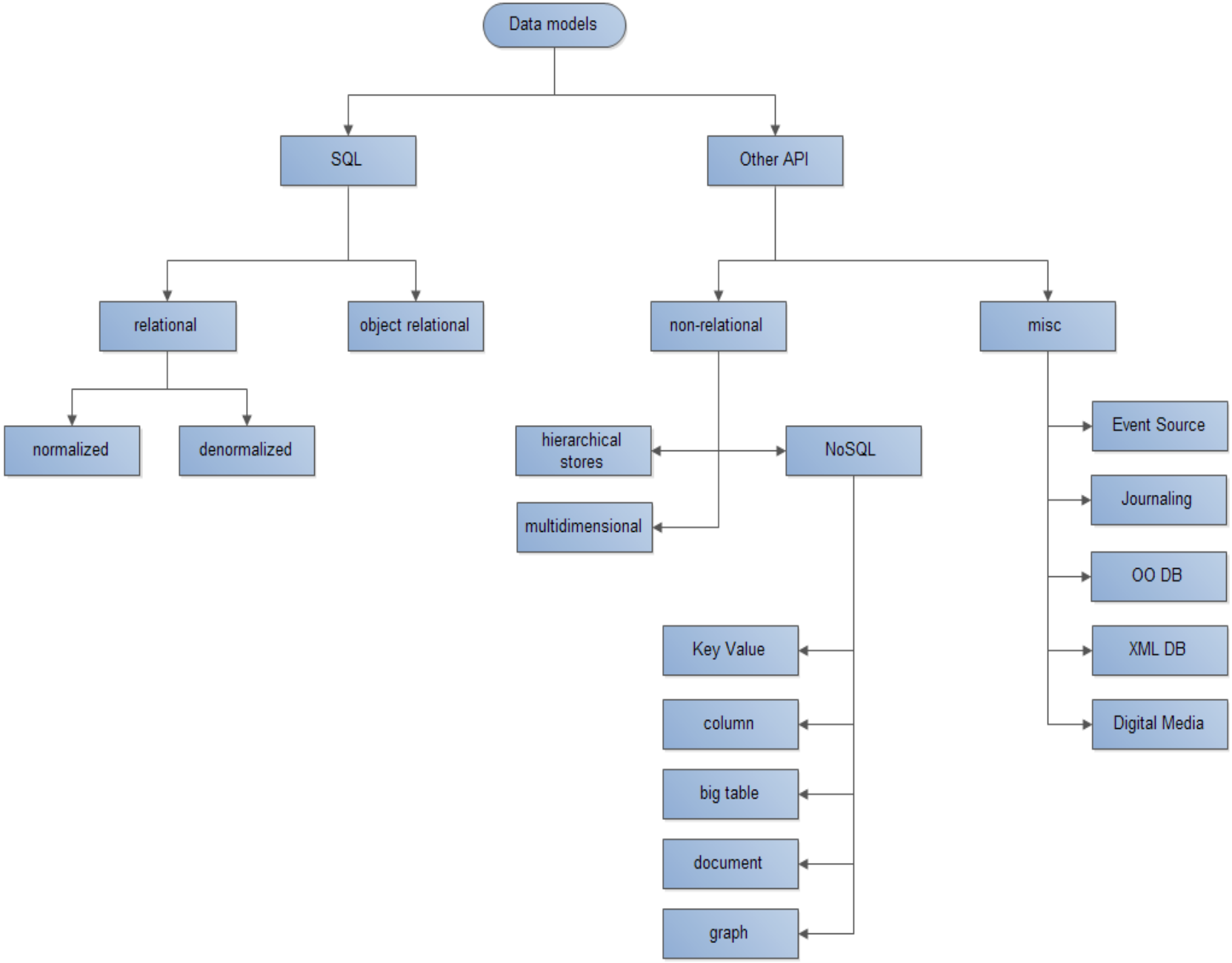
- **Forbes:** Big Data Is Big Market & Big Business - $50 Billion Market by 2017 ($5 Billion 2012 ;-))
- **IDC:** Big Data as a Part of Future „Third Platform“:
  - 1. Internal IT-Services
  - 2. External SaaS Services (Cloud – Mobile - „Internet of Things“)
  - 3. Big Data Processing

- **But:** Is Big Data like Cloud Computing? Big Hype – No Business?

- **Or:** Market driven by Business Needs or Drives the Market the Business?

- **Last Question:** What exactly is big?

# Technical Backgroud: Storage



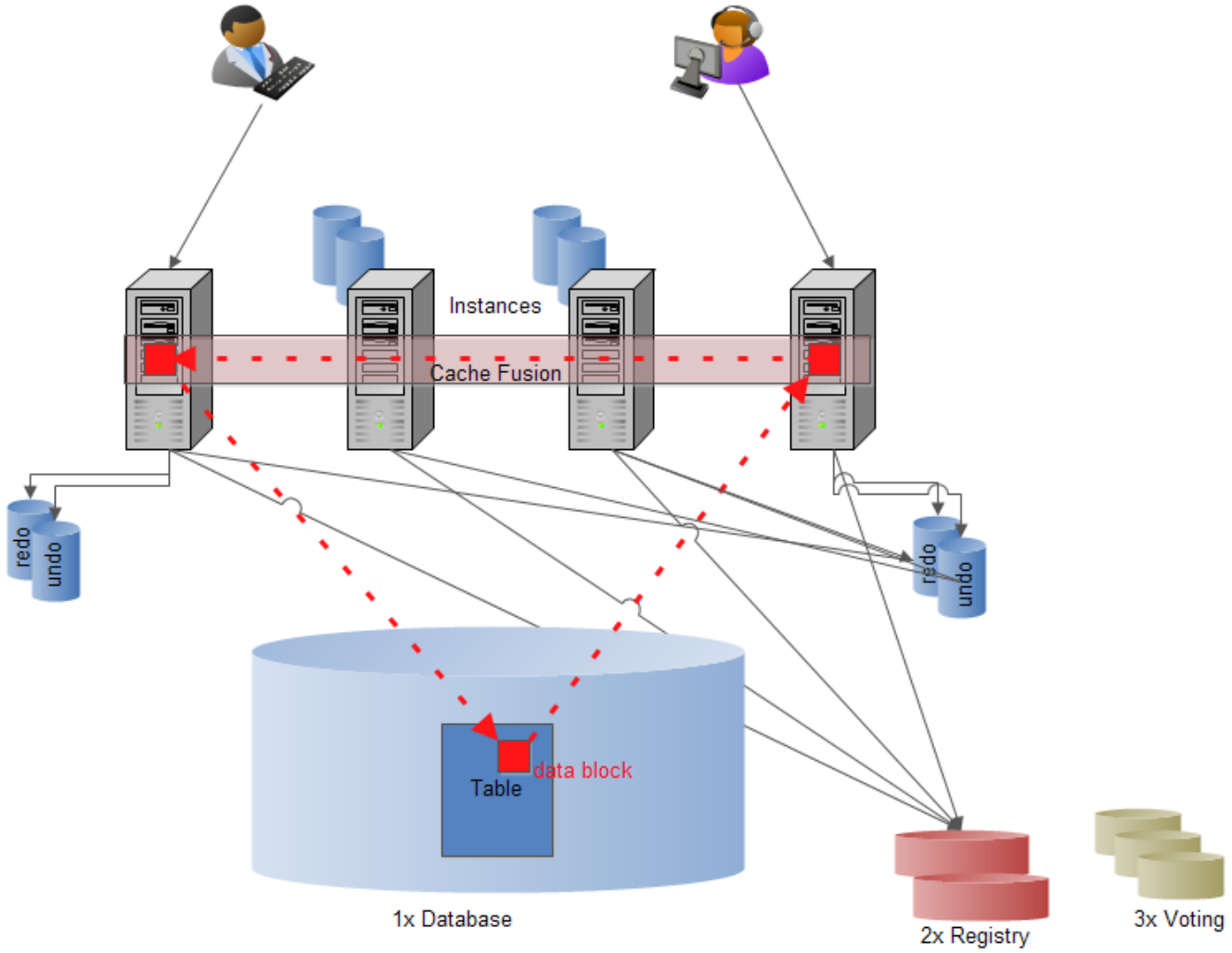
# Technical Backgroud: NoSQL Data Modeling

- NoSQL data modeling often starts from the application-specific queries as opposed to relational modeling:

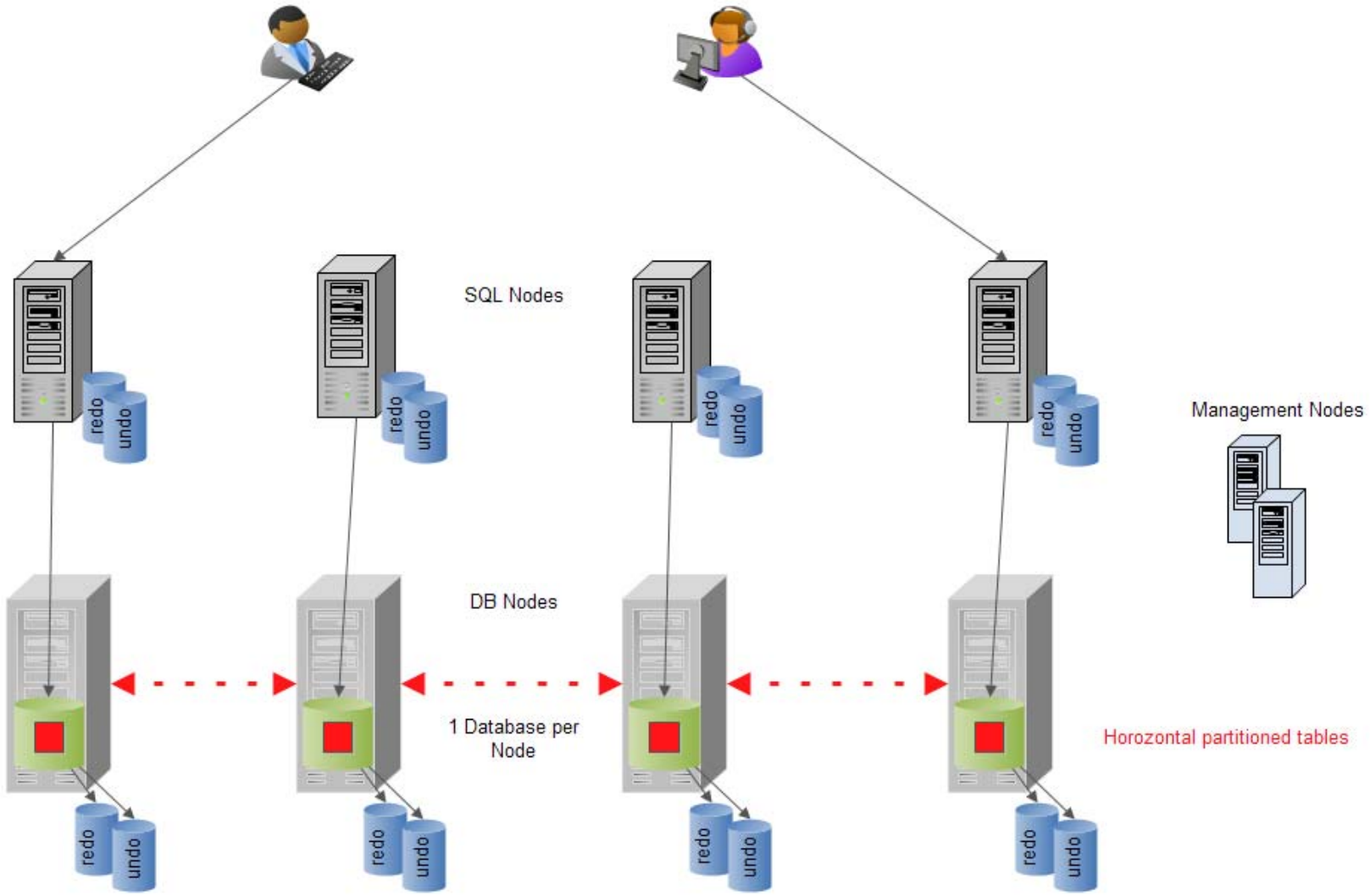
  - Relational modeling is typically
    - driven by structure of available data,
    - the main design theme is *"What answers do I have?"*
      - *Answere oriented*
    - More generic

  - NoSQL data modeling is typically
    - driven by application-specific access patterns, i.e. types of queries to be supported.
    - The main design theme is *"What questions do I have?"*
      - Question oriented
    - More application specific

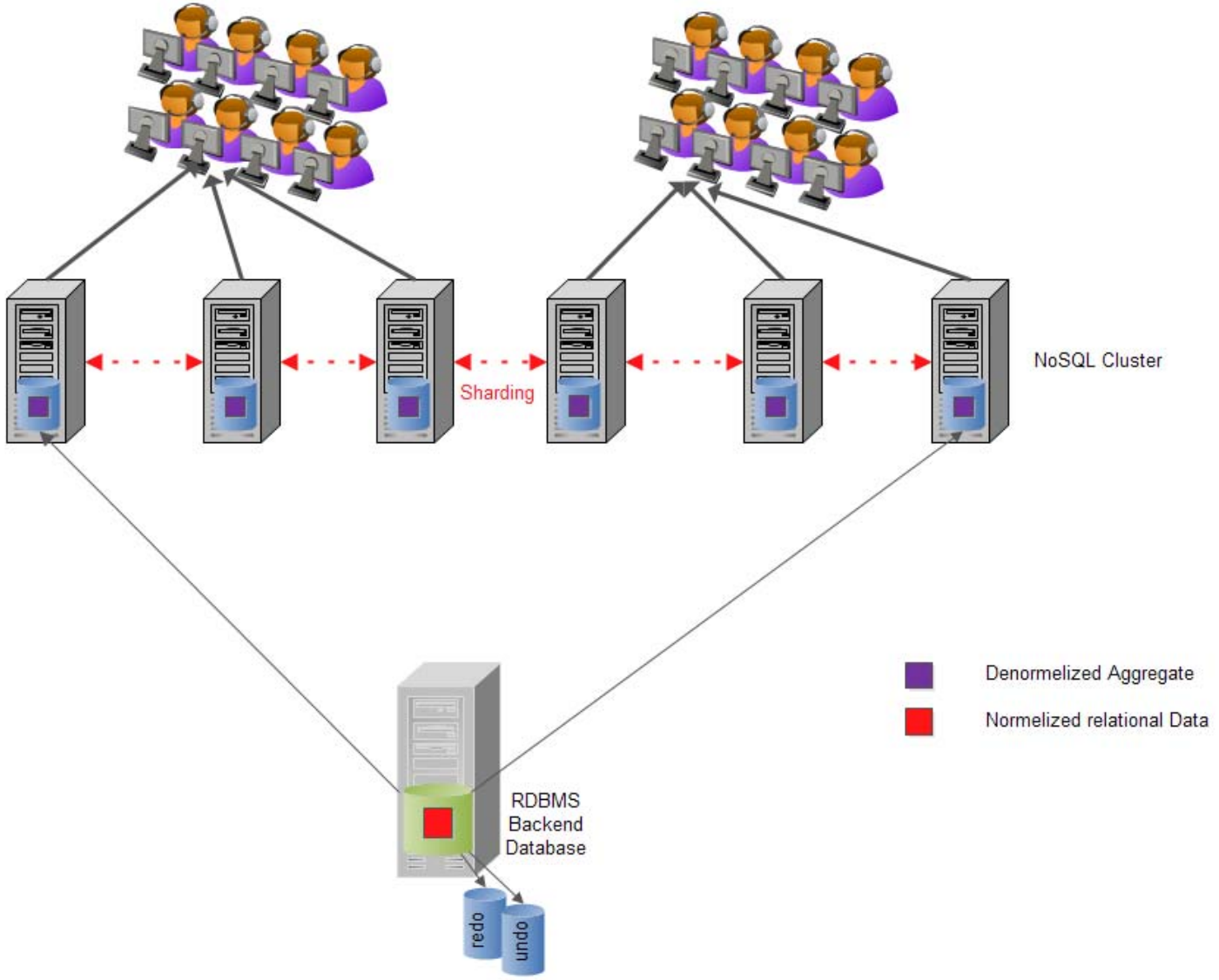
# Technical Backgroud: Big Data Processing Example Oracle RAC



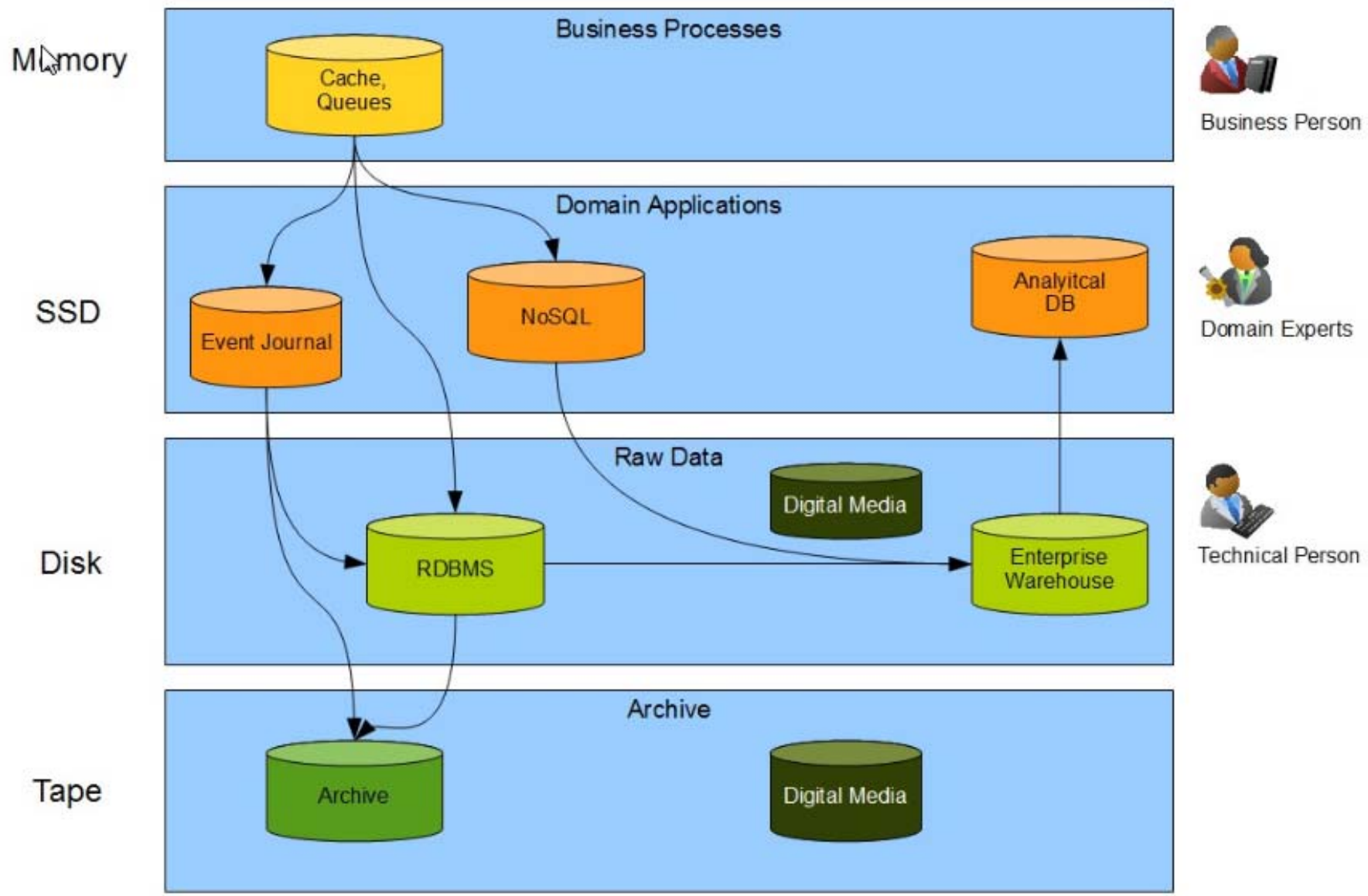
# Technical Backgroud: Big Data Processing Example Alternative 1: MySQL Cluster



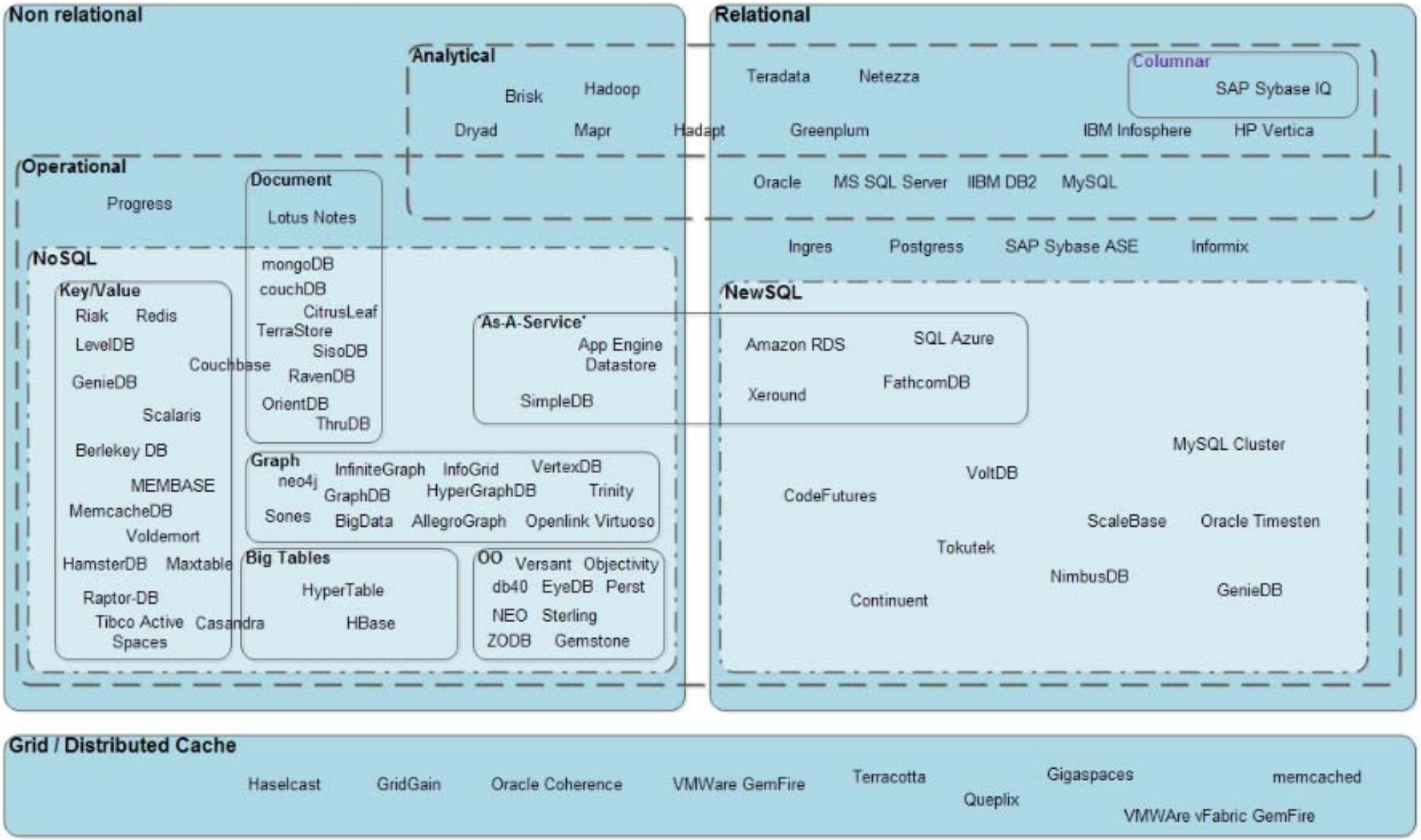
# Technical Backgroud: Big Data Processing Example Alternative 2: NoSQL / NewSQL



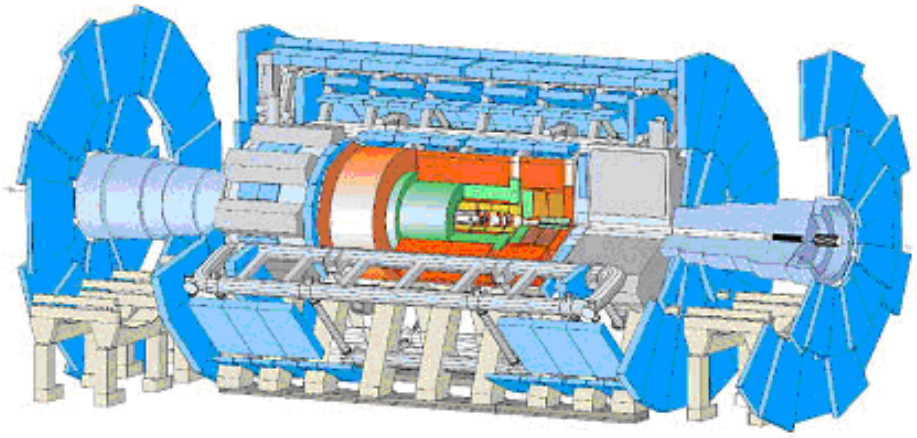
# Technical Backgroud: Storage Layers

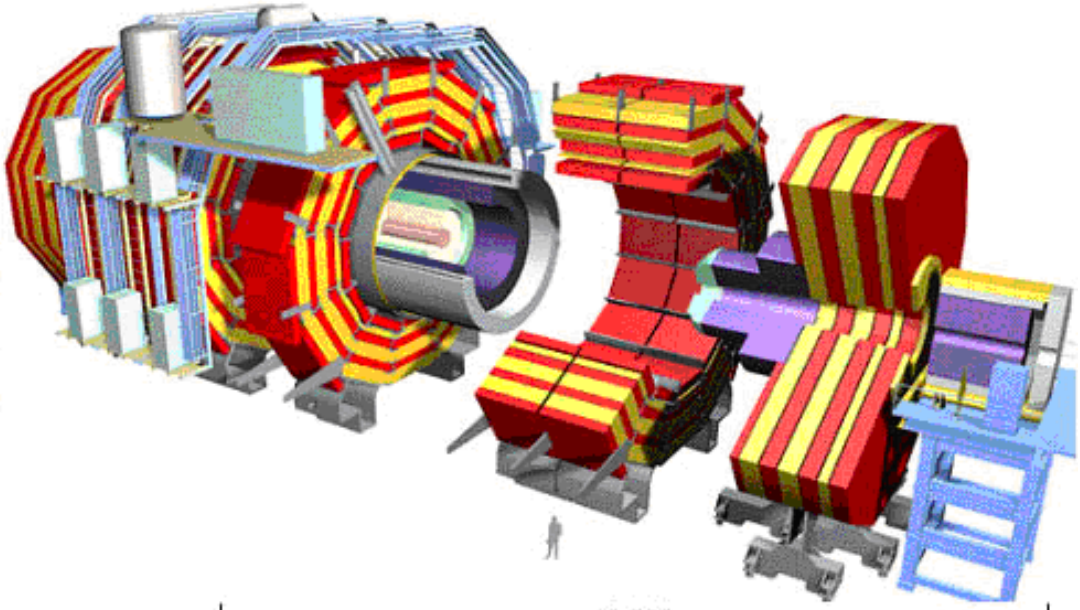


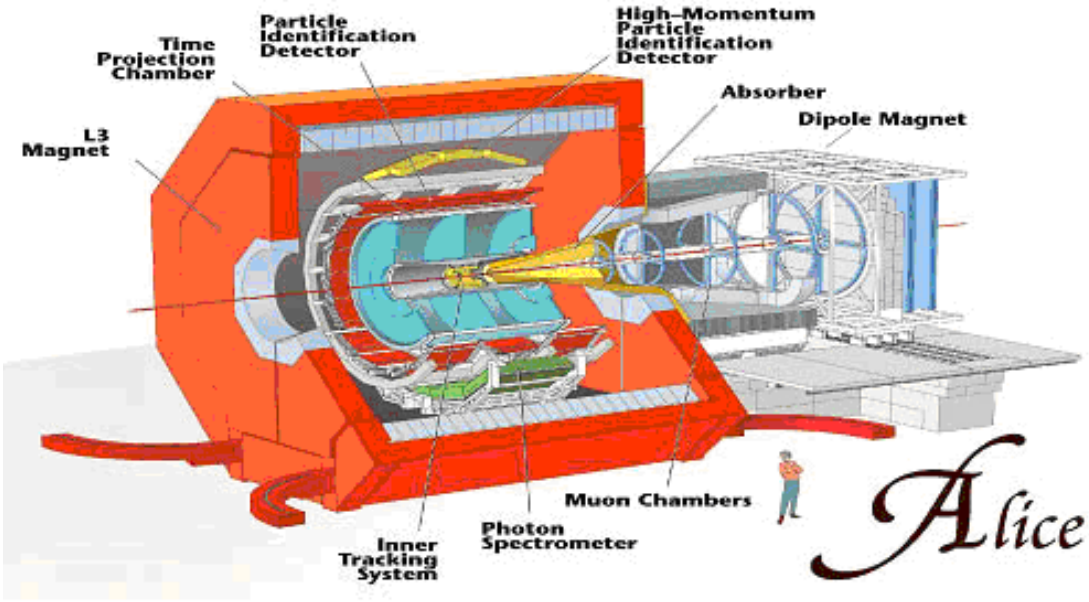
# Technical Backgroud: Products Today

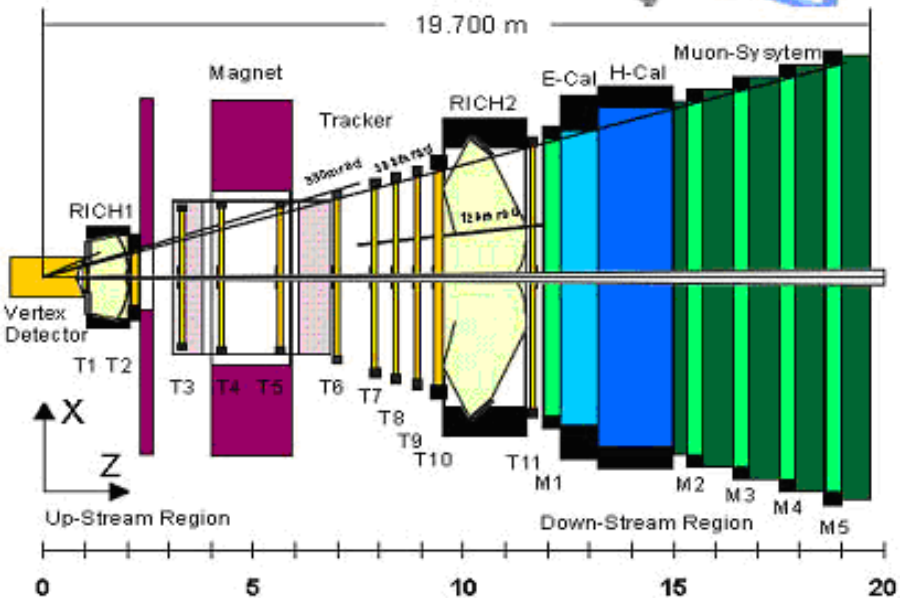


# Business Cases: Classics 1 – Scientific Data Processing







# Business Cases: Classics 1 – Scientific Data Processing

| Test | Usage | Data Rate | Data Volume |
|---|---|---|---|
| ATLAS | Higgs-Bosons | 100 Mb / sec | **1 Petabyte / year** |
| CMS | Higgs-Bosons | **100 Mb / sec** | **1 Petabyte / year** |
| Alice | Big-Bang Simulation | **1.5 Gb / sec** | **1 Petabyte / month** |
| LHCb | Will be explained later on ☺ | | **400 TB / year** |

# Business Cases: Social Media

- **Facebook:** Facebook stores much of the data on its massive Hadoop cluster, which has grown exponentially in recent years.
- Today the cluster holds 30 petabytes of data or, as Facebook puts it, about 3,000 times more information than is stored by the Library of Congress.

- **Key Issue for Companies:** Integration of Social Media for Market Research, Product Placement, Recommendations

# Business Cases: Advanced Analytics

- **Walmart:** Largest Commercial DWH: ca. 500 TB

- **Key Issue:** Retail Analytics – Combines Social Media Data, Geoinformations and Demographic Information for Shop Plannings

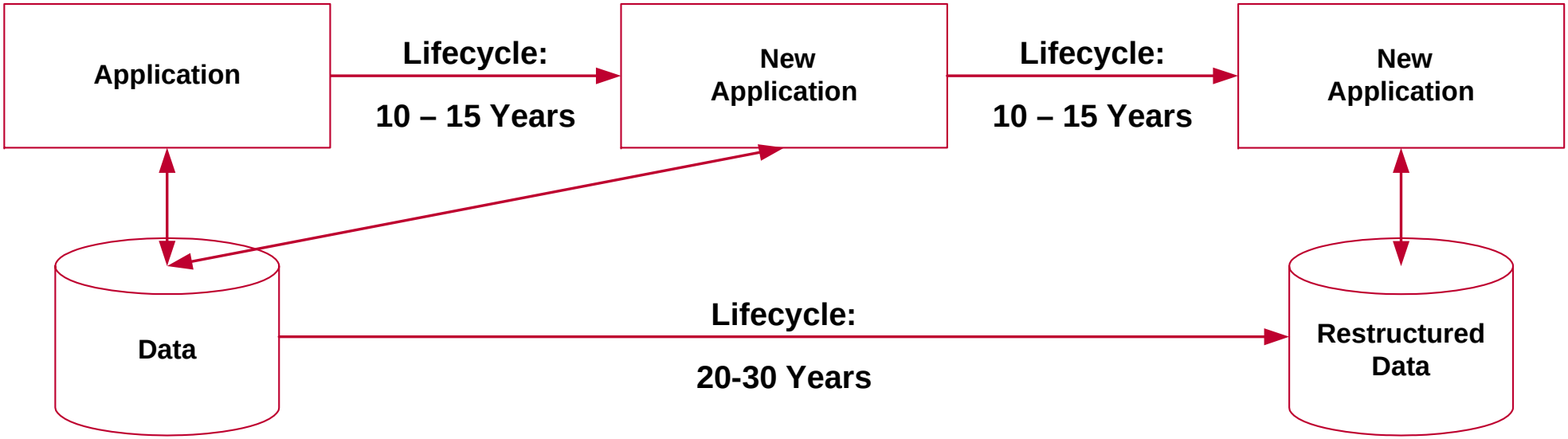
- **Log Analysis**
- **Fraud Detection**
- **Call Center – Call Storage**

- **Big Data Analytics:** As a market…

# Business Cases: War Rooms / Telco / Traffic Analysis

- **War Rooms:** Finding Natural Ressources without digging in the deep
- **Telco / Energy:** Network Optimization
- **Public & Private Transport & Logistics:** Traffic Analysis & Routing Optimization

# Business Cases - Future Trend: Structured & Non Structured Data



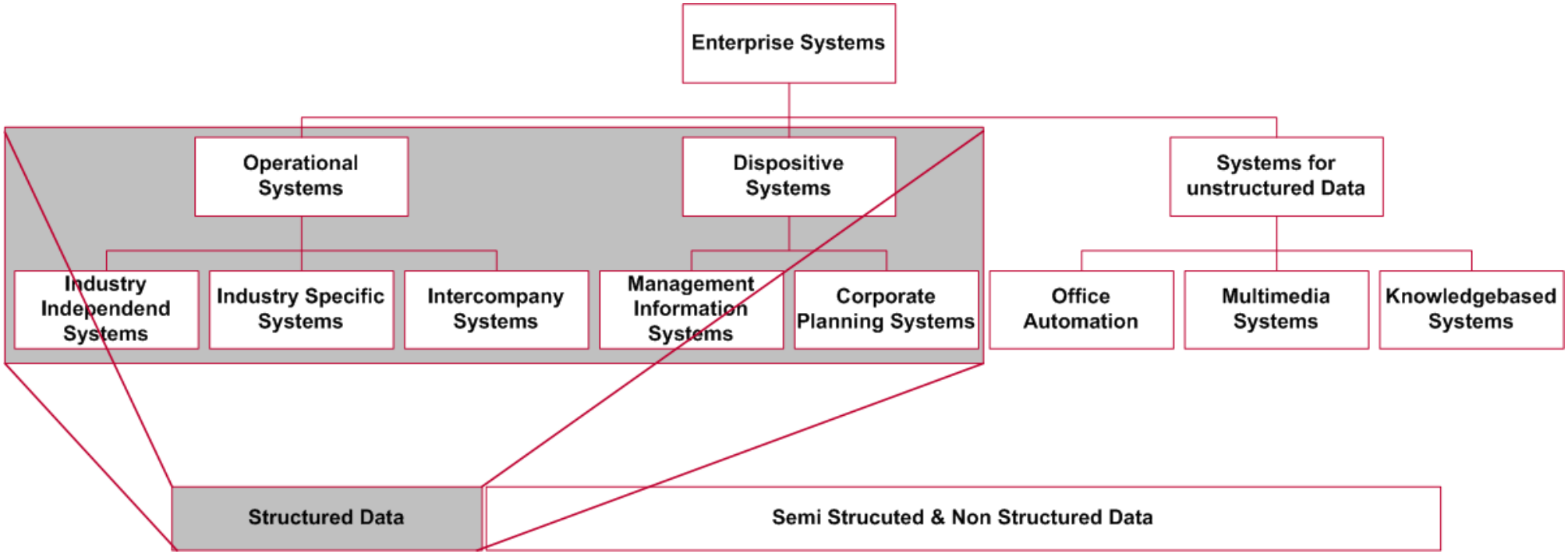

**Applications:** Lifecycle (Build – Run – Dismiss): 10-15 years

**Data:** Lifecycle (Build – Run – Dismiss): 20-30 years

Data ist structured (ca. 20%) or non structured (ca. 80 %)

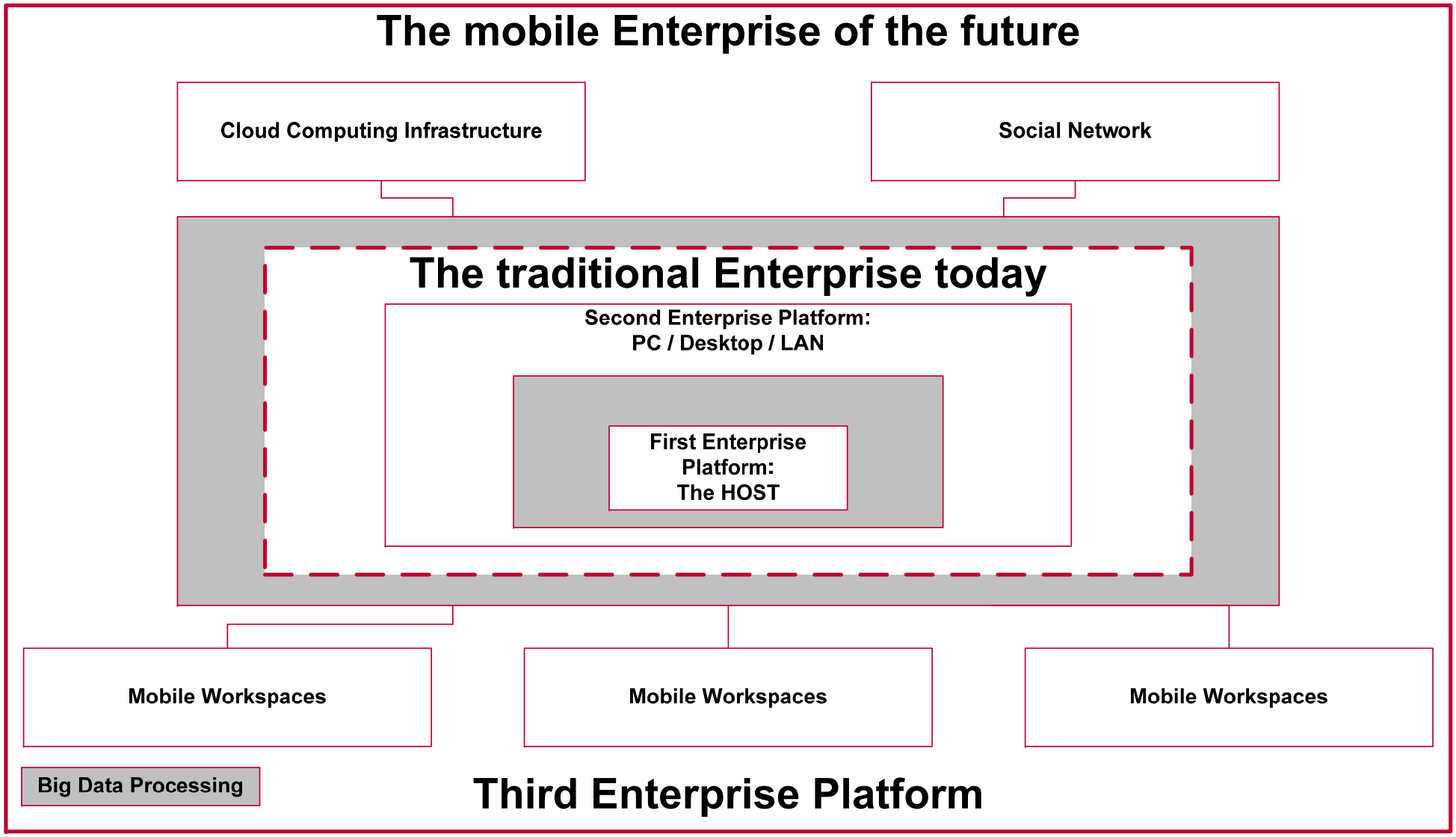
**Consequence:** Separation of systems & management prinicples

# Business Cases - Future Trend: Structured & Non Structured Data




**Consequence:** Big Data as a solution to an open problem: How to access Non Structured Data ?

# Business Cases - Future Trend: „Third Plattform“



# Ready to go?: Considerations for Enterprises

- **What to watch:** How Big ist my Big Data? (Investment, Standards)
- **What to watch:** Maturity of products / providers (ask you consultant)

- **How to watch:** Focus on Usage Pattern
- **How to watch:** Check the important products

- **When to try:** Check your Business Case – Consumer Business? / Geoinformation / Chemical Processchains (Graph Analytics)? / RFID / NFC?
- **When to try:** Is Scaling Out your problem?

# Ready to go?: Considerations for IT

- **Technical Use Case:** Universal Access (RDBMS) versus Predefined Access (NoSQL)
- **Technical Use Case:** Near Real-Time really needed?
- **Technical Use Case:** Combine Analytics of Structured (Yeah!) & Unstructured Data (Text, Video, Picture)

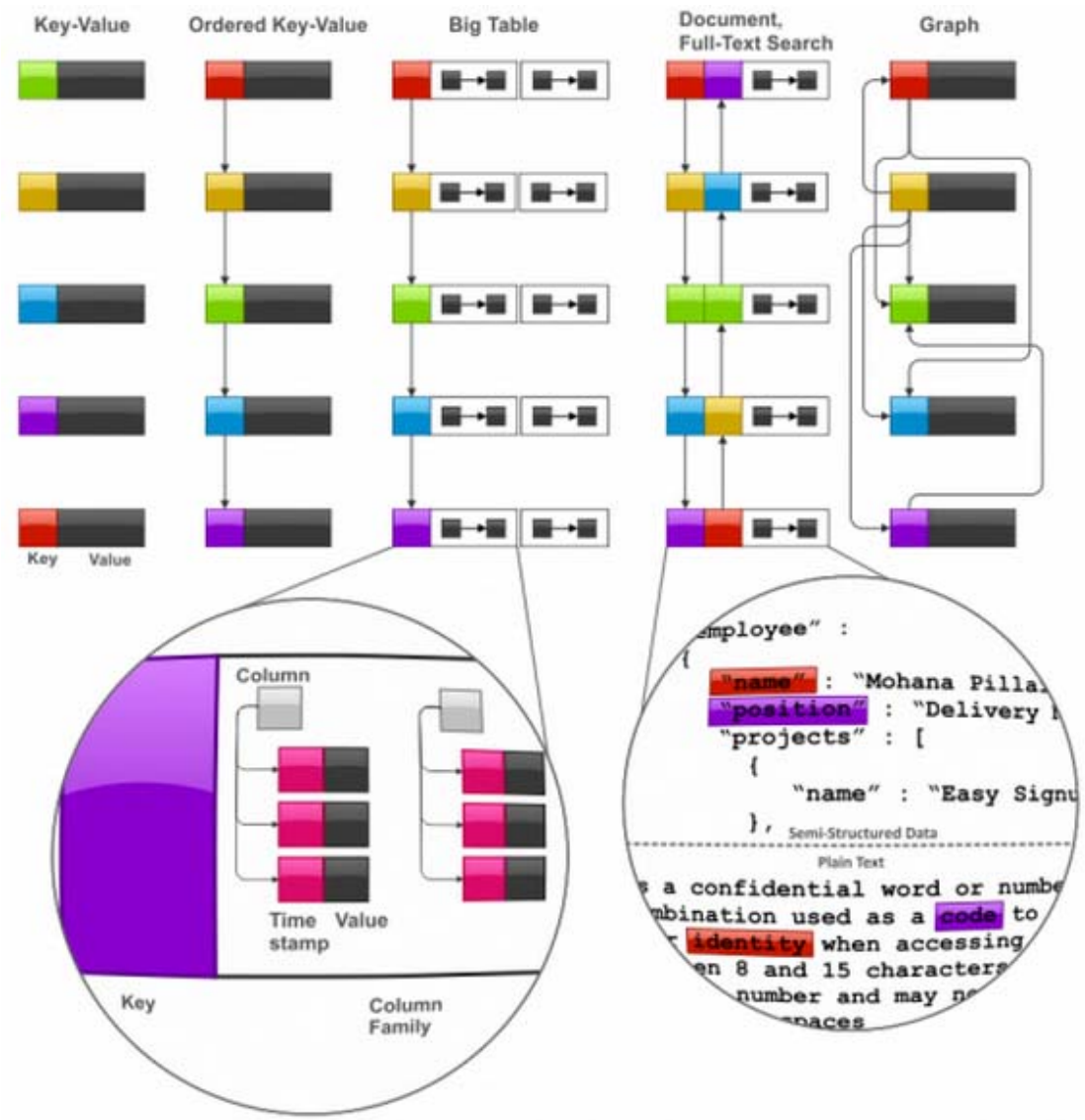
- Watch out for upcoming standards!
- Combine Solutions!

# Thanks for your attention

# Additional Slides

- **Look Ma – No Tables!**

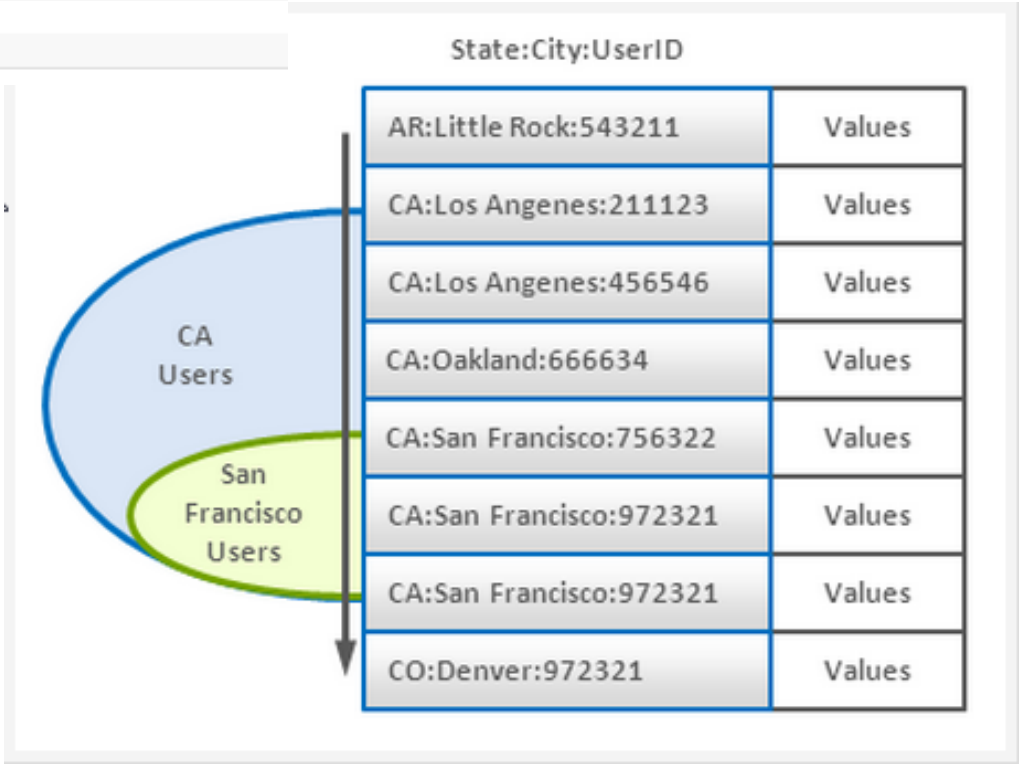
# Technical Backgroud: Look Ma – No Tables!



# Technical Backgroud: Oh – But here they are ☺

- Composite key is a very generic technique, but it is extremely beneficial when a store with ordered keys is used.
- Composite keys in conjunction with secondary sorting allows to build kind of multidimensional index which is fundamentally similar to the Dimensionality Reduction technique

```
1  SELECT Values WHERE state="CA:*"
2  SELECT Values WHERE city="CA:San Francisco*"
```



# Ready to go?: Some Considerations

- There is a dark side to most of the current NoSQL databases.
  - They talk about performance, about how easy schemaless databases are to use.
  - About nice APIs.
  - They are mostly developers and
  - **not operation and system administrators**.
  - No-one asks those. **But it's there where rubber hits the road.**
- The **three problems** no-one talks about – almost none, are:
  - **ad hoc data fixing** – either no query language available or no skills
  - **ad hoc reporting** – either no query language available or no in-house skills
  - **data export** – sometimes no API way to access all data

Blog answeres:

- a NoSQL store doesn't imply no RDBMS at all.
- Good discussion of running both types of DBs in one system: http://johnpwood.net/2009/09/29/using-multiple-database-models-in-a-single-application/