A Perspective on Big Data Technology and Research Direction of ETRI

Seung Ku HWANG

ETRI **Big Data Software Research Lab.**

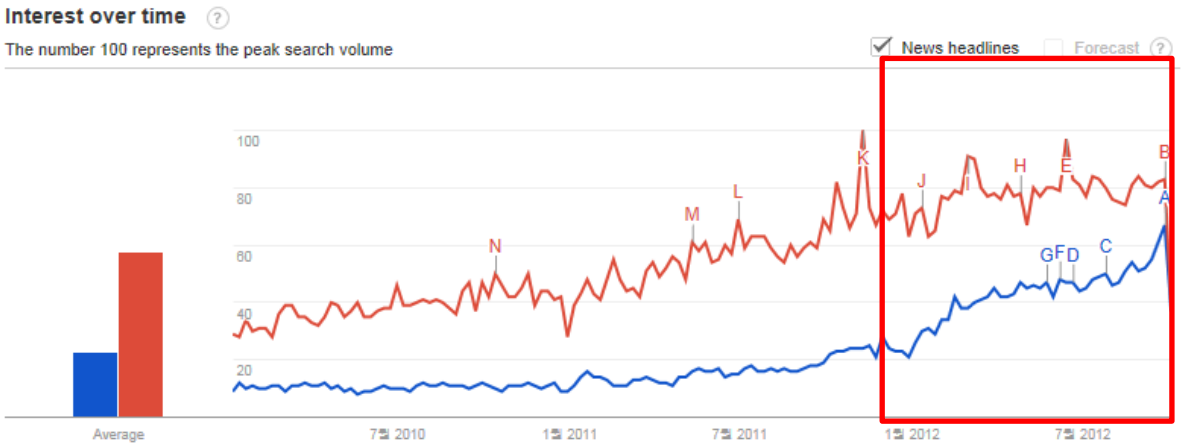
PRAGMA Workshop
2012. 10. 10.

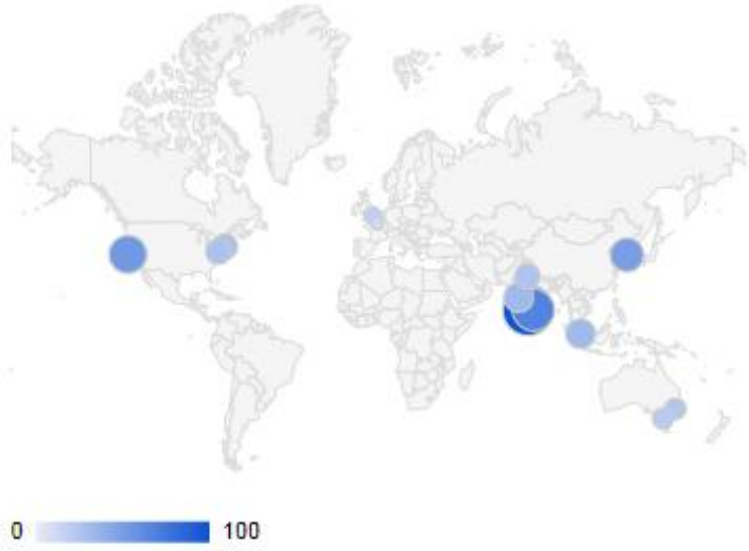
Contents

I. BIG DATA TREND

BIG DATA : IS IT REAL OR HYPE?

- Is there something really new and important, or is it just hype?
- Is this going to change the how we do things for years to come, or is it just a distraction?
- How do we really use it?





PARADIGM SHIFT

Capital

People

Data

"**Data** are becoming the new raw material of business :
an economic input almost on a par with capital and labour."

- The Economist, 2010

DATA REVOLUTION

Data will separate the winners and losers in every single industry.
The world's next, great natural resource, except it's not limited

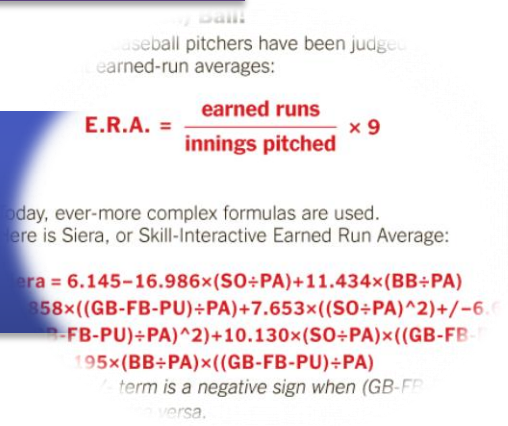
- IBM CEO Ginni Rometty, 2012

Data Revolution with the evolution of computing technology

DATA REVOLUTION - PUT THE DATA WORK!

Target Marketing : Diaper-Beer -
25~35yrs, first baby, late night, …

Management by Data :
New Criteria, New Algorithm
to create new Value



Predictive Asset Management :
- Reduce downtime to improve Productivity & save operation cost

DATA SOURCES

- Social Networks
- Blogs
- Chat Rooms
- Product Review Sites
- Web Pages
- email
- Documents

Data Sources

Unstructured

80%

- Images
- Videos
- Music
- Sensors
- Call Data Records
- Log Files
- Machine Generated

80/20

Structured : RDBMS, ERP/CRM, EDW

HOW BIG IS THE DATA PRODUCED…

LHC(Large Hadron Collider) – Particle Accelerator

40 TB/s

Boeing Jet Engine

10TB/30min/Engine Operation

Social Networks

Twitter : 2,300 Tweets/s(2011.6) ➔ 8 TB/day

(New York Stock Exchange ➔ 1 TB/day)

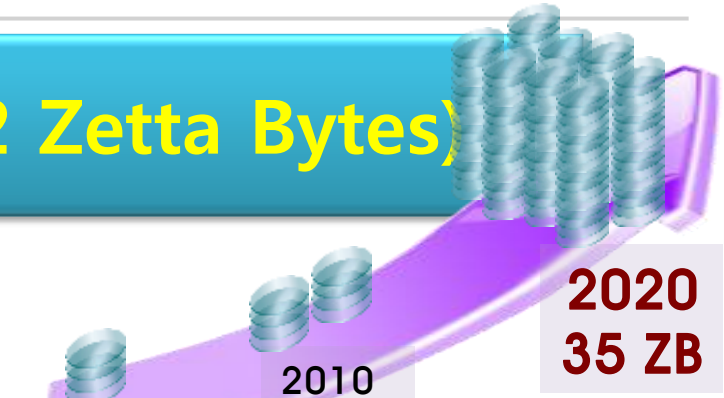
Facebook : 60~70TB/day

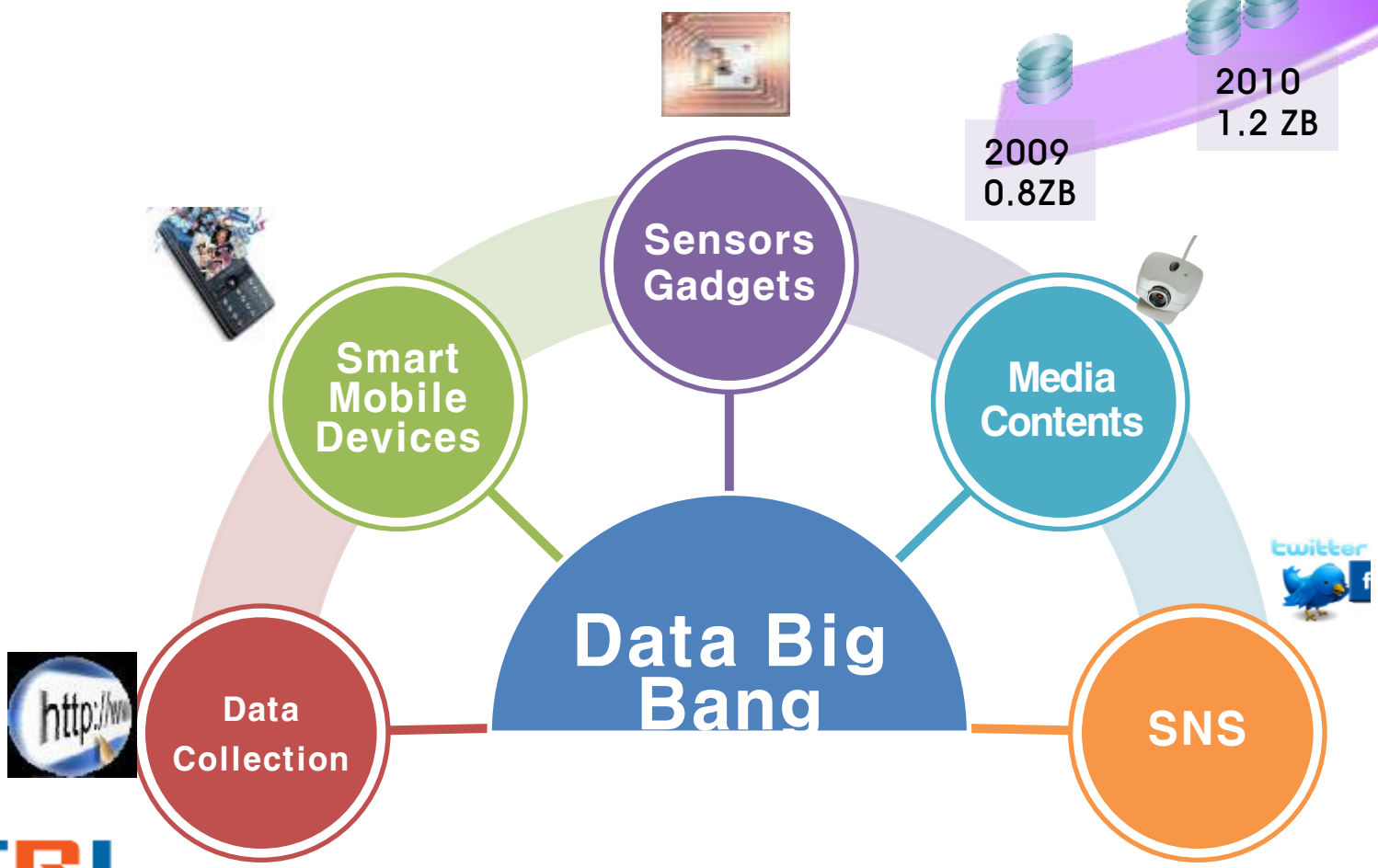
WalMart Transaction

1M Customer Transactions/Hour, DB Size – 2.5PB

DATA EXPLOSION ➔ BIG DATA

2010 Zetta Byte Era (2010 1.2 Zetta Bytes)





WHAT DO YOU MEAN BY BIG DATA?

3V: Volume, Variety, Velocity (Complexity, Value)

VOLUME

- **Many terabytes or petabytes of information. Exceed human capacity for reading and comprehension, creating demand for automated or highly assisted computer techniques for data exploration, navigation and discovery**
  - Turn 12 terabytes of Tweets created each day into improved product sentiment analysis
  - Convert 350 billion annual meter readings to better predict power consumption

VELOCITY

- **How fast the data arrives, how fast it is available, and how fast it can be evaluated and processed to meet the intended purpose**
  - Scrutinize 5 million trade events created each day to identify potential fraud
  - Analyze 500 million daily call detail records in real-time to predict customer churn faster

VARIETY

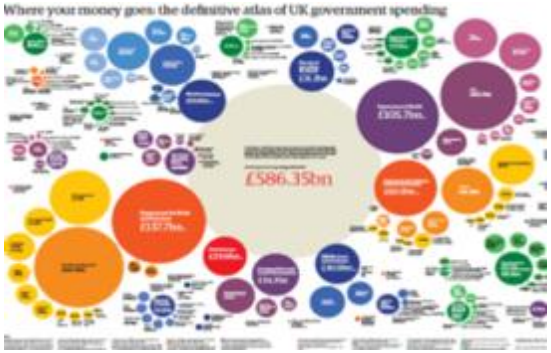
- **The myriad data types and formats, such as tabular data, hierarchical data, documents, email, social data, free text, metering data, video, image, audio, stock ticker data, financial transactions, …**
  - Monitor 100's of live video feeds from surveillance cameras to target points of interest
  - Exploit the 80% data growth in images, video and documents to improve customer

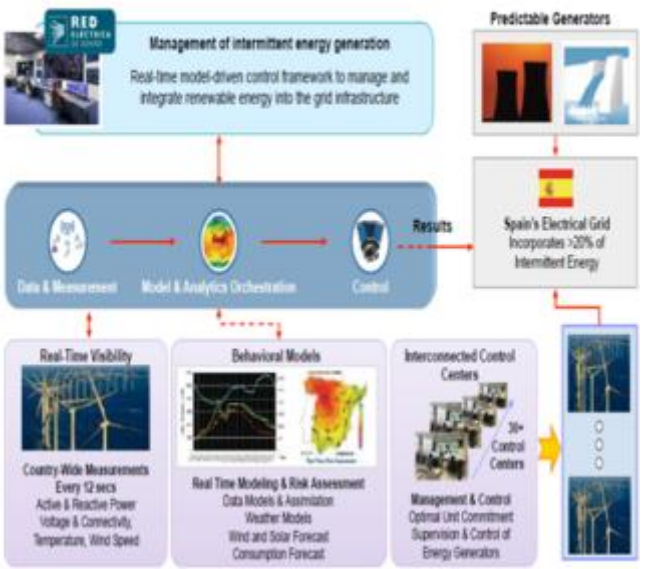
ROLES OF BIG DATA

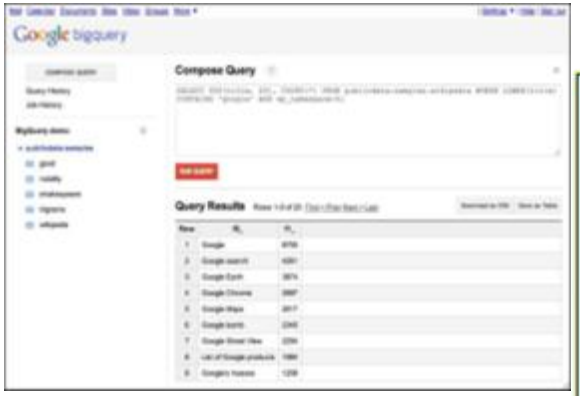
Features of Future Society	Roles of Big Data	
Uncertainty	Insight	- Future outlook, Forecast, Prospect - Scenario Simulation for various Possibility
Risk	Responsiveness	- From Issue Monitoring, Analysis to Fast Decision, Real Time Response - National Risk Management : Water Management, Disease Control,… - National Competitiveness
Smart	Competitiveness	- Opportunity to provide Personalized Intelligent Service - Product Competitiveness based on Trend Analysis - Optimal Decision from Social Needs, reputation,
Convergence	Creativity	- New Value Creation : Healthcare, Automobile, Environments - New Convergence Market

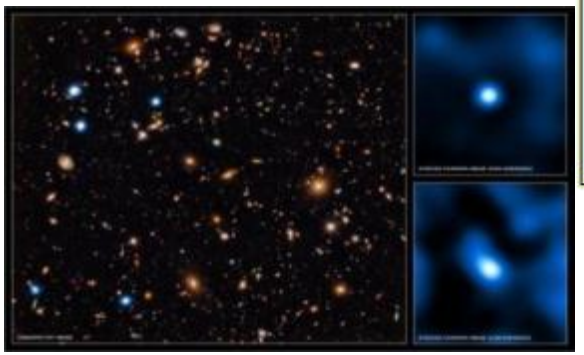
(Source: NIA Report 2011.12)

BIG DATA USE CASES



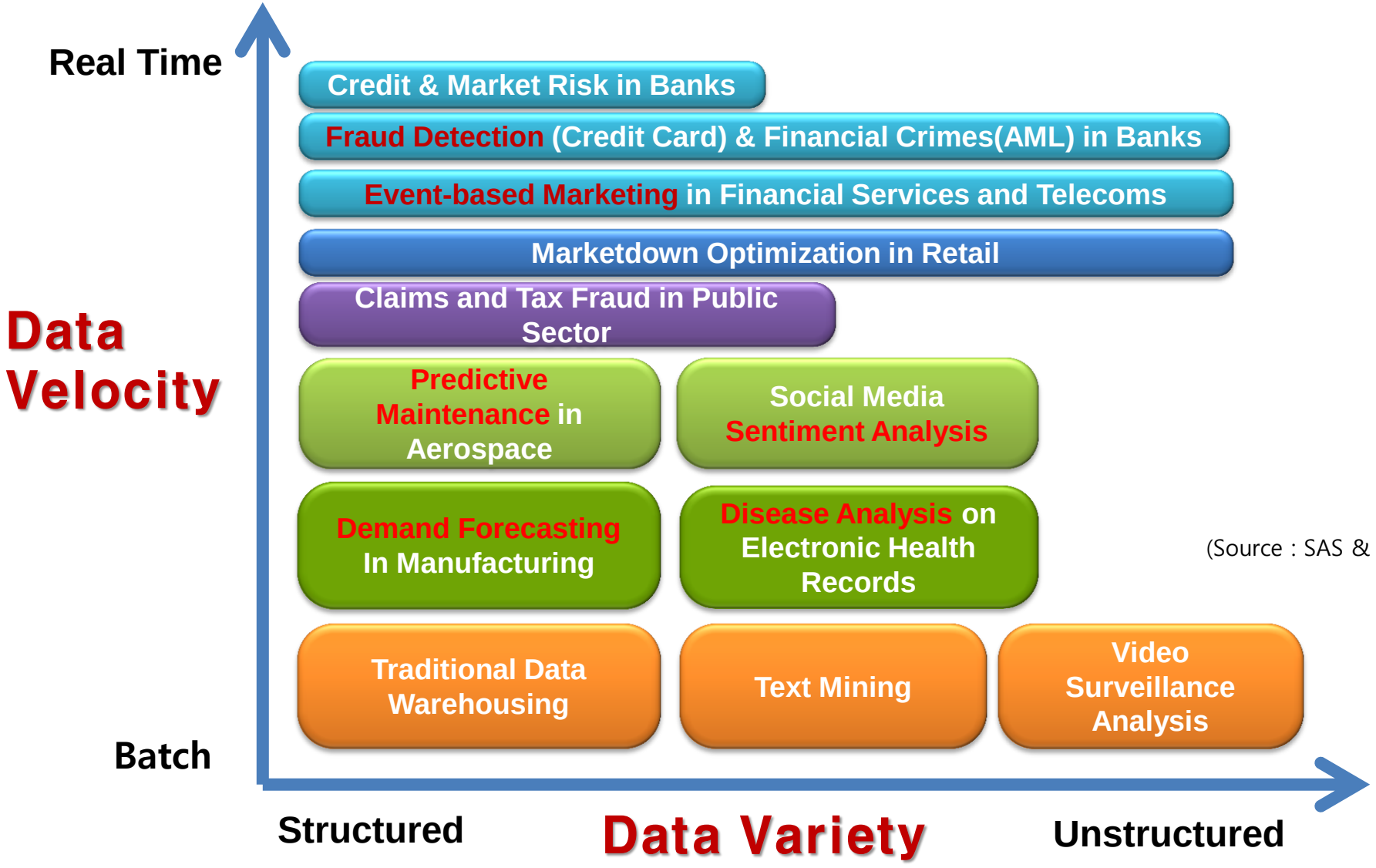






ETRI

BIG DATA POTENTIAL USE CASES



(Source : SAS & IDC)

BIG DATA – CAPTURING ITS VALUE

More than double the total annual health care spending in Spain

More than GDP of Greece

US health care
- $300 billion value per year
- ~0.7 percent annual productivity growth

Europe public sector administration
- €250 billion value per year
- ~0.5 percent annual productivity growth

Global personal location data
- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

US retail
- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

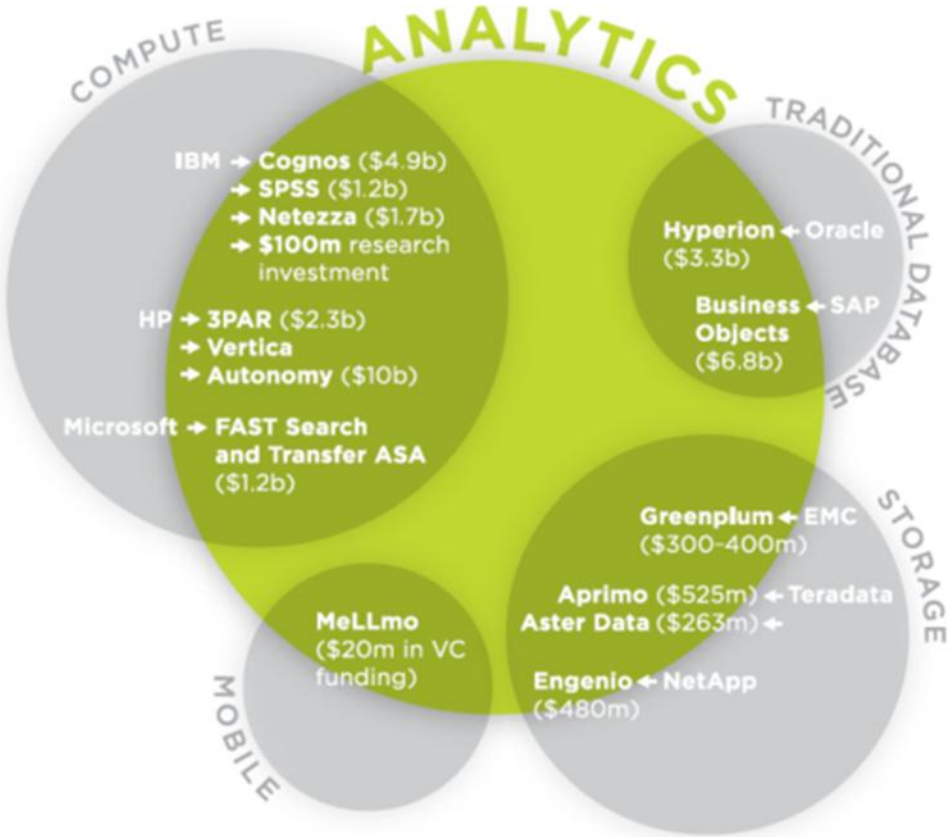
Manufacturing
- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

SOURCE: McKinsey Global Institute analysis

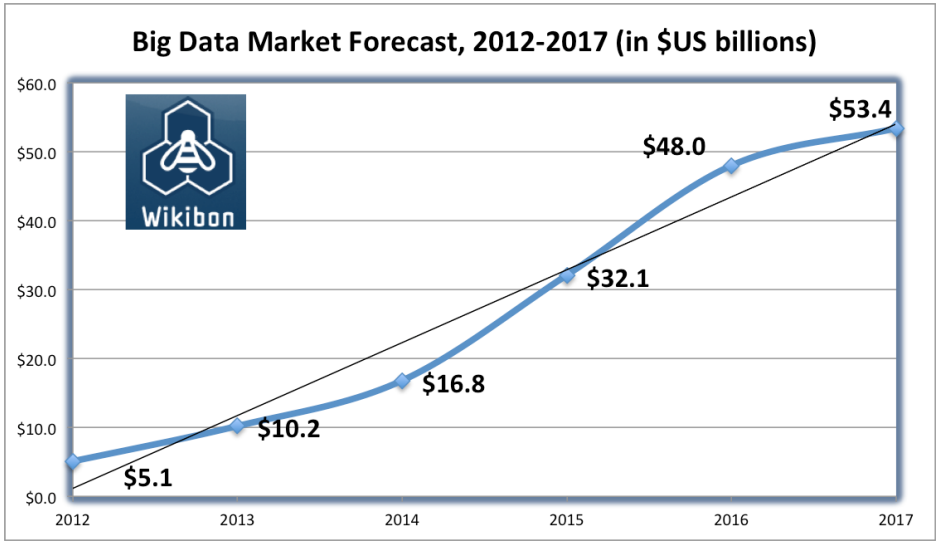
BIG DATA MARKET TREND

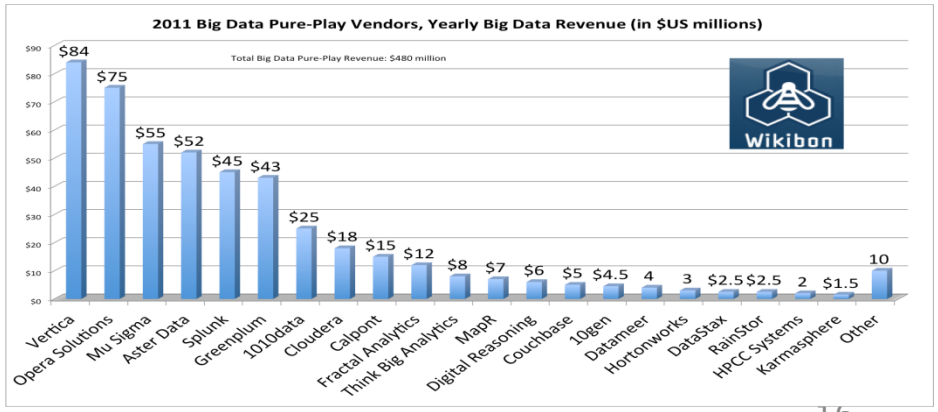
Oracle, IBM, Microsoft, SAP spent more than $15B to acquire Data Management & Analytics firms

o Market Size (Wikibon)
$5B(2012) ➔ $53.4B(2017) , CAGR 58%

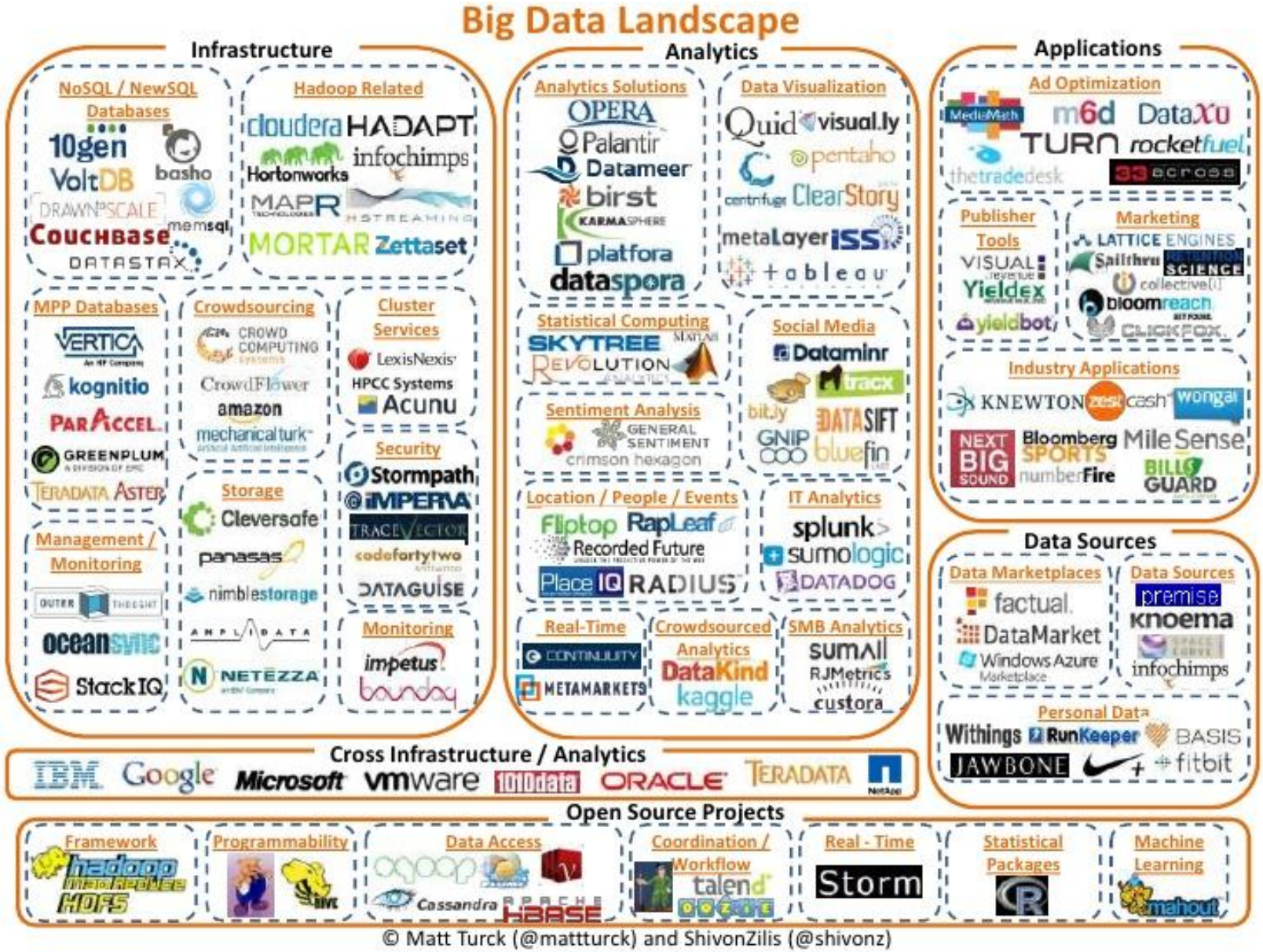


Market acquisitions and funding point to data analytics
(Source : CSC)



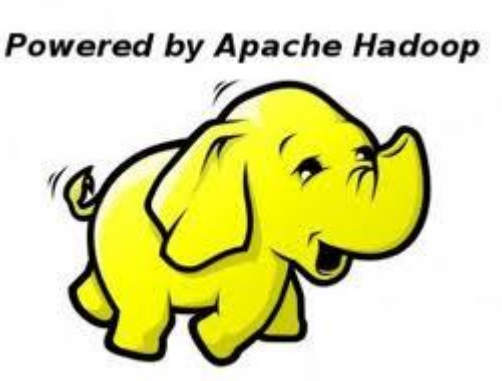


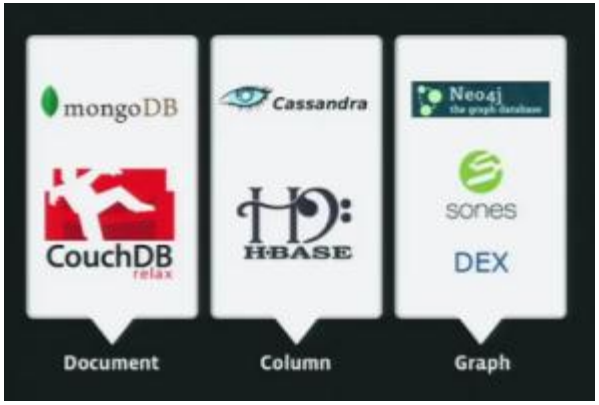
BIG DATA ECOSYSTEM



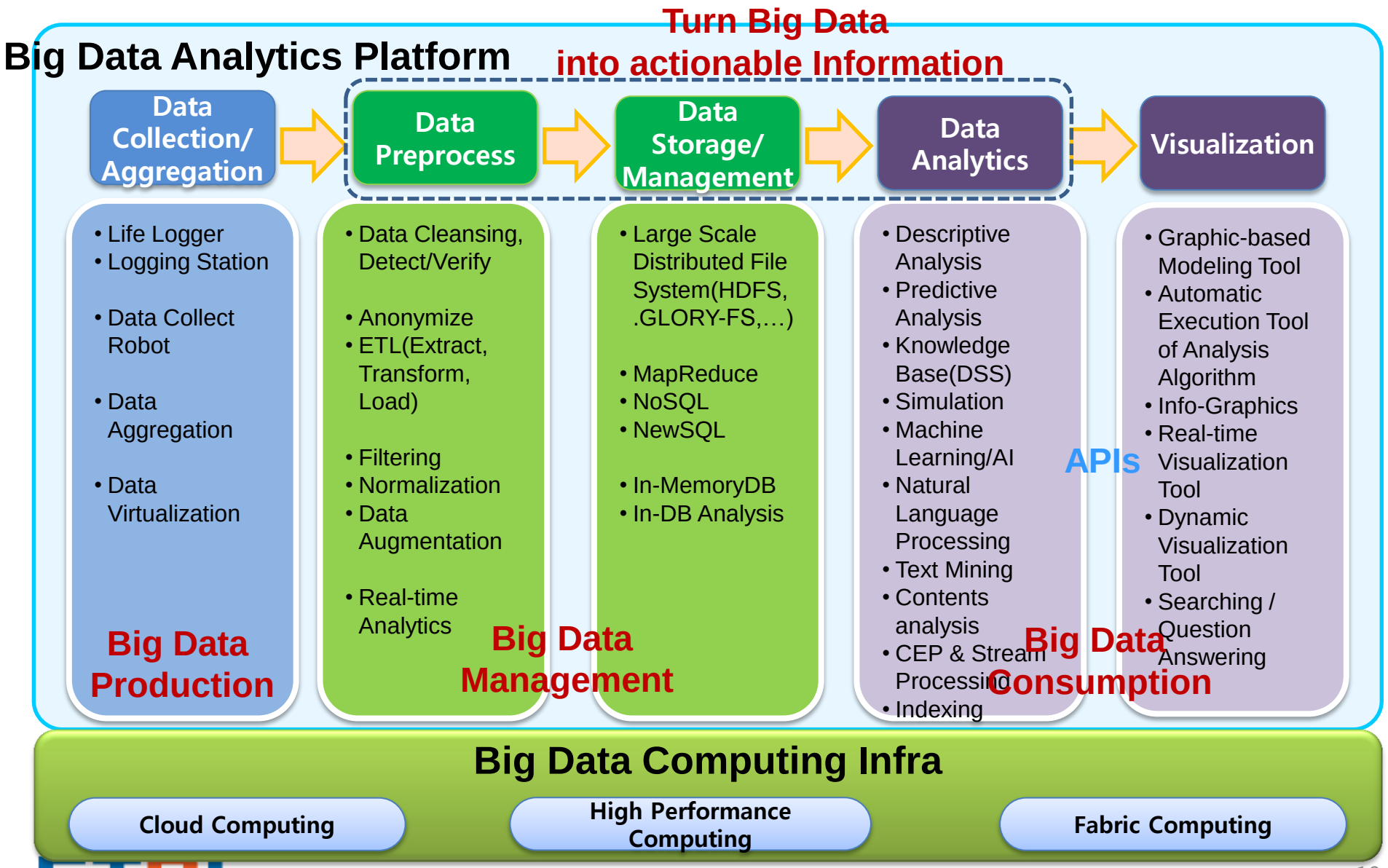
© Matt Turck (@mattturck) and ShivonZilis (@shivonz)

II. BIG DATA TECHNOLOGY





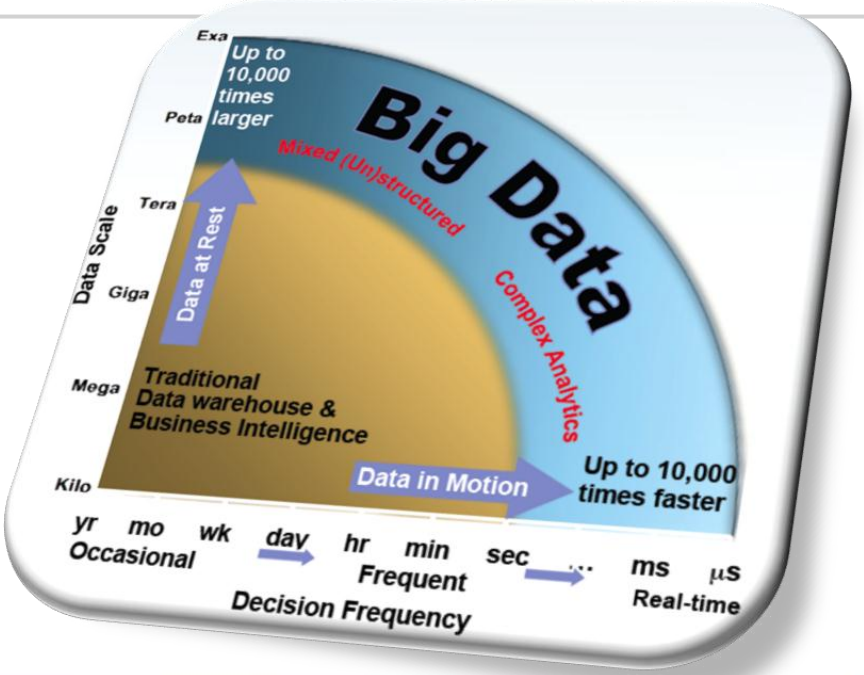
BIG DATA PLATFORM



BIG DATA ANALYTICS

Analytics is the discovery and communication of meaningful patterns in data.

- ***The quantitative change has begun to make a qualitative difference!***

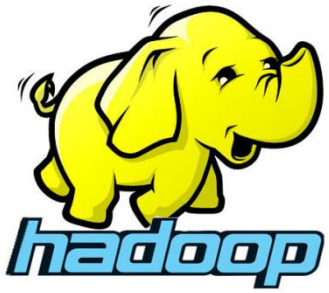


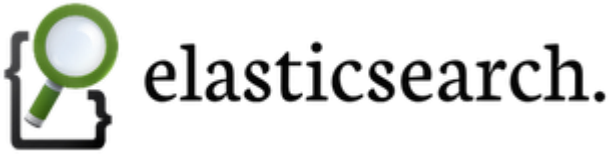
Source: VoltDB, Inc.

Data Age

Interactive	Real time Analytics	Record Lookup	Historical Analytics	Exploratory Analytics
Milliseconds	*Hundredths of seconds*	*Second(s)*	*Minutes*	*Hours*
. Place trade . Serve ad . Enrich stream . Examine packet . Approve trans.	. Calculate risk . Leaderboard . Aggregate . Count	. Retrieve click stream . Show orders	. Backtest algo . BI . Daily reports	. Algo discovery . Log analysis . Fraud pattern match

OPEN SOURCE BIG DATA TECHNOLOGIES





OPEN SOURCE BIG DATA TECHNOLOGIES

R is an **open source programming language** and software environment designed for statistical computing and visualization.

cascading

An open source **software abstraction layer for Hadoop**, Cascading allows users **to create and execute data processing workflows** on Hadoop clusters using any JVM-based language.

SCRIBE

Scribe is **a server developed by Facebook** and released in 2008. It is intended for **aggregating log data streamed in real time** from a large number of servers.

elasticsearch.

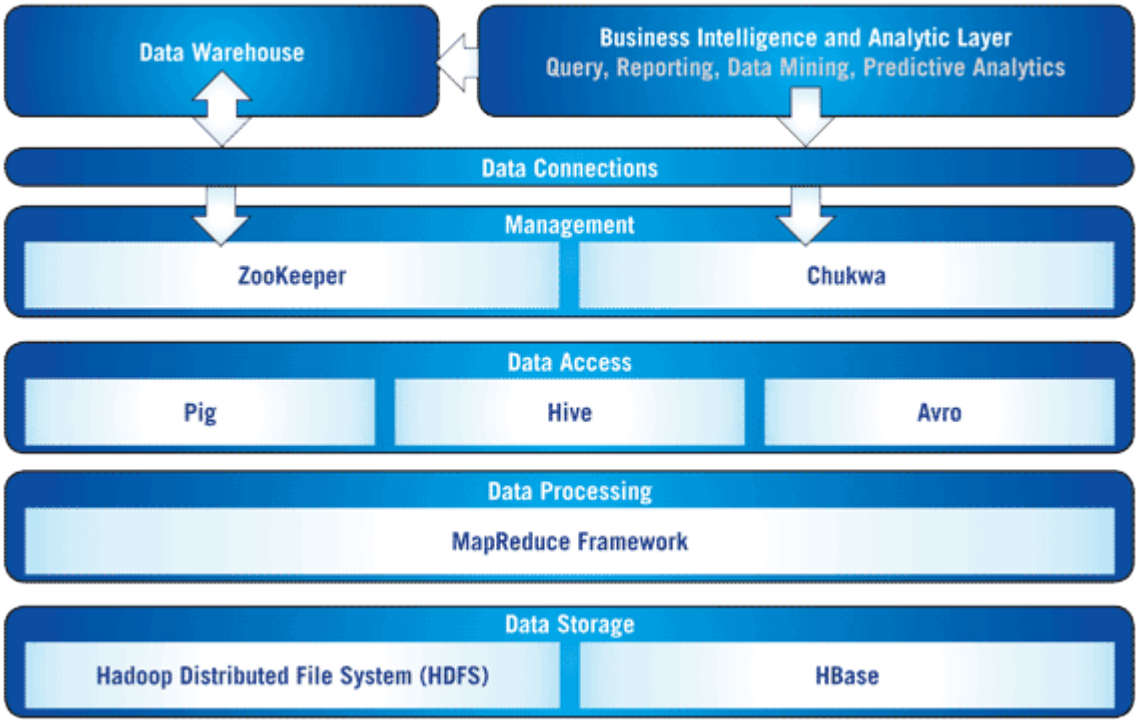
ElasticSearch is a distributed, RESTful open source **search server**. It's **a scalable solution that supports near real-time search and multitenancy** without a special configuration.

ETRI

Hadoop Ecosystem

Hadoop is a free, **Java-based programming framework that supports the processing of large data sets in a distributed computing environment**. It is part of the **Apache** project sponsored by the Apache Software Foundation.

FIGURE 1: THE HADOOP SYSTEM



(출처 : Information Management)

HDFS : Hadoop Distributed File System (Creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, extremely rapid computations)

MapReduce : A programming model and software framework

HBase : Scalable, column-oriented distributed database (NoSQL)

Pig : High-level programming language and execution framework for parallel computation

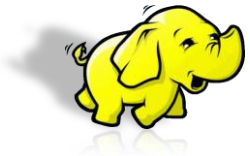
Hive : Data warehouse infrastructure that provides ad hoc query and data summarization

Avro : Data serialization system

ZooKeeper : coordination, configuration and group services for distributed applications

Chukwa : Large-scale monitoring system

CASE STUDY : YAHOO! HOMEPAGE

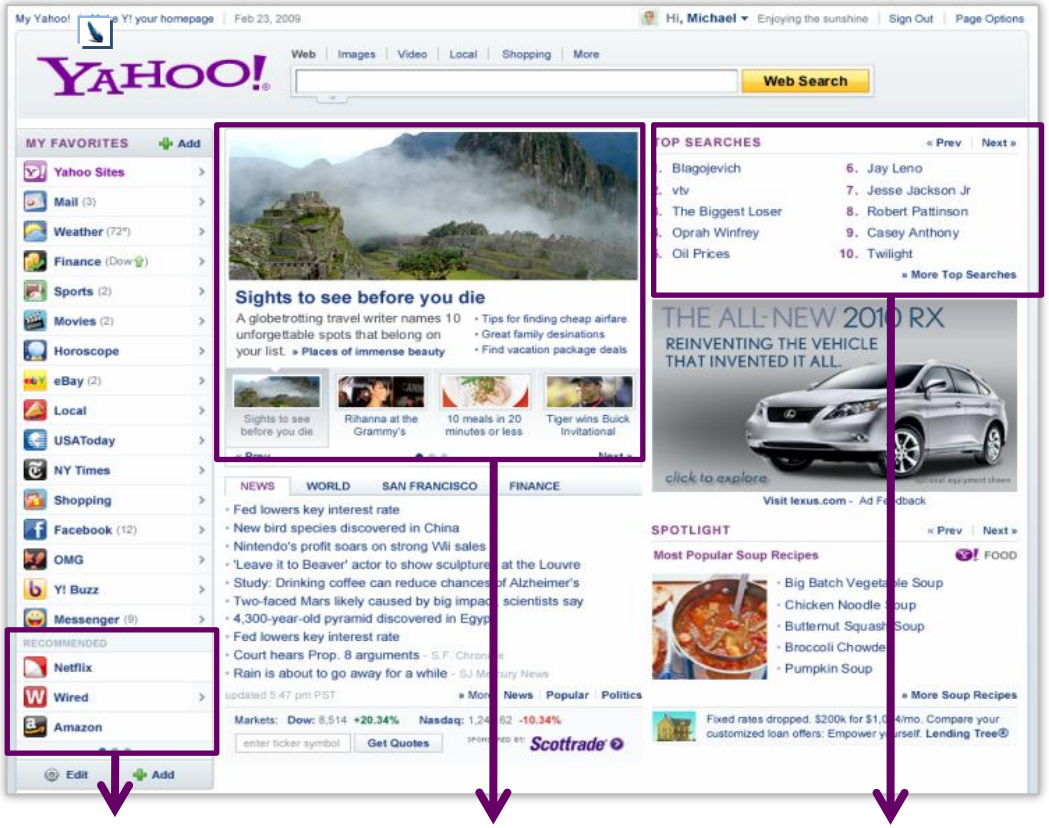


Personalized

for each visitor

Result:

twice the engagement



+79% clicks
vs. randomly selected

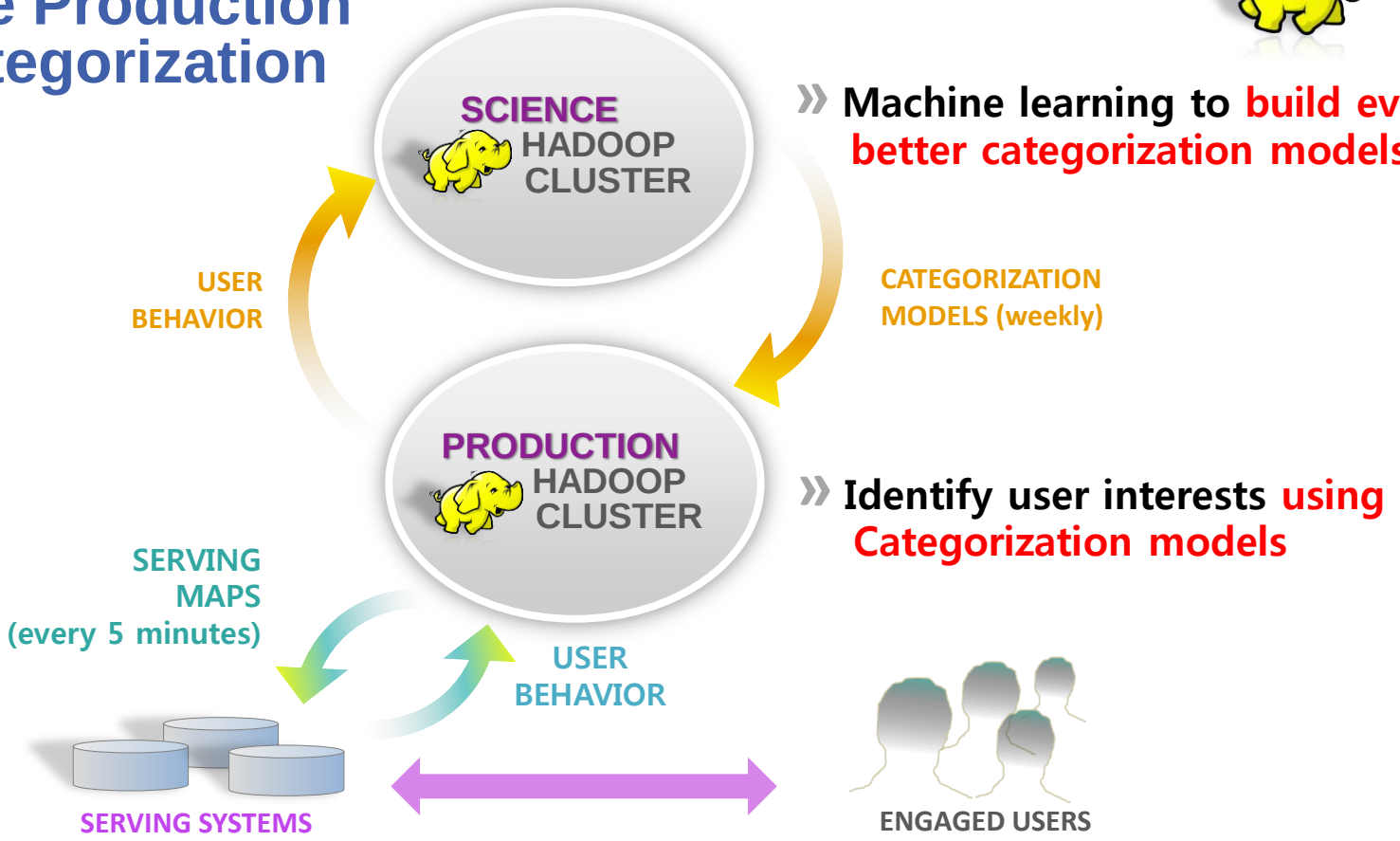
+160% clicks
vs. one size fits all

+43% clicks
vs. editor selected

(Source : Yahoo 2011)

CASE STUDY : YAHOO! HOMEPAGE

- Serving Maps
  - Users - Interests
- Five Minute Production
- Weekly Categorization models



» Machine learning to build ever better categorization models

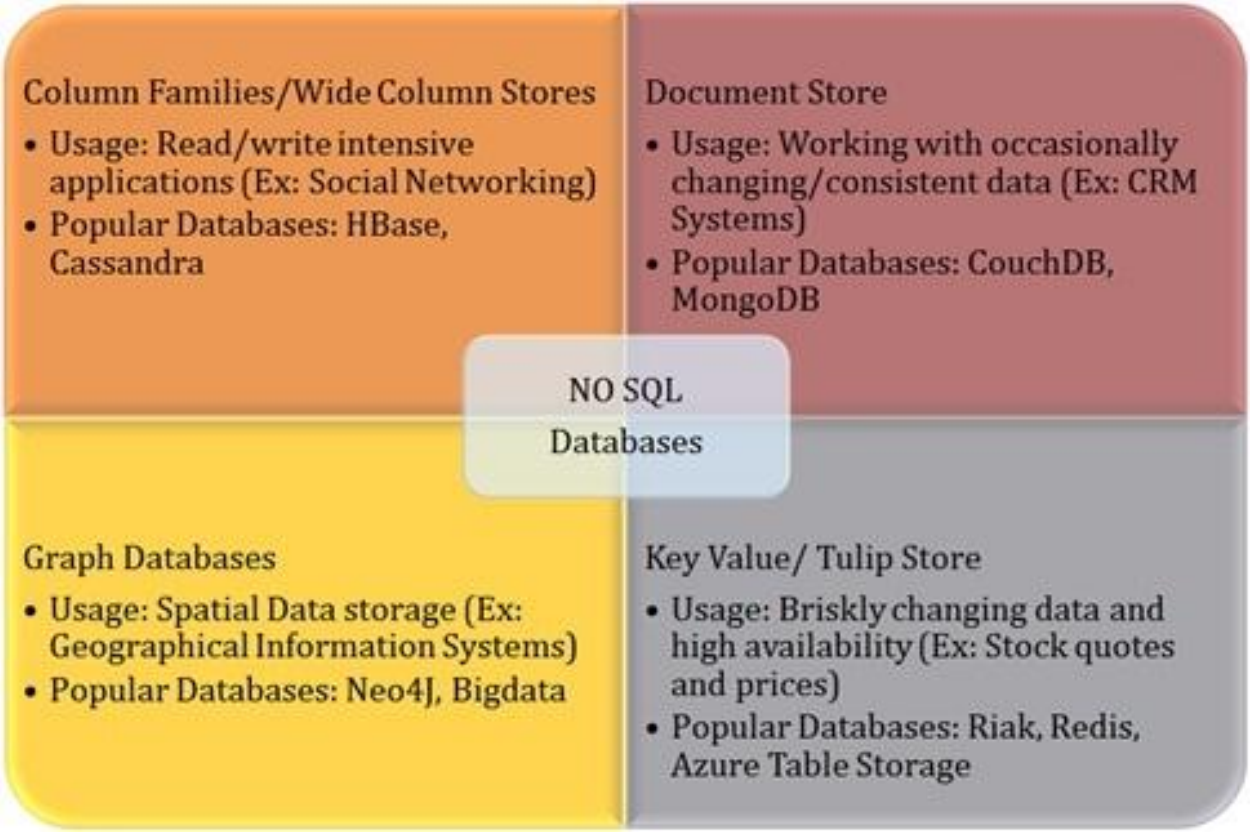
» Identify user interests using Categorization models

Build customized home pages with latest data (thousands / second)

(Source : Yahoo 2011)

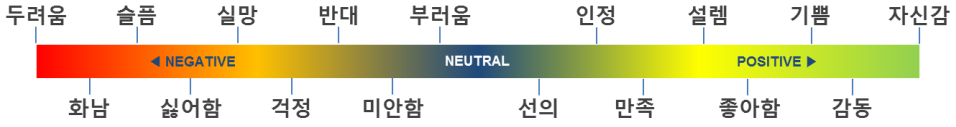
NOSQL

No SQL databases are the data stores that are **non-relational (without any fixed schemas & joins), distributed, horizontally scalable** and often **don't adhere to the principles (ACID: atomicity, consistency, isolation, durability)** of traditional relational databases.



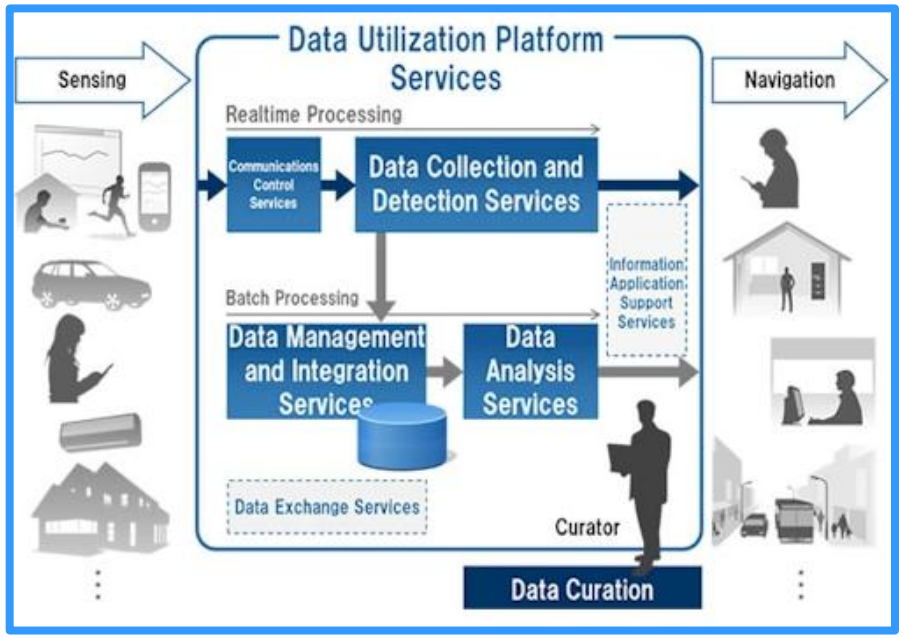
http://www.codeproject.com/Articles/279947/Migration-of-Relational-Data-structure-to-Cassandr

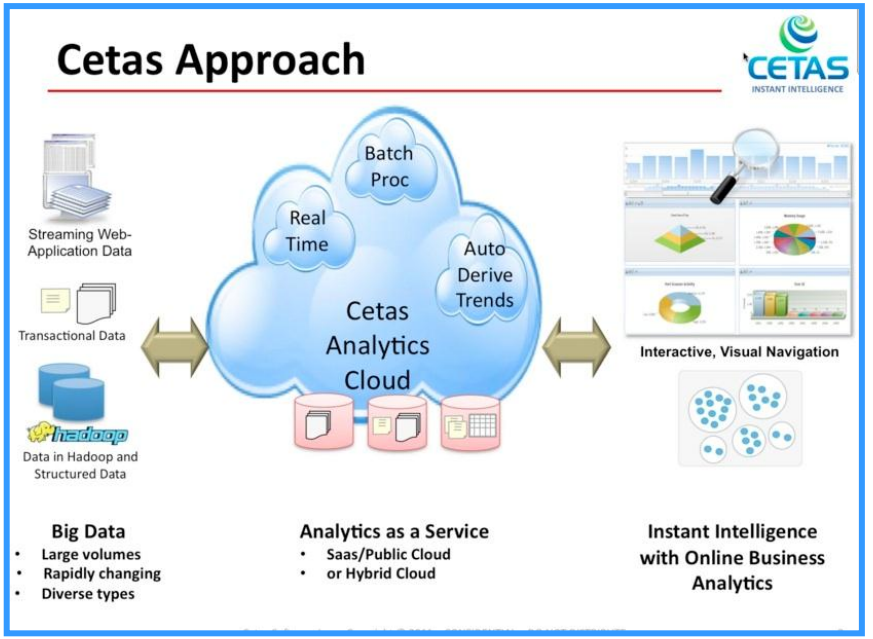
SOCIAL MEDIA TEXT MINING



BIG DATA IN THE CLOUD

Cloud analytics is a service model in which elements of the data analytics process are provided through a public or private cloud.



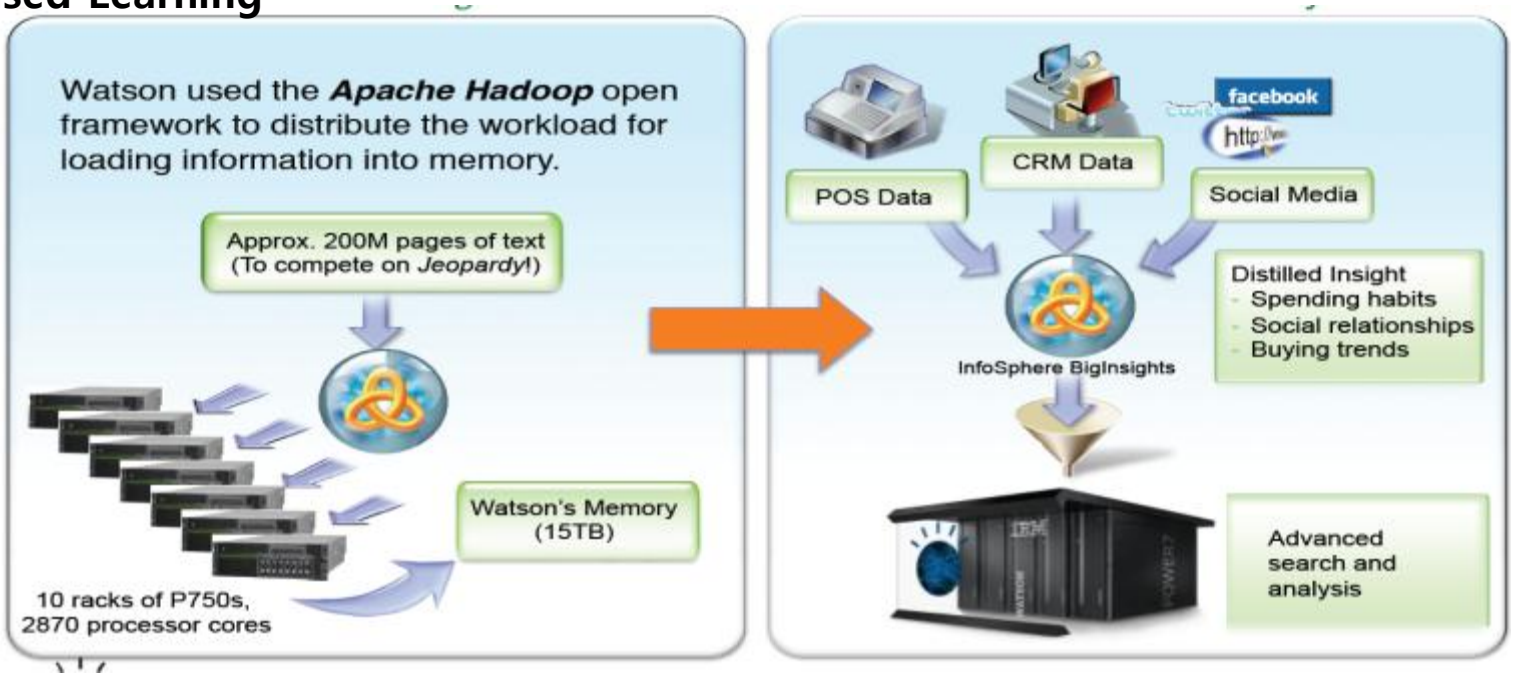


INTELLIGENCE

o Case Study : IBM Watson

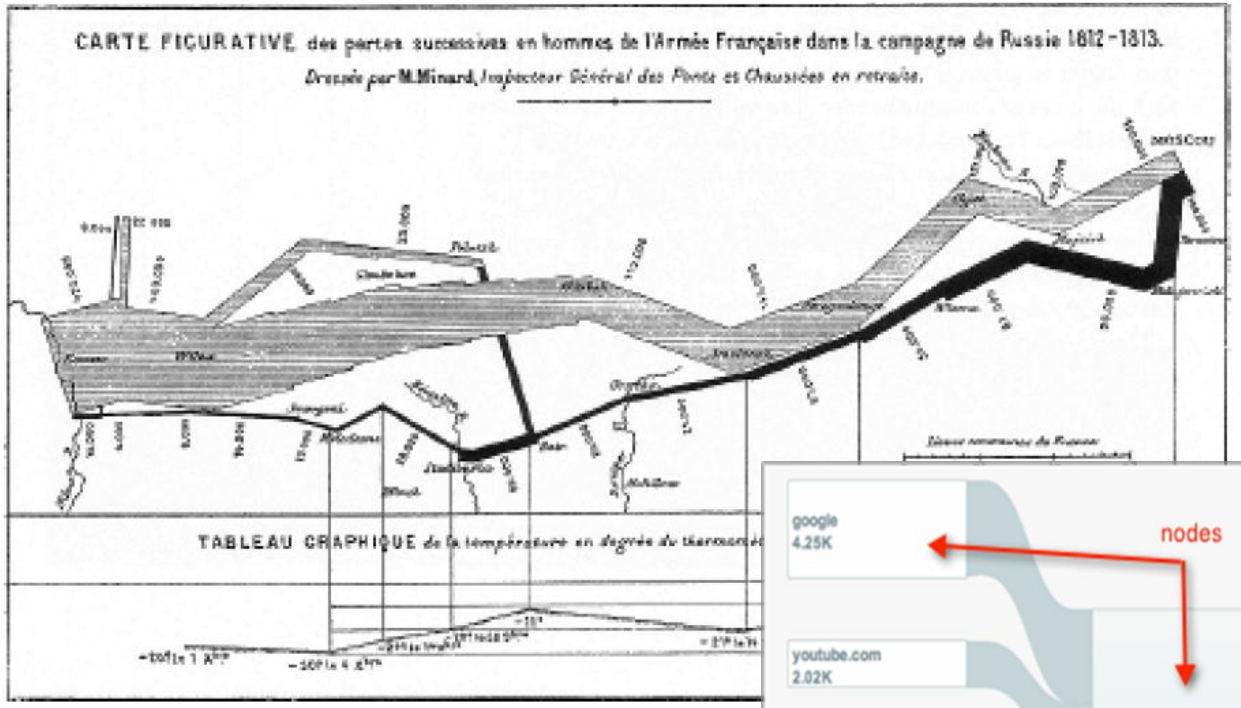
①High Performance Computing, ②Big Data Analytics, ③Artificial Intelligence

- Key Capabilities : Natural Language Processing, Hypothesis Generation, Evidence-based Learning



- **Watson in finance** : To assist financial services professionals in consuming and analyzing available data to improve decision making
- **Watson in healthcare** : To help physicians to make evidence based diagnosis and treatment decisions to reduce medication errors

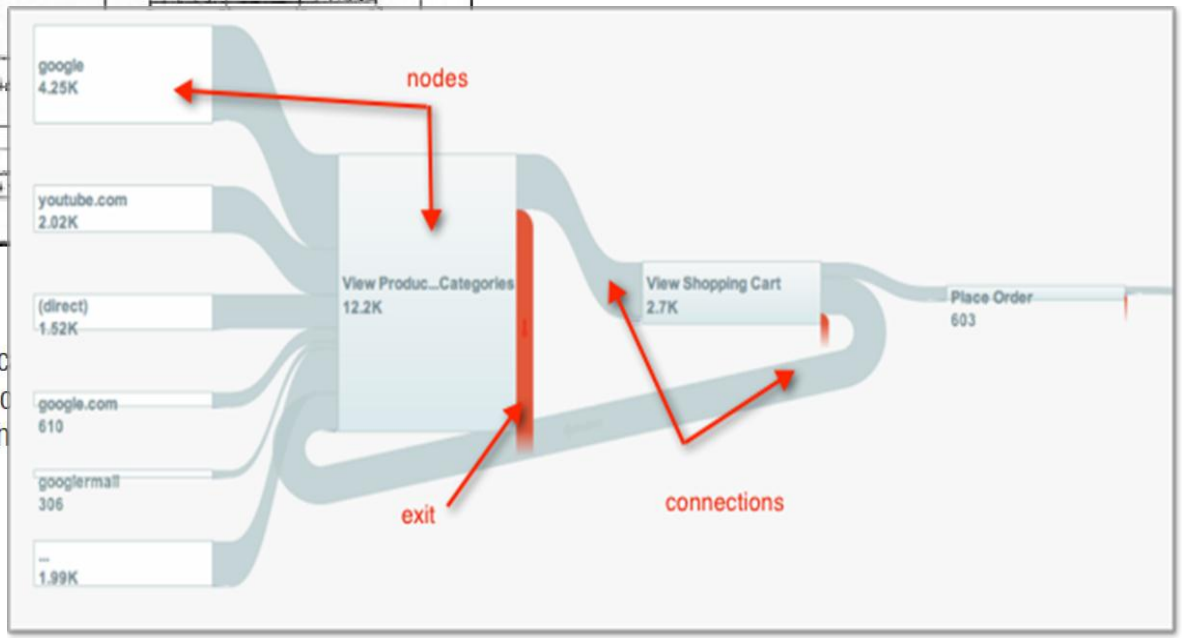
VISUALIZATION & VISUAL ANALYTICS



1869

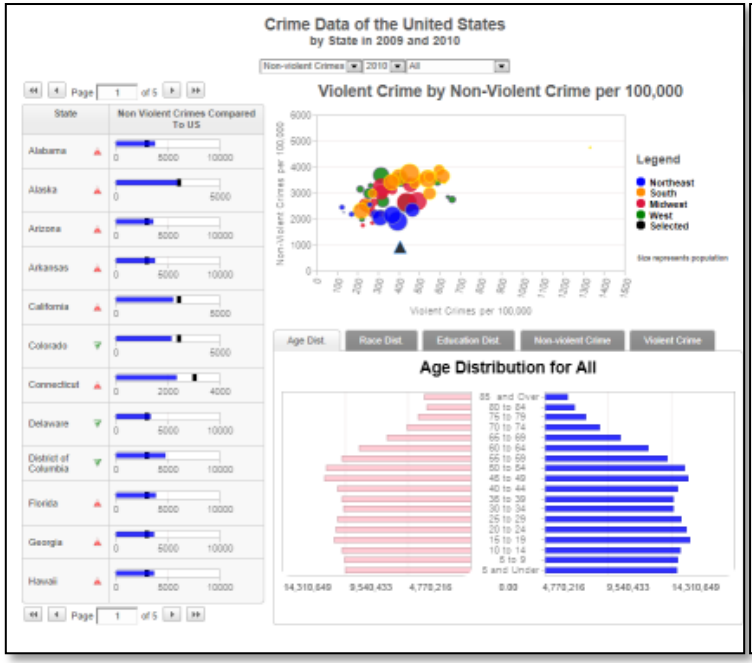
The best graphic ever made?

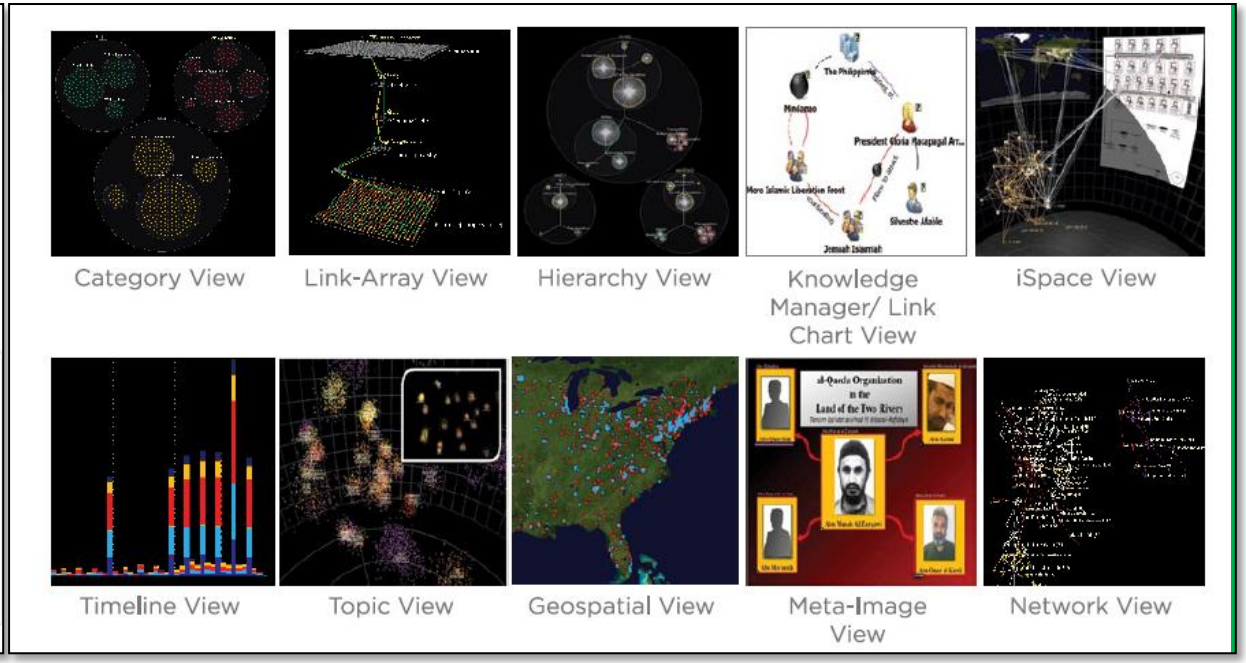
The French engineer, Charles Minardi (1781-1870), illustrated graphic
of Napoleon against Russia in 1812. The width of the course is propo
surviving soldiers in the war campaign. In beige for the way in and in



Google

VISUALIZATION & VISUAL ANALYTICS



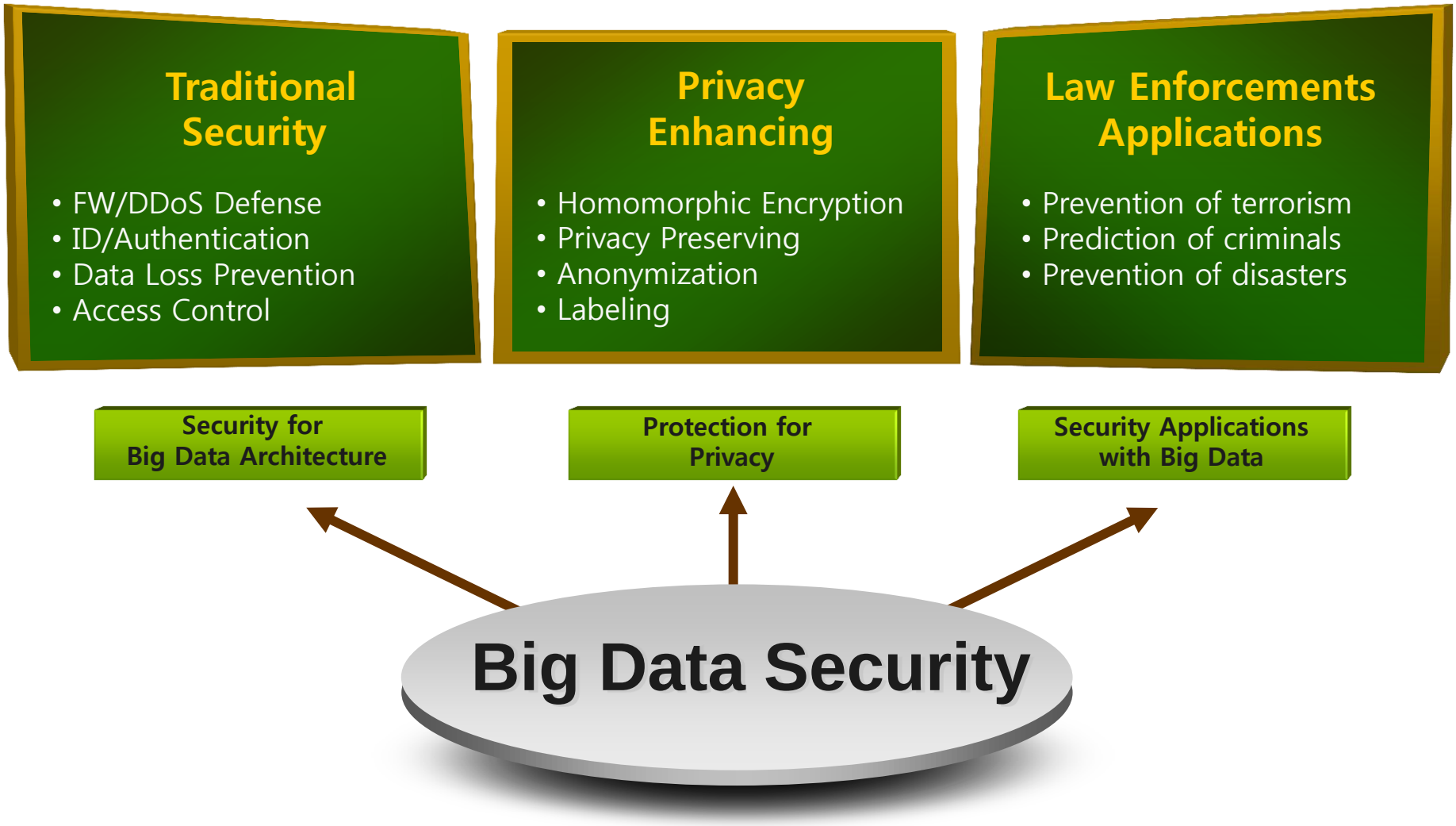


(Source : LogiXML)

"The essence of information visualization is to accelerate human thinking with tools that amplify human intelligence" – Ben Shneiderman

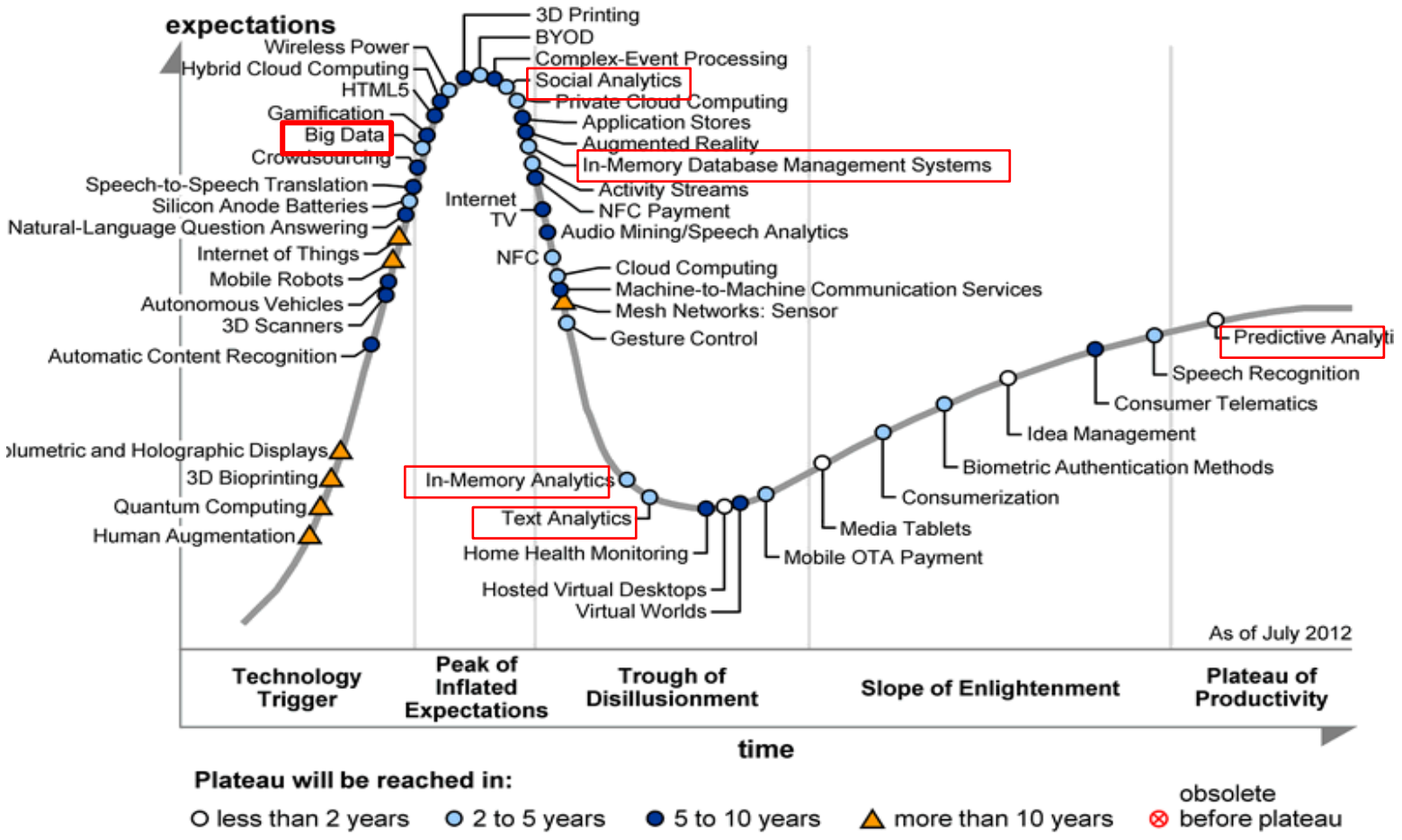
Visual Analytics : Analytical reasoning supported by highly interactive visual interfaces

BIG DATA SECURITY/PRIVACY



III. RESEARCH DIRECTION OF ETRI

BIG DATA : EMERGING TECHNOLOGY



* Source : Gartner(2012) “Hype Cycle for Emerging Technologies, 2012”

BIG DATA TECHNOLOGY CHALLENGES

Volume : High-Performance Processing of Massive Datasets
- Computing, Storage, Distributed Data Cloud

o Sharing and Scalability among Clouds : **"Distributed Cloud"**
o Resource Service Cloud with Linked Data : **"Data Cloud"**

Velocity : Real-time, Streaming Data Processing

o Many-core, **In-memory Computing** : **"High-efficient System SW"**
o New Architecture with Scalable Resources : **"Fabric Computing System"**

"The leading edge of Big Data is streaming data"
- The Data Warehouse Institute

Variety : Intelligence – Intelligent Data management and Complex Analysis

o Unified Management of Various Types of Data : **"Unified DBMS(New SQL)"**
o Complex Analytics with Beyond human level Intelligence : **"AI based Analytics/Forecasting "**

BIG DATA TECHNOLOGY VISION

Smart Revolution enabled by Big Data Platform Technology

Super Convergence

- Smart Grid
- Natural Resource
- Electronic Vehicles
- Healthcare Medical
- Insight/ Forecasting
- Personal Service

Stream Computing

Security Safety

(3) Big Data SW Intelligence

- Data Collection/ Management
- Data Anaysis
- Modelling/ Simulation

Reality

Knowledge Visualization

Intelligent Computing

Scale

Performance

(1) Cloud Computing

- Data Virtualization
- Distributed Computing

(2) High-Performance Computing

- Fablic Computing
- Manicore/In-Memory Computing

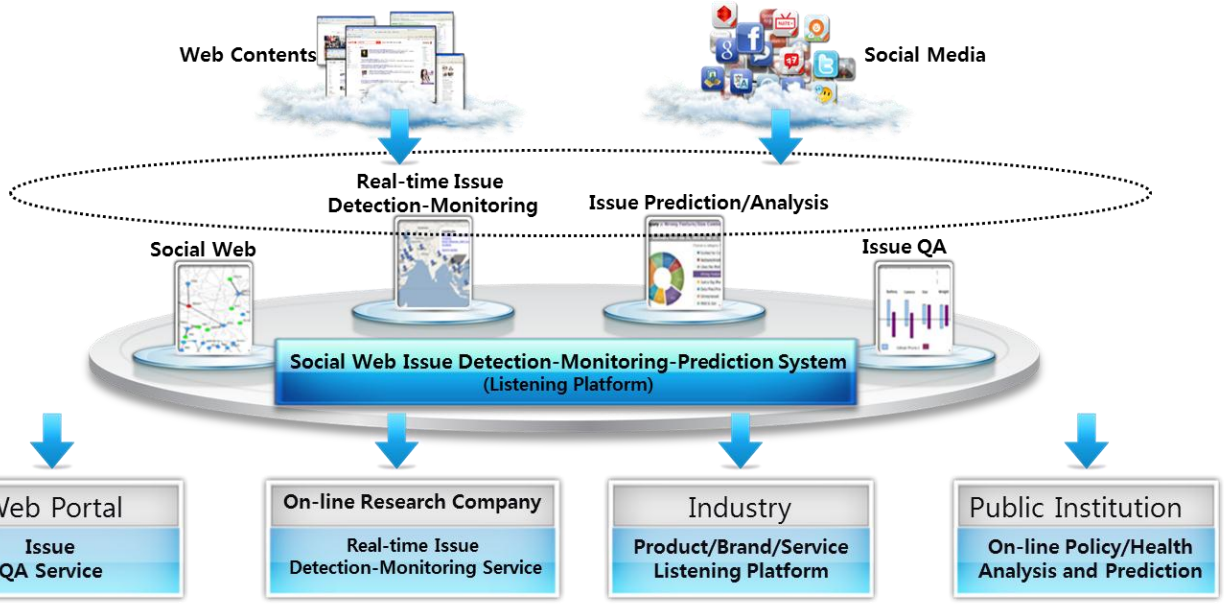
ETRI R&D – Social Big Data Analytics

❑ Social Big Data Analytics

- The Social Big Data Analytics **analyzes and predicts the spread of issues with the passage of time through social media**
- **To deliver 'Hidden Insight' to human experts for data-driven decision making**

❑ Research Issues

- **Text Analytics**: Natural Language Processing, Fine-grained Opinion Mining and Event & Temporal Information Extraction
- **Social Web Analytics**: Detection & Monitoring of Complex Social Issues
- **Predictive Analytics for Social Web**: Predictive Models of Human Interaction



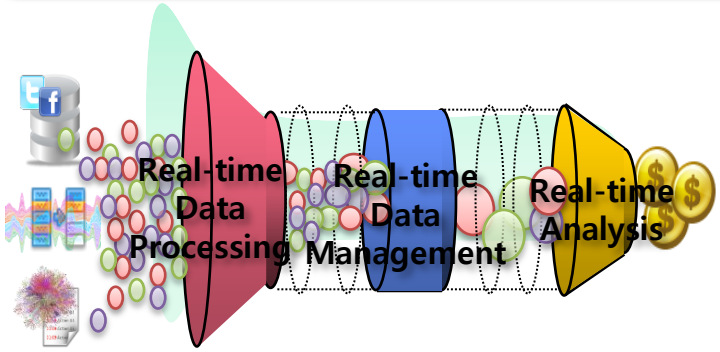
ETRI R&D – Distributed In-Memory Data Management Platform

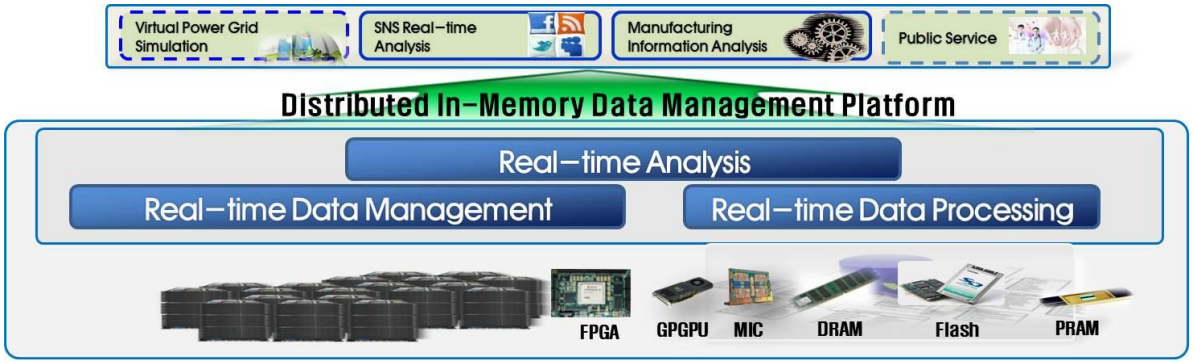
❑ Distributed In-Memory Data Management Platform

- In-Memory Data Processing and Management Technology providing real-time processing and high scalability based on heterogeneous memories(Volatile, Nonvolatile)
- To provide extreme transaction processing and real-time analysis environment

❑ Research Issues

- **Distributed Stream Processing :** Programming Model for Processing Various Exploding Stream on Memory
- **High Speed Stream Processing :** Processing on HW Accelerator (FPGA, GPGPU)
- **High Speed and Low Power DBMS :** Data management on Heterogeneous Memory (DRAM, NVRAM)
- **Distributed Data Management** : In-Memory Data Distribution & Replication based on key-value model





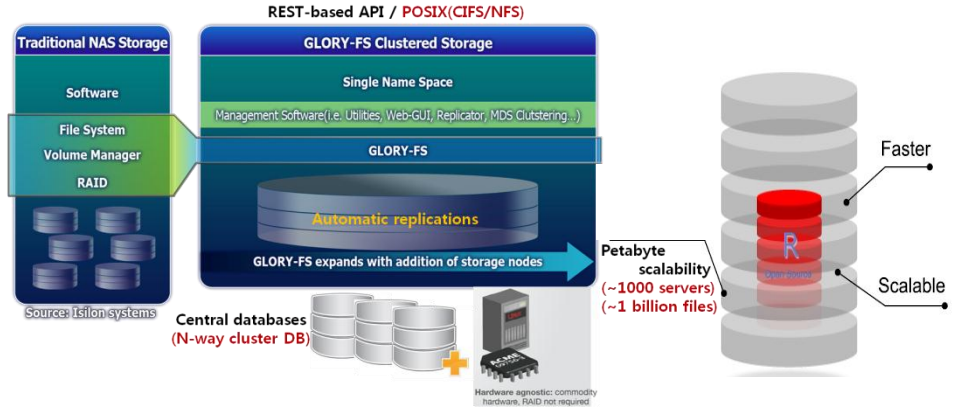
ETRI R&D – Big Data Distributed File System

❑ High-scalable/High-performance Cloud File System - “GLORY-FS”

- To develop petabyte-scale **big storage SW technology** using commodity HW, being well-suited to cloud computing
- To provide unlimited scalability in capacity, performance, reliability of data

❑ Research Issues

- **Petabyte-scale Storage Capacity** : Out-of-band virtualization architecture connecting more than 1,000 storage nodes
- **Commodity Storage HW Fault-resiliency** : Data replication/Parallelized self-recovery
- **World No.1 Performance** : 50,000ops/s(glory-fs) vs. 15,000ops/s(Oracle Lustre)
- **Big Data Lifecycle Management** : Tiered storage(DRAM > SSD > HDD)
- **Analytical app. as well as Online applications** : Unified cloud storage system



ETRI R&D – MAny-core HPC System for Bio Application

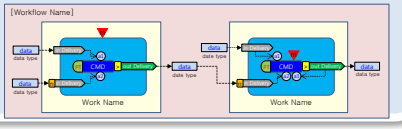
MAHA

- Develop key technologies for scalable accelerator based PetaFLOPS Computing System
- Develop computing system optimized for human genome analysis

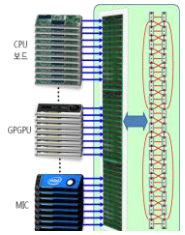
Research Issues

- **Heterogeneous Computing HW**
  - nVIDIA GPGPU + Intel MIC based
  - High speed, low-latency network : PCIeLink 128/256 Gbps

- **High Speed and Low-power IO File System**
  - SSD+HDD file system : Max 700 Gbps/ 1M IOPS
  - 40 SSD storage servers (equal to 600 HDD servers)
- **System Management SW**
  - Heterogeneous, optimal resource management(CPU/GPGPU/MIC) for Bio workflow SW
- **Fast Human Genome Analysis Application**
  - Parallelized genome sequence mapping, protein docking analysis

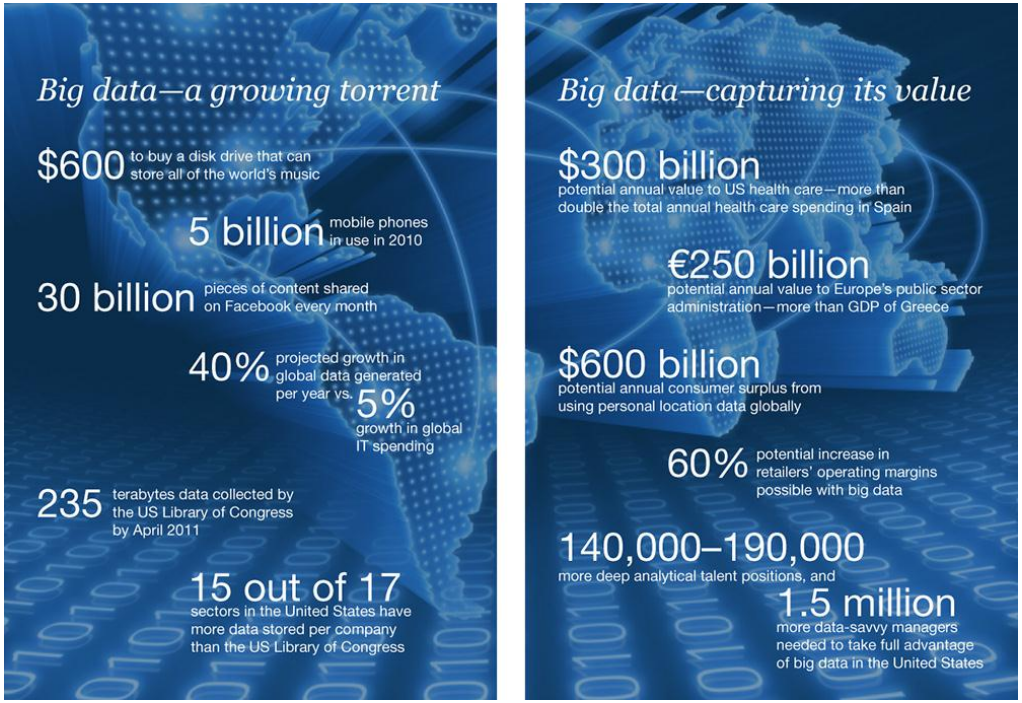


Many core and Accelerator-based Computing for Genomic Data Analysis



High speed system network

IV. CONCLUDING REMARKS



(Source : McKinsey 2011)

CONCLUDING REMARKS

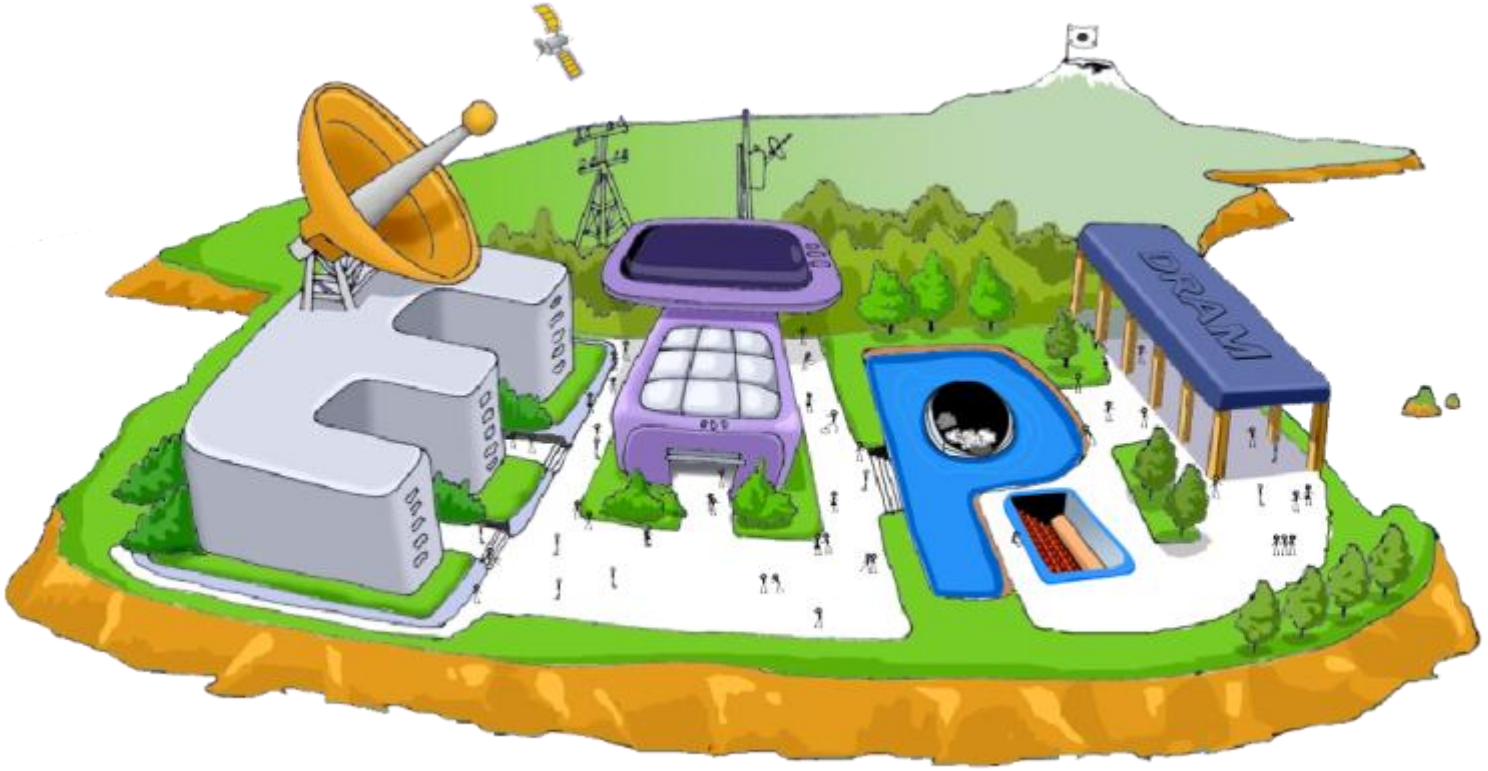
Balanced Data Ecosystem

Data Sovereignty

Technology Innovation

Data Scientist

Data Privacy, Data Trustworthiness,
Protection against misuse



감사합니다 !
Thank you!