The Big Data Challenge

2015

Professor Sasu Tarkoma

What is Big Data?

90% of the world's data has been generated over the last two years

Facebook produces 10 TB and Twitter 7TB of data per day

Facebook has 40 000+ Hadoop nodes

By 2020 we have 35 Zettabytes of data ($35*10^{21}$B)

"Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools." - Wikipedia, 2014

Big Data

- A **massive** volume of structured and unstructured data
- **Cannot** be processed with traditional database and software solutions
- **Traditional** data analysis algorithms run **too slow** over the data
- ***High-volume, high-velocity, high-variety***
- **Big Data Pipeline:**
  - Data acquisition, storage, analysis, post-processing, results

Big Data Demonstrators

IBM Watson

Google Translate

Genome analysis

Financial data analysis

Astronomy

CERN LHC

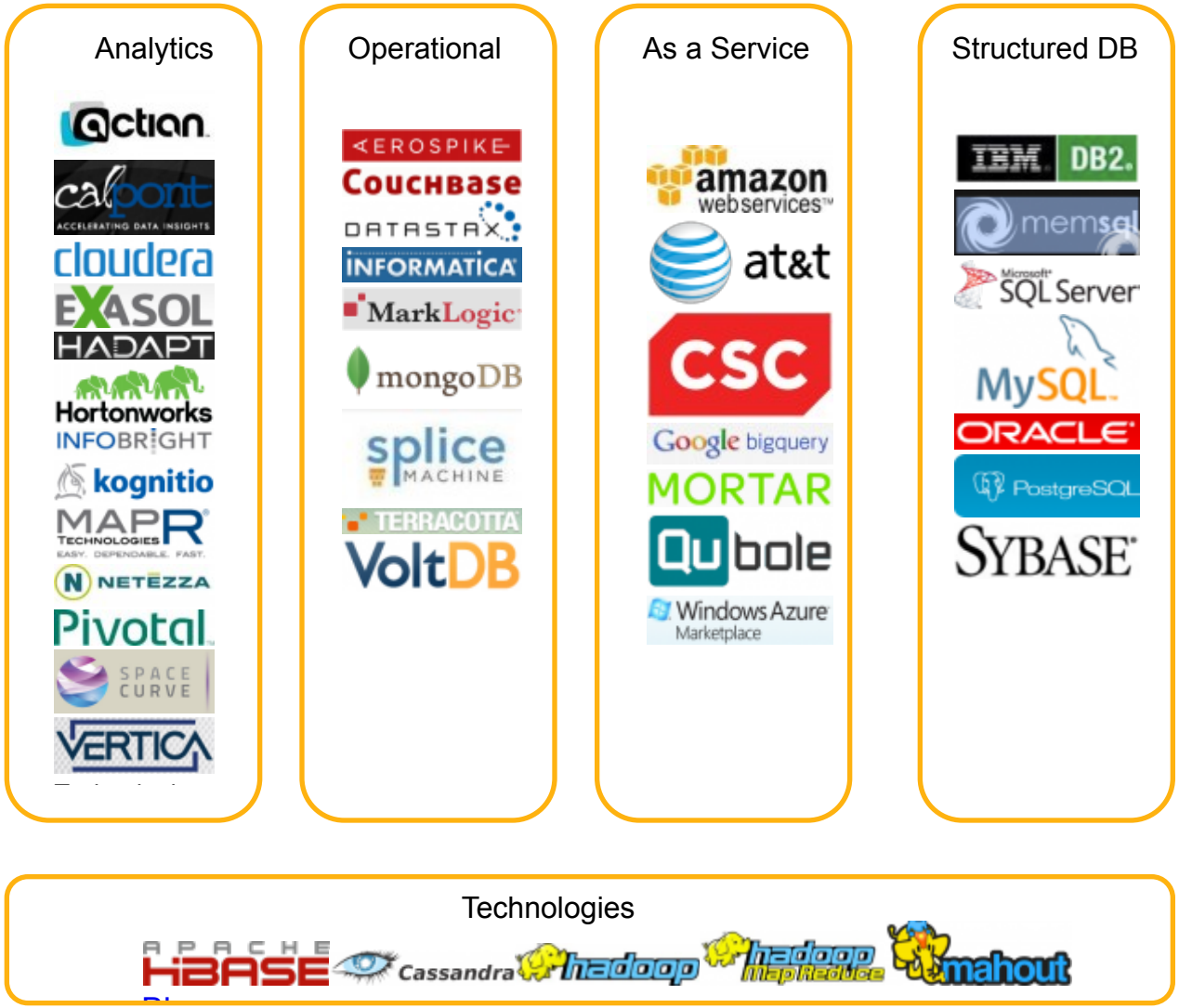
Smart traffic management

…

Big Data Prospects

Domain	Data sources	Prospects
Manufacturing	Sensor information from products and machines	Improved and more intelligent automation, automatic diagnosis and support
Public sector	Smart city, open data, surveys	New services, customized services, improved processes
Retail	Sensor information from products and smart malls, interactive shops, social media	Marketing and advertisement, sentiment analysis, personalized service, smarter shopping
Healthcare	Patient monitoring, environmental monitoring	Personalized medicine, preventive care, improved processes
Science	Various, combination of data sources	Address problems with huge data requirements, data fusion
Positioning services	Location data	Smart traffic, localized services, localized advertising

http://www.bigdatalandscape.com



http://www.bigdatalandscape.com

Big Data Problems

Scalability

Data leaks

Privacy problems

Centralization

Too much noise in the data

Big Data has value, but

also **Small Data** is Valuable!

Big Data Process

1. Acquisition
2. Extraction
3. Integration
4. Analysis
5. Interpretation
6. Decision
7. Understanding decision and starting from 1.

We have a feedback loop and the process is iterative.

Big Data Challenges

- **New computational paradigms**
  - Parallel, cluster, multi-cluster, decentralized
- Computational **barrier**
  - Linear or sublinear algorithms demanded
- **Theory**
  - Lack of algorithms for distributed case
  - Understanding the relation between machine learning tasks and the platform (and system resources)
- **Practice**
  - Not easy to run a cluster
  - How to choose the algorithms and the system parameters

Three Views to Big Data

Batch processing for **high volume**

- Parallel and distributed for static data
- High latency

Real-time processing for **high velocity**

- Parallel and distributed processing of streams
- Low latency

NoSQL for **heterogeneous data**

- Tunable performance and consistency