WORKSHOP ON LARGE-SCALE TOMOGRAPHY

BIG DATA: SIZE DOES MATTER!

LAJOS RODEK
BIG DATA ARCHITECT, EPAM SYSTEMS, SZEGED
LAJOS_RODEK@EPAM.COM

JANUARY 26, 2016

DISCLAIMER

NO SCIENCE TODAY!

AGENDA

1. Introduction to Big Data
2. Big Data in practice
3. Technologies & tools
4. Conclusions

INTRODUCTION TO BIG DATA

DEFINITION OF BIG DATA

"… a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis." (IDC, 2012)

"… high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." (Gartner, 2013)

THE 3 V'S

Volume

- Scale of data
- Large & expanding
- Many data sources

Velocity

- Rate of data arrival
- Rate of processing: offline (batch) vs low-latency vs real-time (stream)
- Rate of changes

Variety

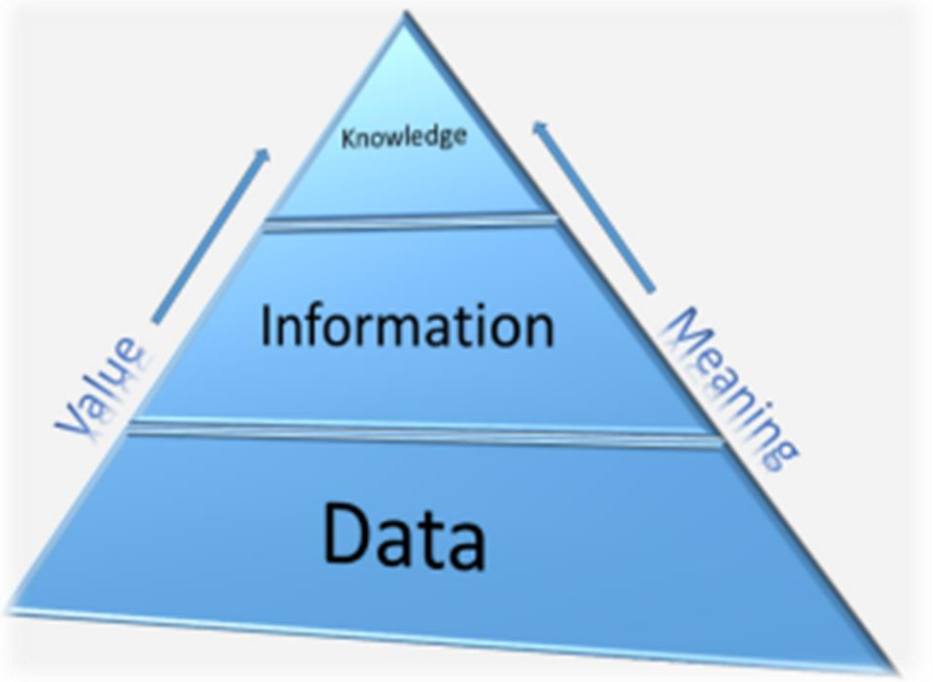
- Structured vs unstructured vs semi-structured data
- Text vs binary data
- "Dark data"

(Doug Laney, META Group / Gartner, 2001)

ONE MORE IMPORTANT V

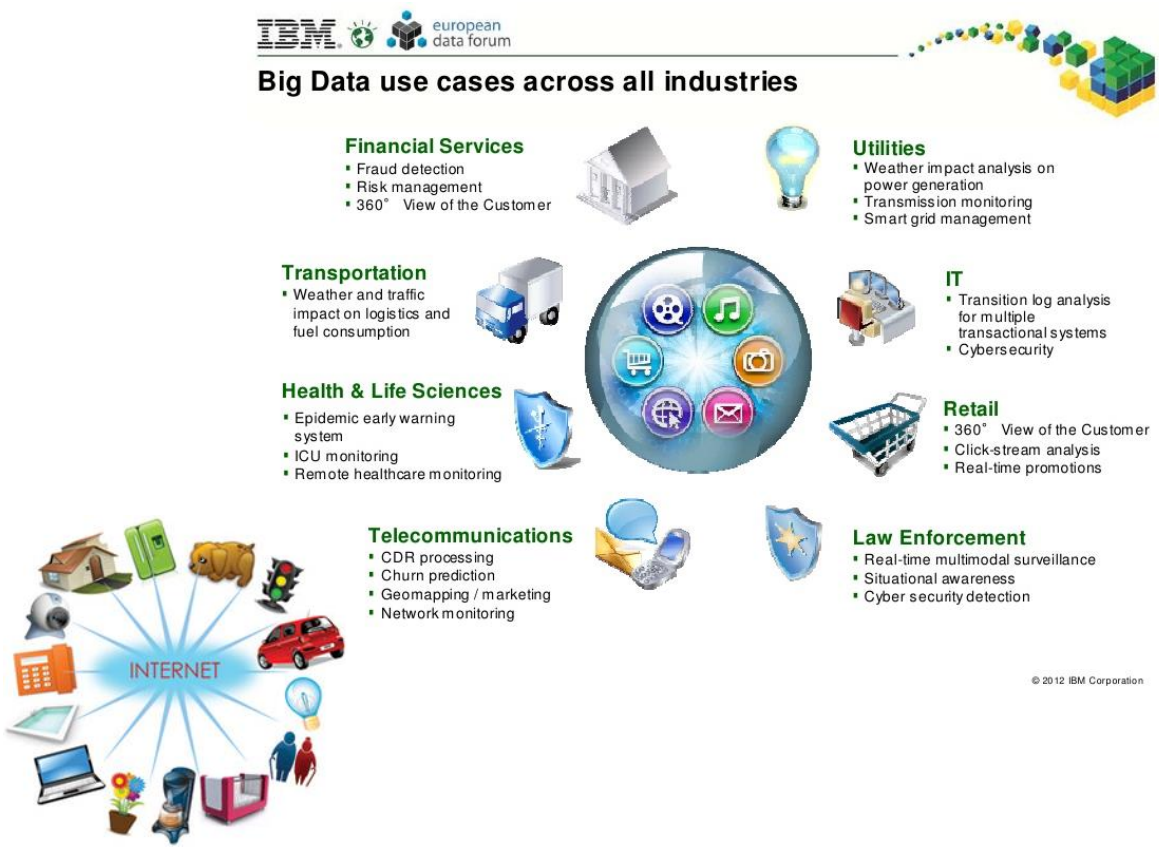
Value

- Relevance
- Outcome
- Actions



USE CASES


TOP BIG DATA USE CASES
Customer
Financial
Marketing
Retail
Security
Pharma
Customer Analytics 48%
45% Experience Analytics
Threat Analysis 30%
Risk Analysis 37%
28% Regulatory Compliance Analysis
23% Location-based Targeting
Campaign Optimization 26%
Fraud Analysis 22%
16% Brand Sentiment Analysis
Product Placement Optimization 16%
9% Other
Drug Discovery 1%


IBM
european data forum
Big Data use cases across all industries
Financial Services
Fraud detection
Risk management
360° View of the Customer
Utilities
Weather impact analysis on power generation
Transmission monitoring
Smart grid management
Transportation
Weather and traffic impact on logistics and fuel consumption
IT
Transition log analysis for multiple transactional systems
Cybersecurity
Health & Life Sciences
Epidemic early warning system
ICU monitoring
Remote healthcare monitoring
Retail
360° View of the Customer
Click-stream analysis
Real-time promotions
INTERNET
Telecommunications
CDR processing
Churn prediction
Geomapping / marketing
Network monitoring
Law Enforcement
Real-time multimodal surveillance
Situational awareness
Cyber security detection
© 2012 IBM Corporation


BIG DATA IN PRACTICE

TYPICAL TASKS

Distributed data storage

- Even geographically → Multiple data centers

Distributed data processing

- Collect
- Transform
- Query
- Analyze & understand

Distributed computing

PRINCIPLES 1.

Robustness & reliability on SW framework level

- Fault tolerance
- Redundant storage

“Keep everything”

- Including raw data

Linear (or better) scalability

- Horizontal (scale out) vs vertical (scale up)
- Scale down
- Dynamic / elastic / autoscaling

PRINCIPLES 2.

Efficiency

- High-throughput
- Low-latency

Data locality

- Execute computation where data are located → No unnecessary data transfers

Running on commodity HW

Dominated by open-source, community-driven SW (vs proprietary)

CHALLENGES 1.

Choosing the right tool

- Abundance of options ☺

Efficient data access

- Denormalization
- Graph schema
- Serialization

Testing

- Verification
- Debugging
- Performance measurement

CHALLENGES 2.

Enterprise integration

- Data hub / lake

Extremely large data size (exponential growth)

- Data federation / virtualization

Data governance

- Data sources, data integration / fusion, data catalogs, metadata management
- Data quality
- Security, privacy, legal compliance
- Retention policy

CHALLENGES 3.

High Availability (HA)

- No single point of failure (SPoF)
- Standby / fallback
- Replication / synchronization

Service Level Agreement (SLA)

- Availability
- Multi-tenancy
- Quotas
- Scheduler policy

CHALLENGES 4.

Administration / operation

- Installation, provisioning
- Monitoring
- Management
- Troubleshooting

Expenses

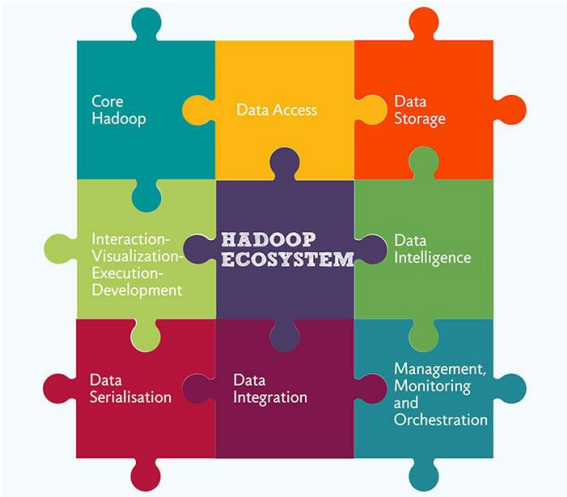
- Infrastructure
- Experienced workforce (e.g. Data Scientist, Data Engineer, Platform Engineer)
- Trainings, learning curve
- Commercial support / consultancy


TECHNOLOGIES & TOOLS

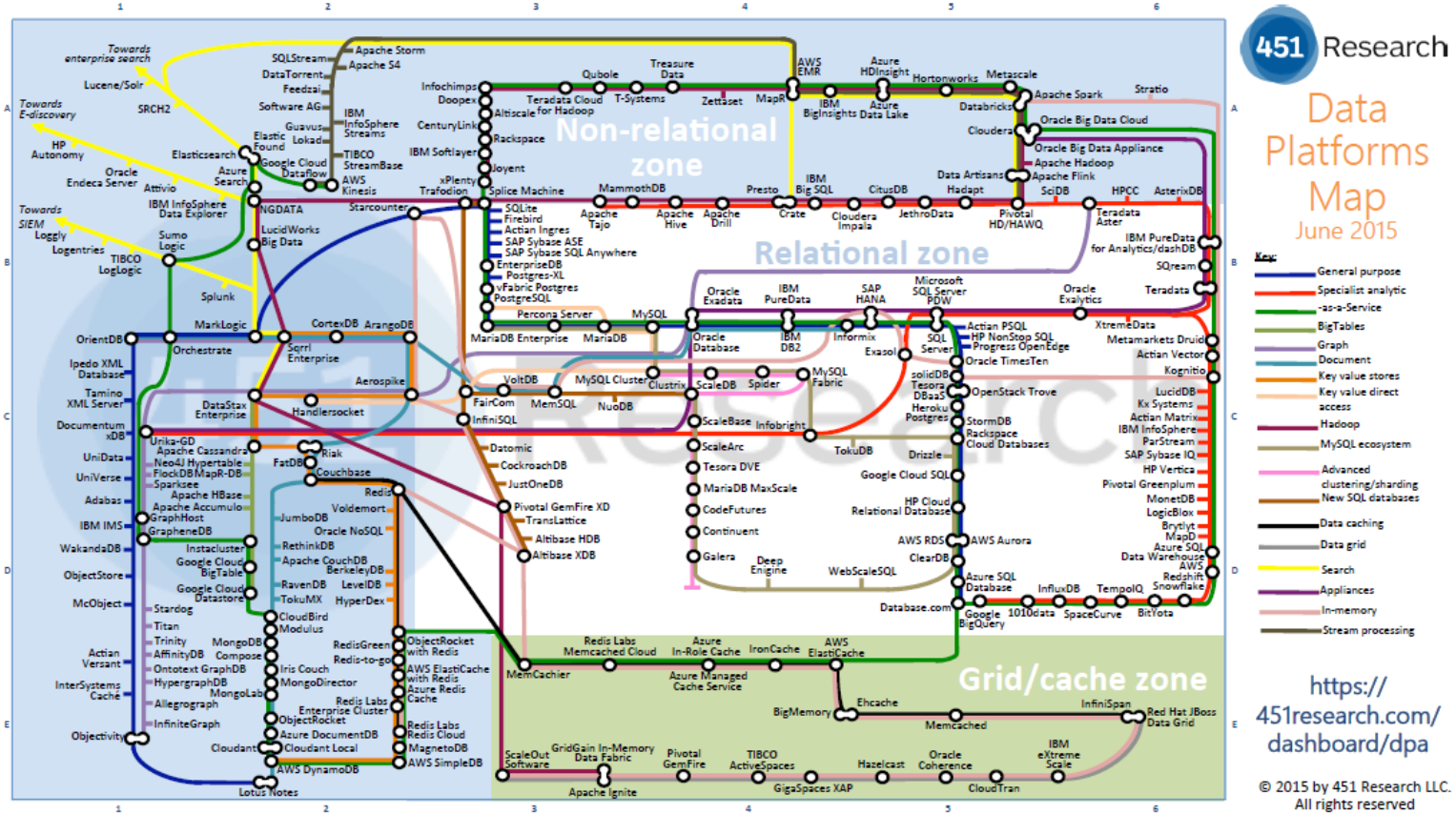
BIG DATA OPEN-SOURCE LANDSCAPE

APACHE HADOOP AND ITS ECOSYSTEM

The Apache Software Foundation
Community-led development since 1999.



STORAGE: RDBMS, NEWSQL, NOSQL, GRID / CACHE



451 Research

Data Platforms Map

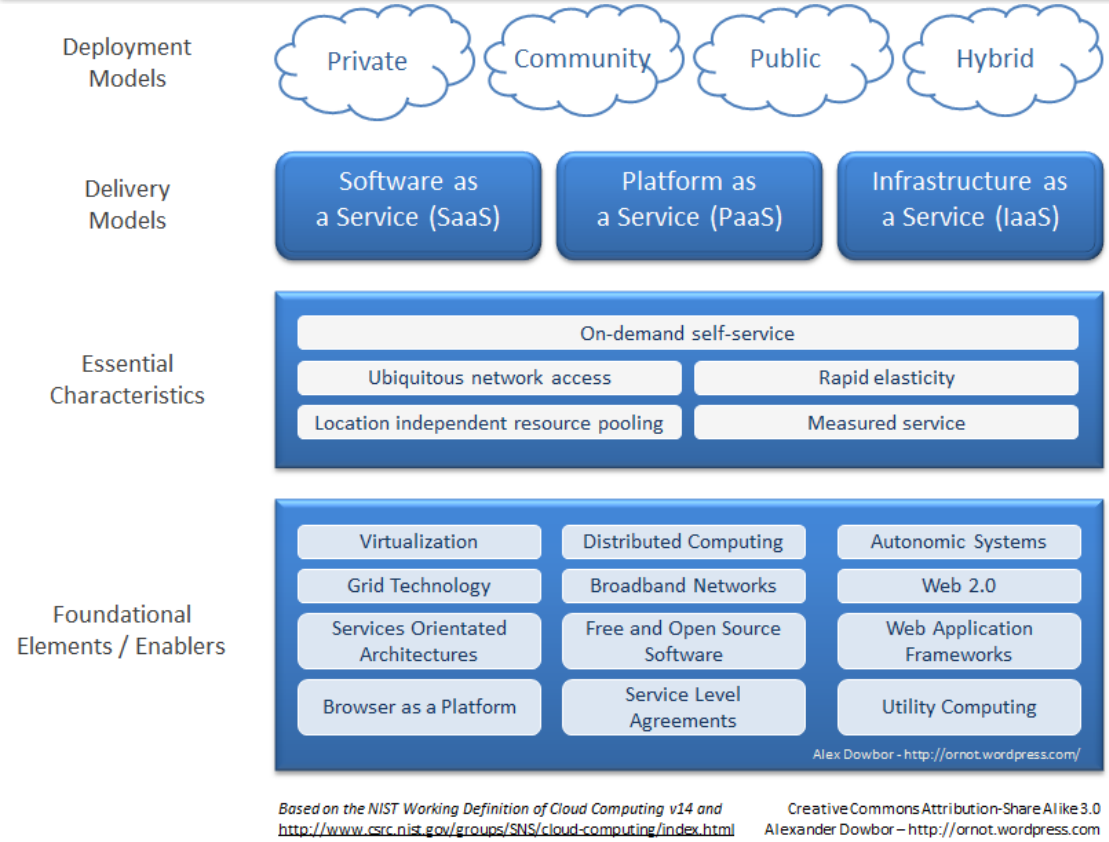
June 2015

Key:

- General purpose
- Specialist analytic
- -as-a-Service
- BigTables
- Graph
- Document
- Key value stores
- Key value direct access
- Hadoop
- MySQL ecosystem
- Advanced clustering/sharding
- New SQL databases
- Data caching
- Data grid
- Search
- Appliances
- In-memory
- Stream processing

https://451research.com/dashboard/dpa

CLOUD



Based on the NIST Working Definition of Cloud Computing v14 and http://www.csrc.nist.gov/groups/SNS/cloud-computing/index.html

amazon web services™

Windows Azure

openstack™ CLOUD SOFTWARE

Google Cloud Platform

rackspace® the open cloud company

Qubole

altiscale

Amazon Elastic MapReduce

APPLICATION DESIGN

Architecture

- Event-driven, reactive
- Lambda, Kappa
- Shared-nothing

Patterns

- MapReduce
- Actor model
- Data pipeline / flow

Algorithms

- Divide and conquer
- Concurrent / parallel

RELATED TOPICS: STORAGE

High-performance drives

- SSD
- RAID

Network storage

- SAN
- NAS

Network / distributed file systems

- NFS, Lustre, GlusterFS, GFS, HDFS, GPFS

“Fast data” (in-memory)

- Tachyon, GridGain / Apache Ignite file system

RELATED TOPICS: PROCESSING

High-performance networking

- InfiniBand, Fibre Channel, fiber-optics
- RDMA, zero-copy

Artificial intelligence

- Machine learning, NLP, data mining, dimension reduction

Analytics & statistics

- DWH, BI, data visualization

Data science

RELATED TOPICS: COMPUTING 1.

Parallel computing

- Multithreading, SMP, OpenMP
- GPGPU → OpenCL, CUDA
- SIMD, VLIW / MIMD, MPP, vector processors

Grid computing

- GigaSpaces XAP, GridGain / Apache Ignite, GemFire / Apache Geode, JPPF, HTCondor

HPC / supercomputers

- PVM, OpenMPI

RELATED TOPICS: COMPUTING 2.

Edge computing

- Sensor networks / IoT, P2P

“Fast data” (in-memory)

- Apache Spark, Apache Flink, SAP HANA


CONCLUSIONS

BIG DATA IS COMPLEX

POSSIBLE CONNECTIONS WITH TOMOGRAPHY

Storage

- Collect
- Query, retrieve
- Link with other data sources, associate metadata

Processing

- Transform, pre-process
- Analyze & understand
- Evaluate

Computing

- Reconstruct


THANK YOU!