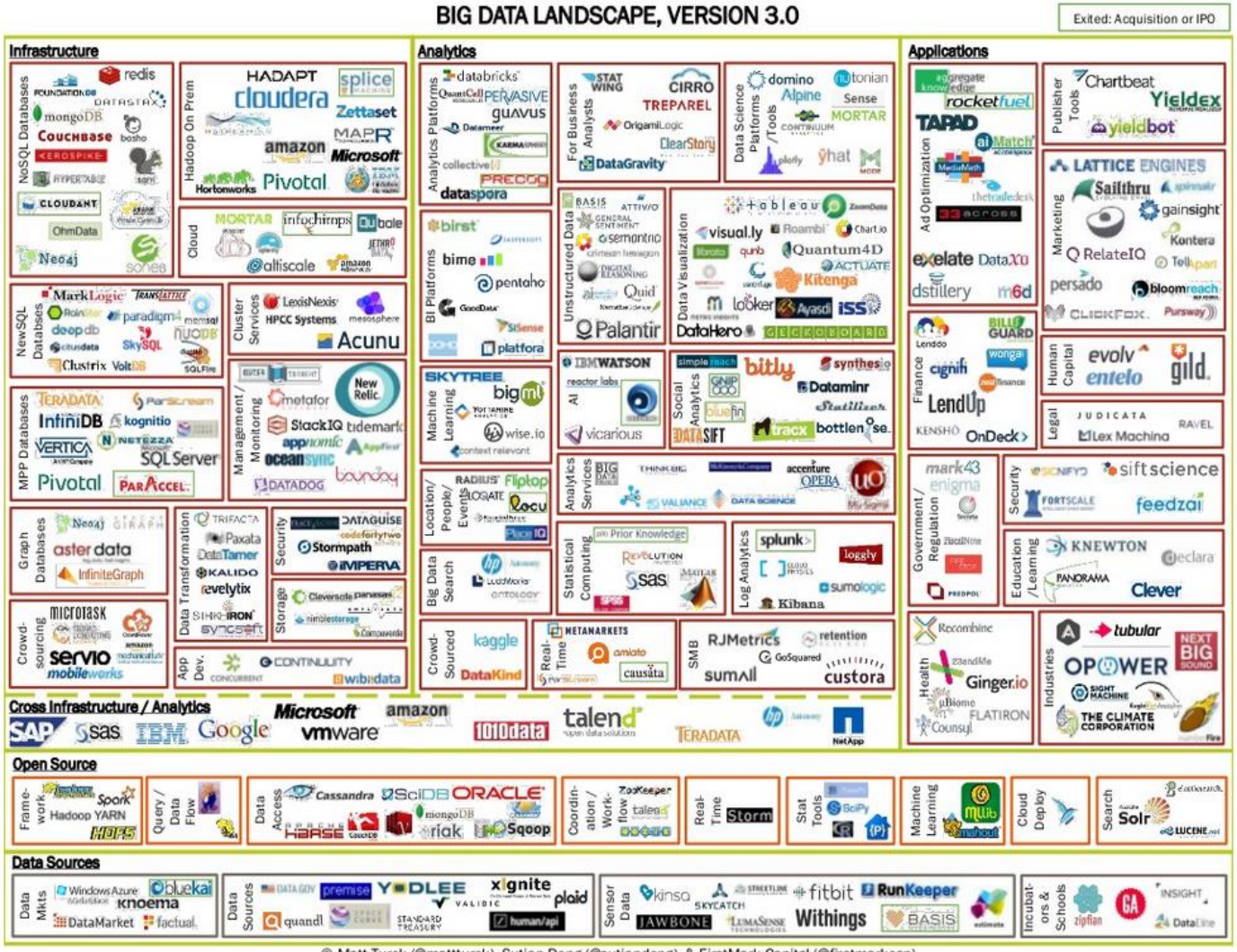
BIG DATA TECHNOLOGIES, USE CASES, AND RESEARCH ISSUES

Il-Yeol Song, Ph.D.
College of Computing & Informatics
Drexel University
Philadelphia, PA 19104

ACM SAC 2015
April 14, 2015
Salamanca, Spain

DREXEL UNIVERSITY
College of
Computing & Informatics



Source: http://dataconomy.com/wp-content/uploads/2014/06/Understanding-Big-Data-The-Ecosystem.png

What is this talk about?

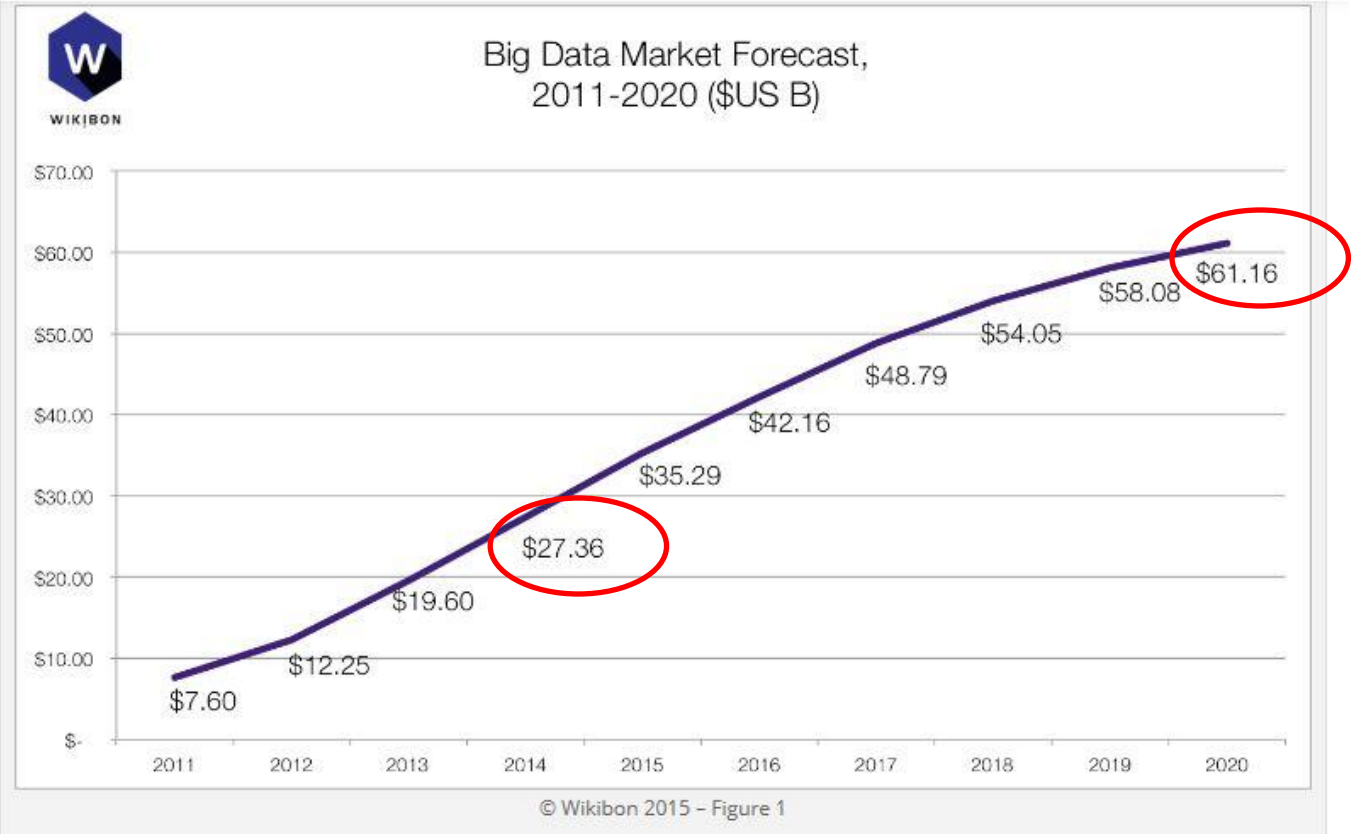
- *An **overview** of big data technologies from a **director's POV***
- *Experiences for over a dozen BD projects as Deputy Director of **NSF Research Center on Visual & Decision Informatics** (CVDI)*
- *PI of a **large-scale smart health project***

Table of Contents

- What is Big Data?
- Data-driven paradigm (DDP)
- Big Data Technologies
  - Hadoop Ecosystems
  - NoSQL databases; NewSQL databases
  - In-memory databases
  - Cloud computing
  - Big data warehousing (ETL, ELT, and Data virtualization, EDW, LDW, Data Integration)
- Values and Use cases
- Research issues
- Conclusions

World Big Data Market (Wikibon, 2015)

(Big Data-related hardware, software, and professional services: **$27.36 billion, 2014)**



Source: http://premium.wikibon.com/executive-summary-big-data-vendor-revenue-and-market-forecast-2011-2020/

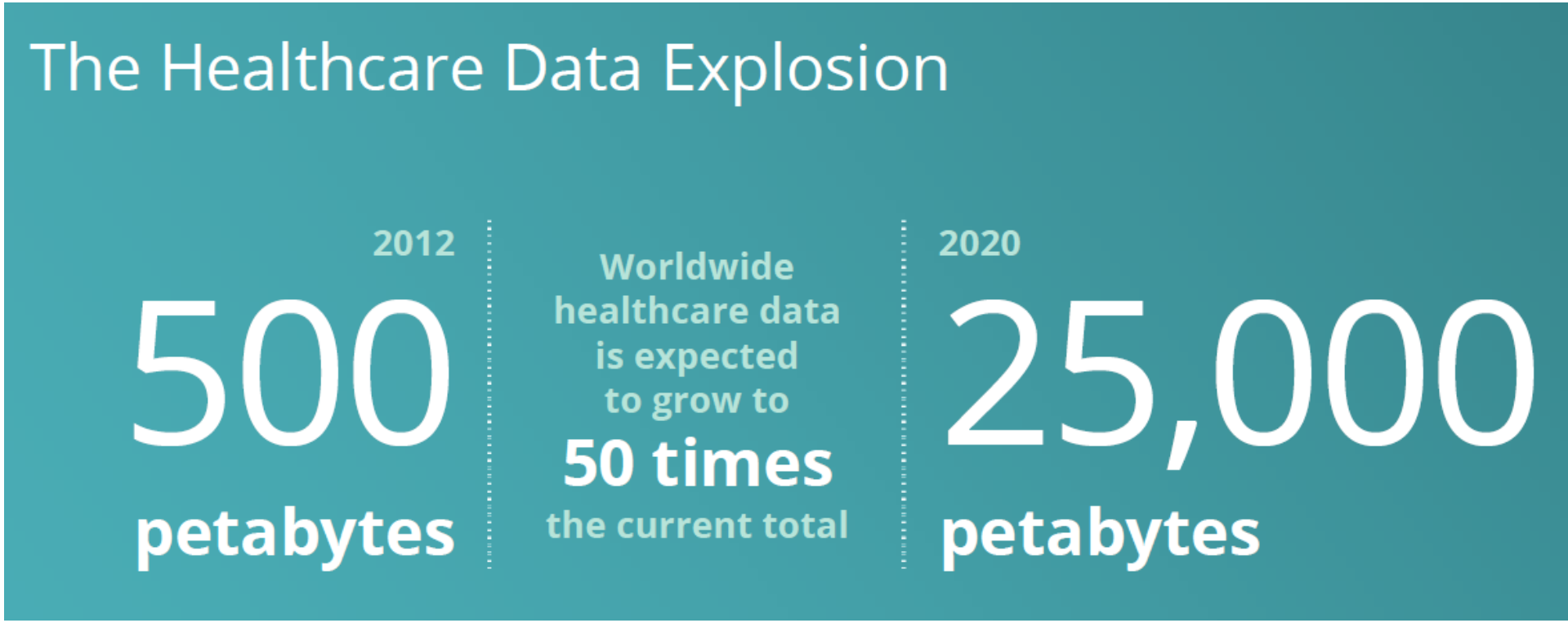
Why Big Data?

73 percent of organizations have invested or plan to invest in big data in the next two years (Gartner, 2014)

90% of the data in the world today has been created in the last two years alone.

("If you think Big Data's Big Now, Just Wait", Ron Miller, TechTrends, 2014)

Example: Healthcare Data



Source: "Big Data in Healthcare Hype and Hope", (Feldman, Martin, and Skotnes, Dr. Bonnie 360^0, Oct. 2012)

Example: Healthcare Data

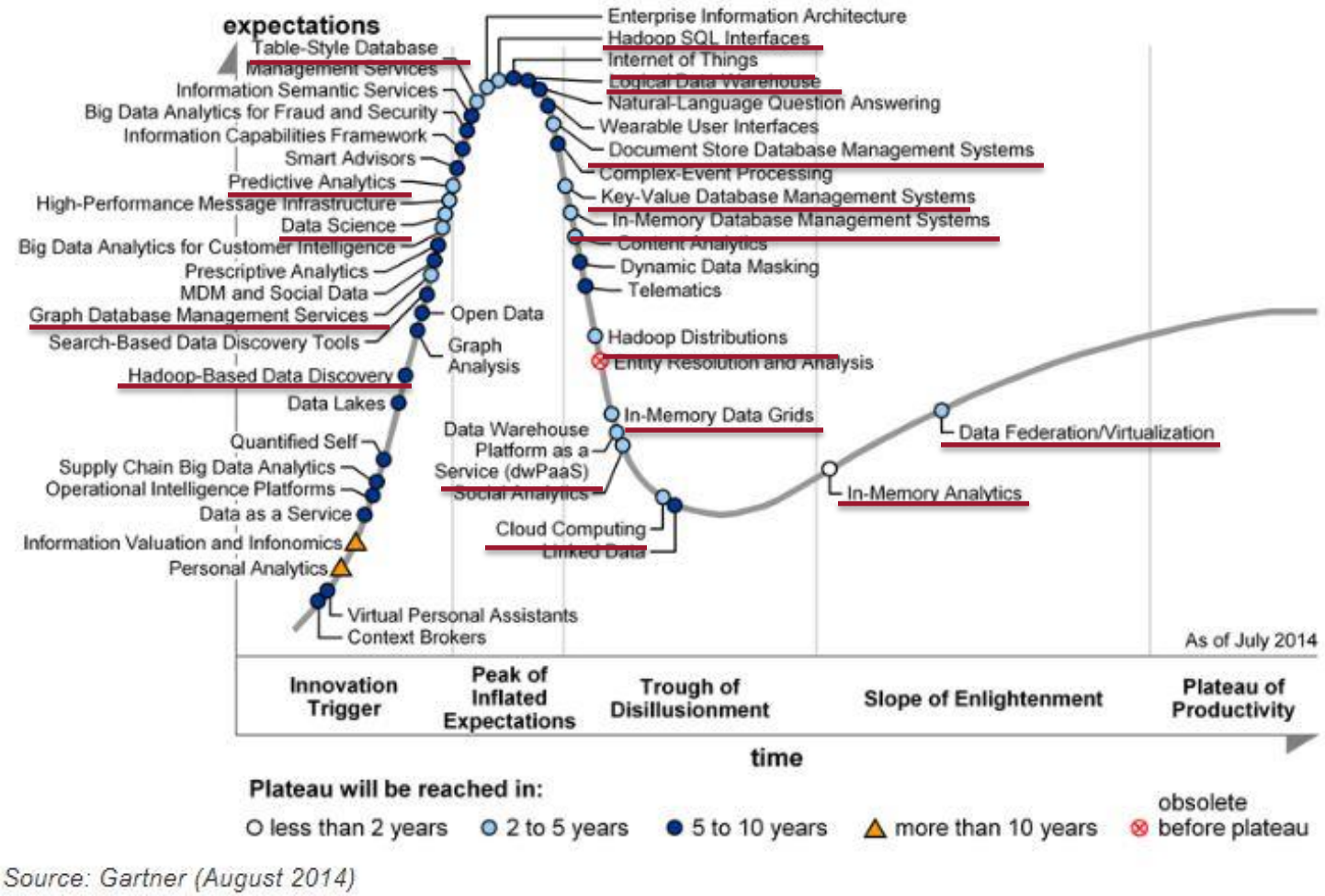
Source: "Big Data in Healthcare Hype and Hope", (Feldman, Martin, and Skotnes, Dr. Bonnie 360^0, Oct. 2012)

What are the problems of Big Data?

"Through 2017, 60% of big-data projects will fail to go beyond piloting and experimentation and will be abandoned."

(Gartner 2014)

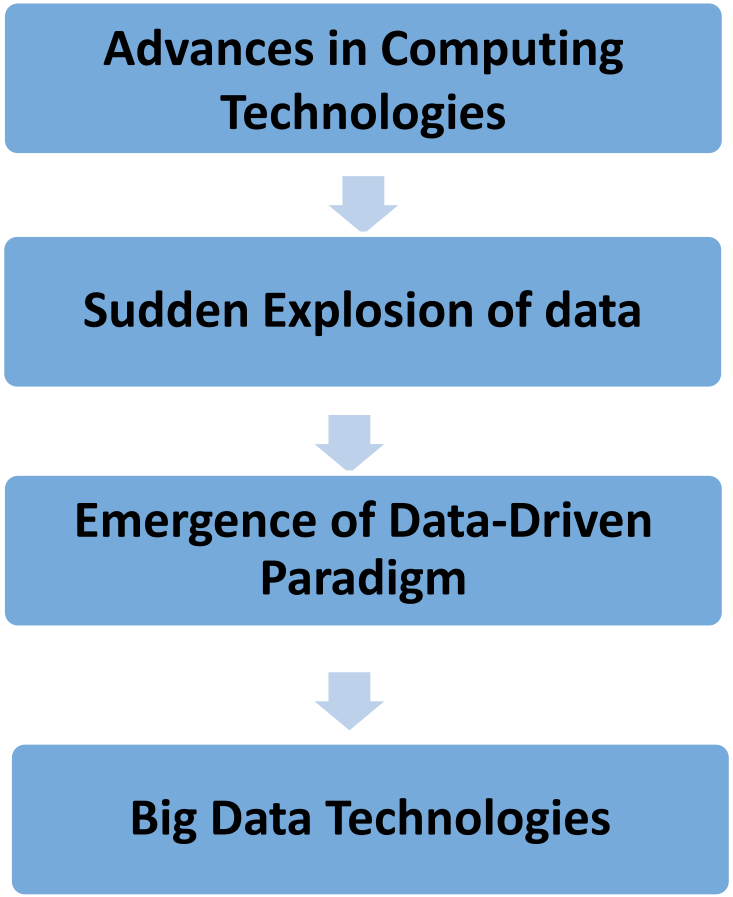
Hype Cycle for Big Data (2014)



Source: Gartner (August 2014)

Source: (Gartner, G00261655, 2014)

Why Big Data?

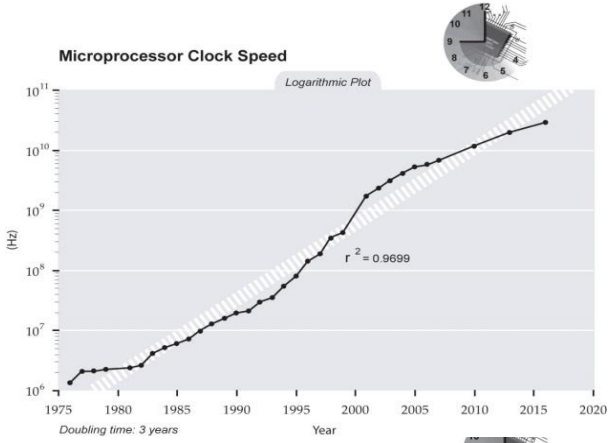

Advances in Computing Technologies
Sudden Explosion of data
Emergence of Data-Driven Paradigm
Big Data Technologies

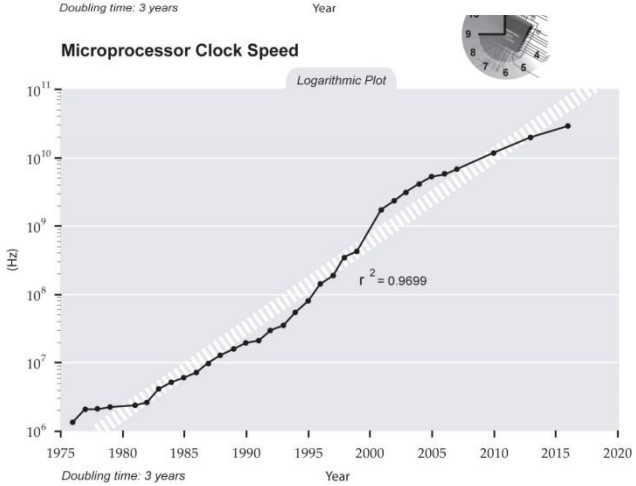
Why Big Data?

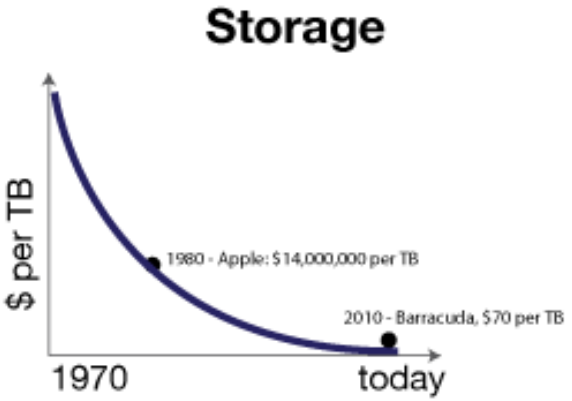
- *Advance in computing technologies*
  - processors,
  - parallel processing technology,
  - increased memory & low storage cost

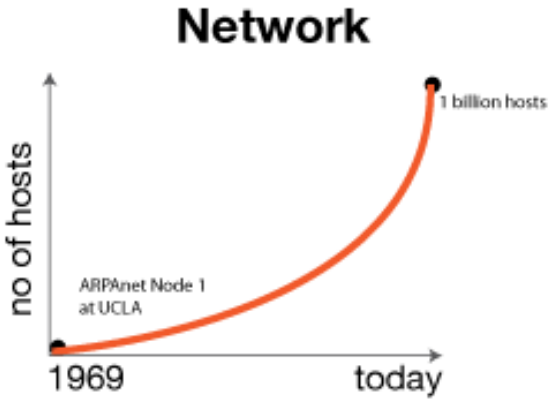
Why Big Data?

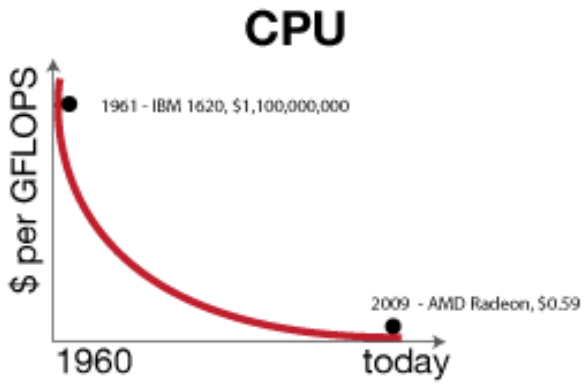
- Advances in storage, network, CPU, and Bandwidth

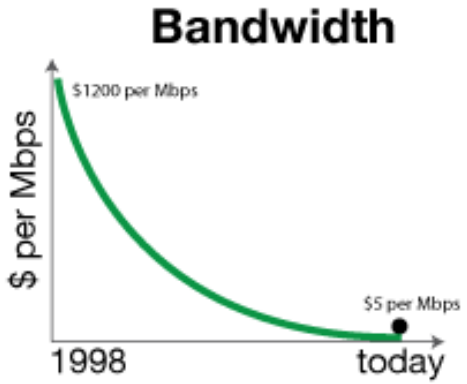












Source: http://radar.oreilly.com/2011/08/building-data-startups.html

Why Big Data?

- *Sudden explosion of data:* ***datafication***
  - Web, Clicks, sensor data, social networks, Web pages, e-commerce data, photo archives, video archives, scientific data (genomics, weather, astronomy, biological research, geographic data), etc.
  - 90% of digital information was generated for the last 2 years and 80% of them are *unstructured data* such as text, image, and video (IBM).

Why Big Data?

- ***Five Types of Big Data (Gartner, 2013)***
  - **Operational data:** those from transaction systems, monitoring of streaming data and sensor data;
  - **Dark data:** those you already own but don't use: emails, contracts, written reports and so forth;
  - **Commercial data:** Structured or unstructured data purchased from industry organizations, social media, etc.
  - **Social data**: those from Twitter, Facebook and other interfaces;
  - **Public data:** numerous formats and topics, such as economic data, socio-demographic data, weather data, etc.

Why Big Data?

- ***Emergence of Data-driven paradigm***
  - How to *utilize those raw data* to *learn new insights, predict trends and changes, introduce innovation and market leads*, and create new opportunities.

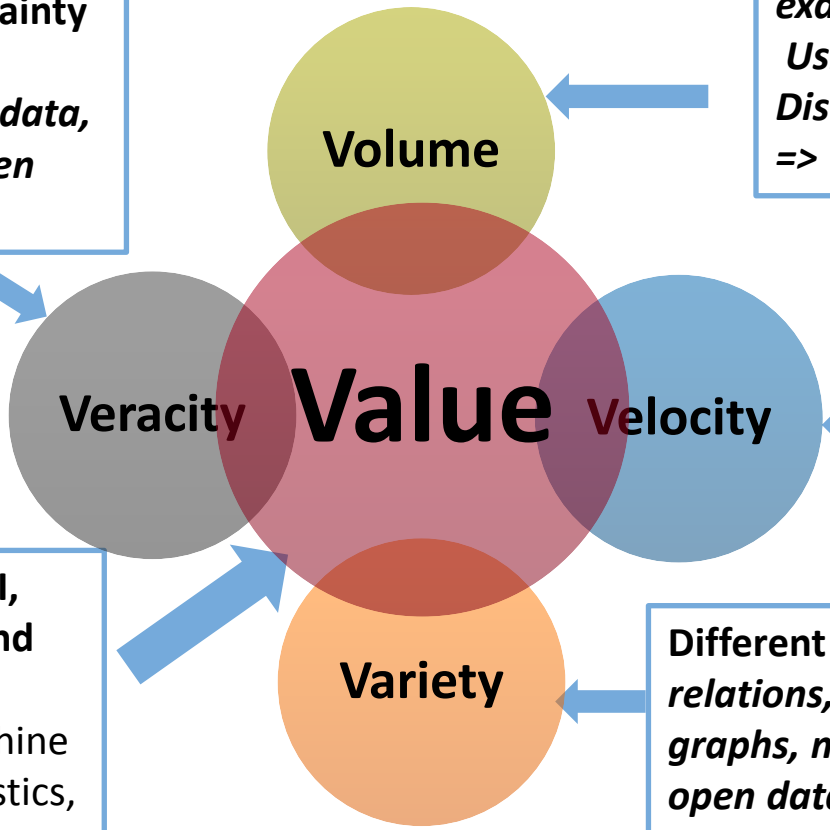
What is Big Data? (3V)

High volume, velocity, and /or variety information assets that demand new, innovative forms of processing for enhanced decision making, business insights or process optimization.

(Big Data: A Definition; Gartner)

What is Big Data?

4Vs



Large volume, cloud, HDFS, EDW, NoSQL
Scale: ***terabytes , petabytes, exabytes***
Using Commodity hardware, Distributed storage (HDFS)
=> Technology solution

Quality, reliability, uncertainty and meaning *in*
data itself (e.g., weather data, translation of hand-written data)

Speed to create/capture/process/ /store
Processing mode: Real-time; streaming; in-memory computing
=> Semi-Technology solution

Actionable knowledge, ROI, Relevancy to customers, and business value
Analysis: SQL queries, machine learning, data mining, statistics, visualization, optimization, decision analysis

Different data types and sources *(e.g., relations, documents, web, XML files, graphs, multimedia, IOT, dark data, open data, external data) -=>data integration/ETL/ELT/Data Virtualization*
=> SW solution

Challenges with Big Data?

- **Volume** and **Velocity** are challenging!
- But **Variety** and **Veracity** are far more challenging!
- But ***Value*** is the most challenging!
  - ***Opportunitie***s for innovative solutions
  - Impacts on technology, society, and business

Data-Driven Paradigm

- How to utilize those raw data
  - *to learn new insights*,
  - *predict trends and changes*,
  - *introduce innovation and market leads*, and
  - *create new opportunities*?

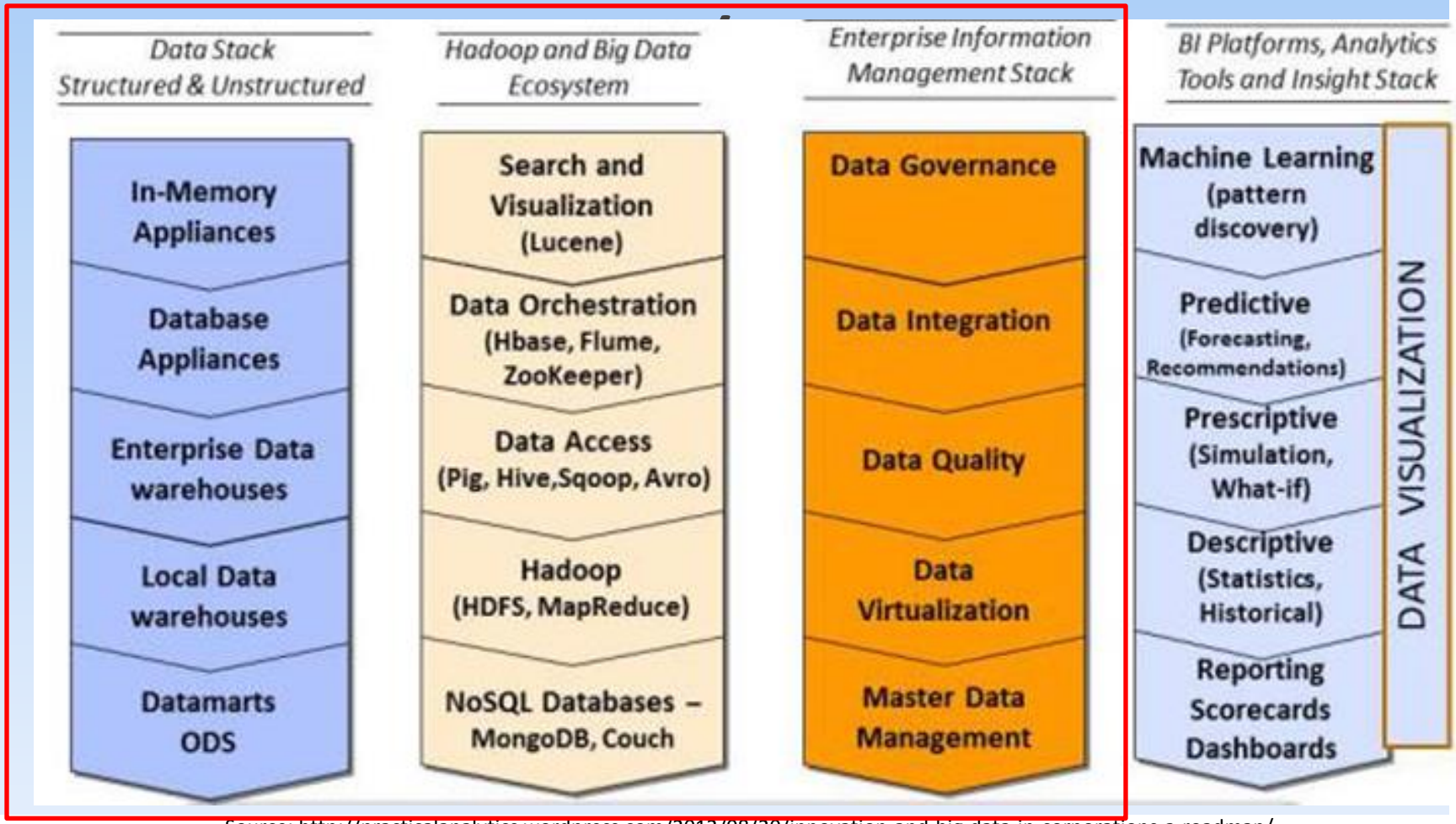
Data-Driven Paradigm (DDP)

- Big data will change the way we live, work and think (Mayer-Schonberger & Cukier, 2013):

Data-Driven Paradigm (DDP)

- Changes introduced by DDP (Mayer-Schonberger & Cukier, 2013):
  - Use the whole data, rather than sampling (eg, Walmart)
  - Accept the probability of prediction (eg, Moneyball)
  - Big data outperforms experts (eg, DeepBlue, Amazon reviewers)
  - Accept correlation; identification of causality may not be possible (eg. UPS, Amazon sales)
  - *Datafication* (eg, car seat data)
    - Quantify/measure as many granular data as possible
    - Let the data speak for itself
  - Movement of value from physical items to brands, ideas, and intellectual rights

Big Data Analytics Infrastructure (Ravi Kalakota)



Source: http://practicalanalytics.wordpress.com/2012/08/20/innovation-and-big-data-in-corporations-a-roadmap/

Problems of Current BDA Infrastructure

- **Too many siloed technology stacks**
  - Difficult to integrate and provide a single unified view of data
  - Increase IT costs
  - Increase complexity in management and maintenance
  - Increase security risks
  - Face performance and scalability issues

- **Expects consolidated stacks and unifying technologies**

Table of Contents

- What is Big Data?
- Data-driven paradigm (DDP)
- Big Data Technologies
  - **Hadoop Ecosystems**
  - NoSQL databases; NewSQL databases
  - In-memory databases
  - Cloud computing
  - Big data warehousing (ETL, ELT, and Data virtualization, EDW, LDW, Data Integration)
- Values and Use cases
- Research issues
- Conclusions

Hadoop & Map/Reduce

- Map/Reduce is a parallel programming framework with automatic parallelization
- Hadoop: Open source version of Map/Reduce
  - Map/Reduce algorithm
  - Extreme **scalability** using low cost commodity hardware
  - **Fault tolerance**
  - Process and store large amounts of structured, unstructured and semi-structured data
- Ex: Yahoo Hadoop cluster with 1000s of server nodes

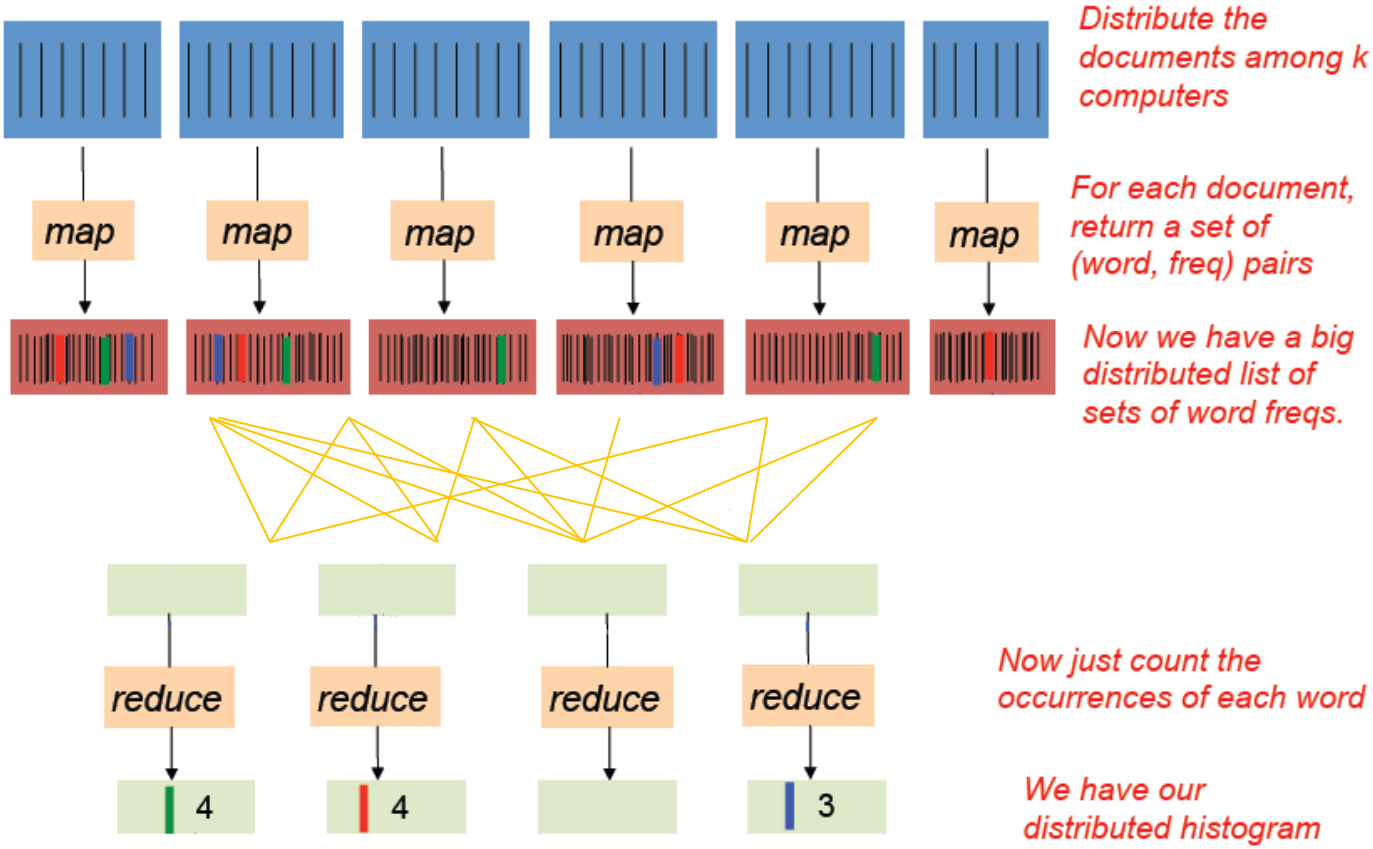
Two Core Components of Hadoop 1

(1) MapReduce: A scalable parallel processing framework

(2) HDFS: distributed, **scalable**, and **fault-tolerant** file-system written in Java

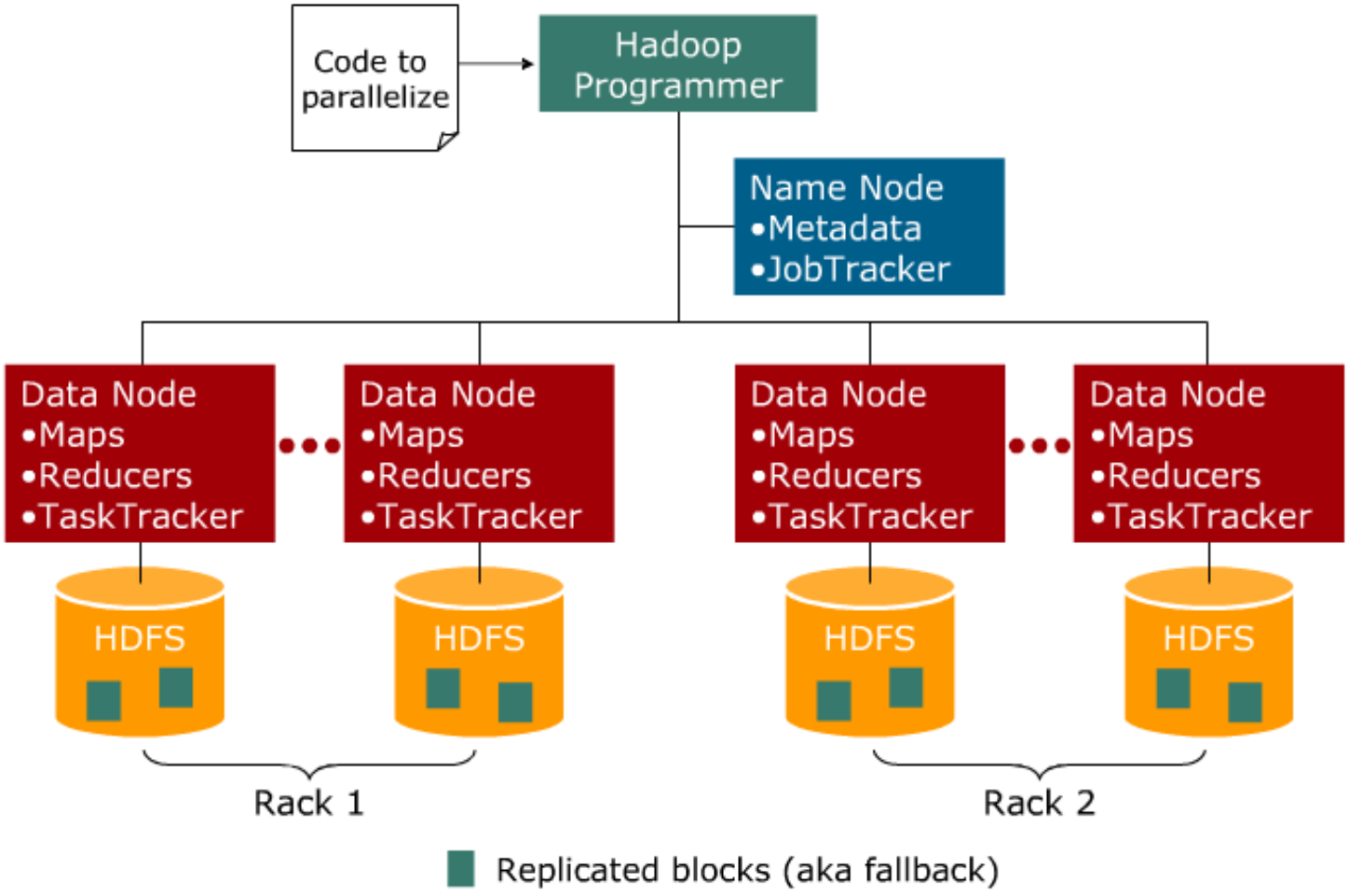
Map/Reduce Parallel Processing Flow

- A simple Example: Compute the word frequency across input documents.



Source: Bill Howe, Retrieved from https://d396qusza40orc.cloudfront.net/datasci/lecture_slides/week3/016_parallel_thinking.pdf

Hadoop Architecture



Why Hadoop?

- The ability to handle **large sets of data** quickly.
- **Low cost**: open source and commodity hardware
- **Distributed** computing power
- **Scalability** by adding nodes
- **Storage flexibility**
- **Triple storage**
- **large and active ecosystems**

Hadoop

- It is **open source** and free, but also available from vendors like IBM, Cloudera, and EMC Greenplum
- Hadoop1 runs M/R algorithm for batch processing
- HDFS is **a file system**, not a DBMS
- Hadoop uses **schema-on-read**.

Typical Hadoop Data

(1) **Clickstream data** (path optimization, basket analysis, next product to buy analysis)

(2) **Sentiment data** from social media stream data (analysis on products, brands, trends, companies, competitions, gap analytics)

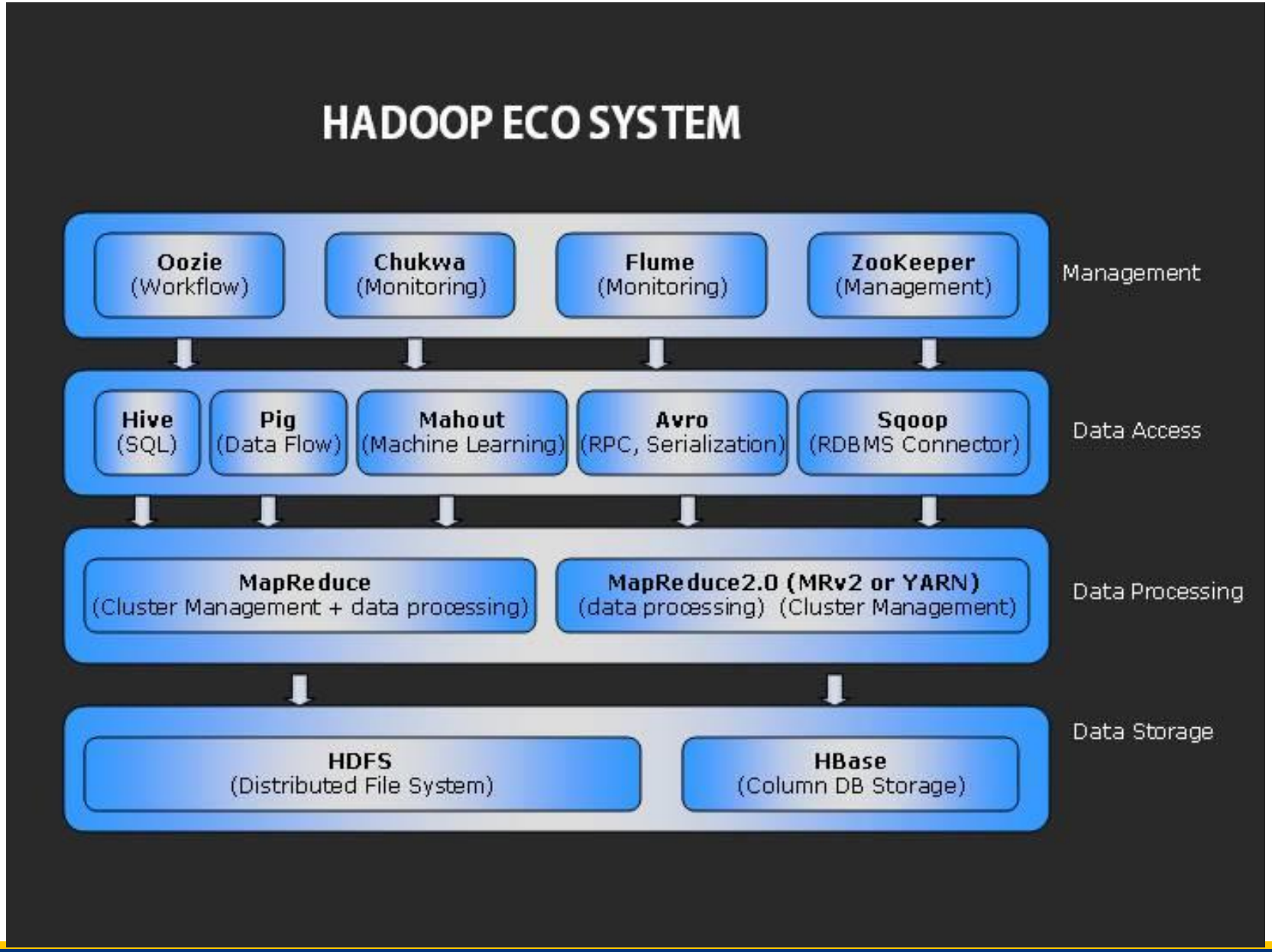
(3) **Structured and Unstructured data** (creating insurance fraud model from both structured and unstructured claim information)

(4) **Geolocation data** from sensors/mobile phones

(track customer movement using moving wifi)

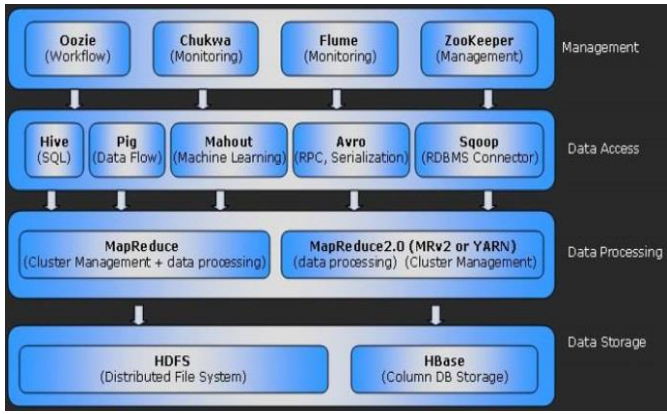
(5) **Server log data** (security analysis on a network)

(6) **Machine and sensor data**


HADOOP ECO SYSTEM
Oozie
(Workflow)
Chukwa
(Monitoring)
Flume
(Monitoring)
ZooKeeper
(Management)
Management
Hive
(SQL)
Pig
(Data Flow)
Mahout
(Machine Learning)
Avro
(RPC, Serialization)
Sqoop
(RDBMS Connector)
Data Access
MapReduce
(Cluster Management + data processing)
MapReduce2.0 (MRv2 or YARN)
(data processing) (Cluster Management)
Data Processing
HDFS
(Distributed File System)
HBase
(Column DB Storage)
Data Storage

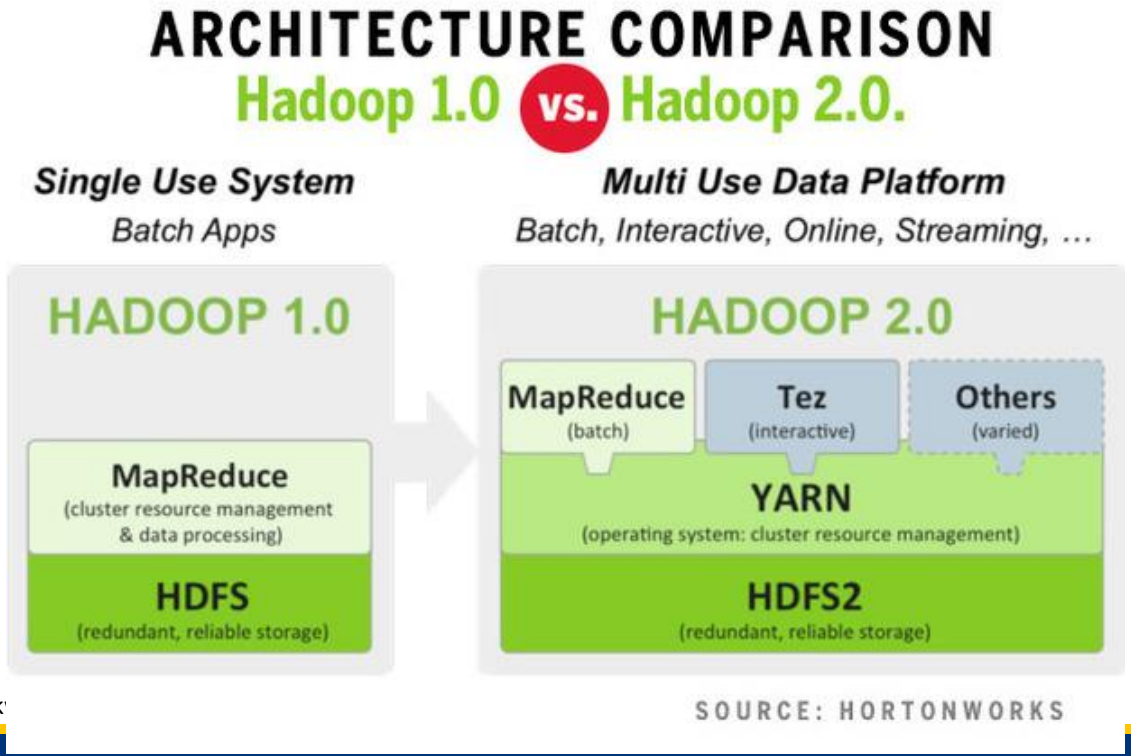
Popular Hadoop Tools

- **HBase**: A Non-relational columnar NoSQL DB that uses HDFS
- **Hive**: *SQL-like* access to data;
- **Pig**: a scripting language to perform ETL for data stored in HDFS
- **Mahout**: scalable machine learning algorithms
- **Storm**: a distributed real time computation system for stream data (Twitter uses Storm to identify trends in near real time)



Hadoop 1 and Hadoop 2

- Hadoop 1 (MapReduce1) vs. Hadoop 2 (Yarn or MapReduce 2)
  - Hadoop1 : Single Use System (**batch** applications only)
  - Hadoop2: Multi Use Data Platform (**batch, interactive, online, streaming, graph**, etc.) **Extends to non M/R applications**



Source: http://www.network

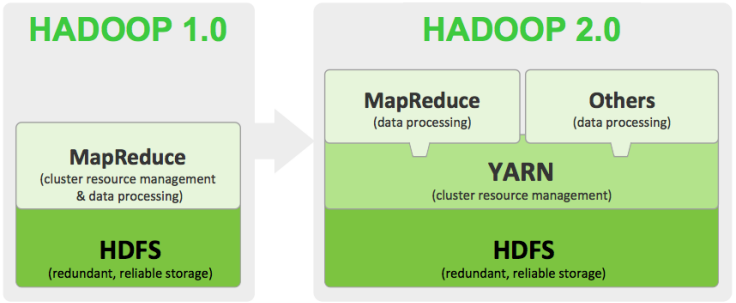
Three Core Components of Hadoop 2

(1) MapReduce: A scalable parallel processing framework

(2) HDFS: distributed, scalable, and fault-tolerant file-system written in Java

(3) YARN (Yet Another Resource Negotiator):

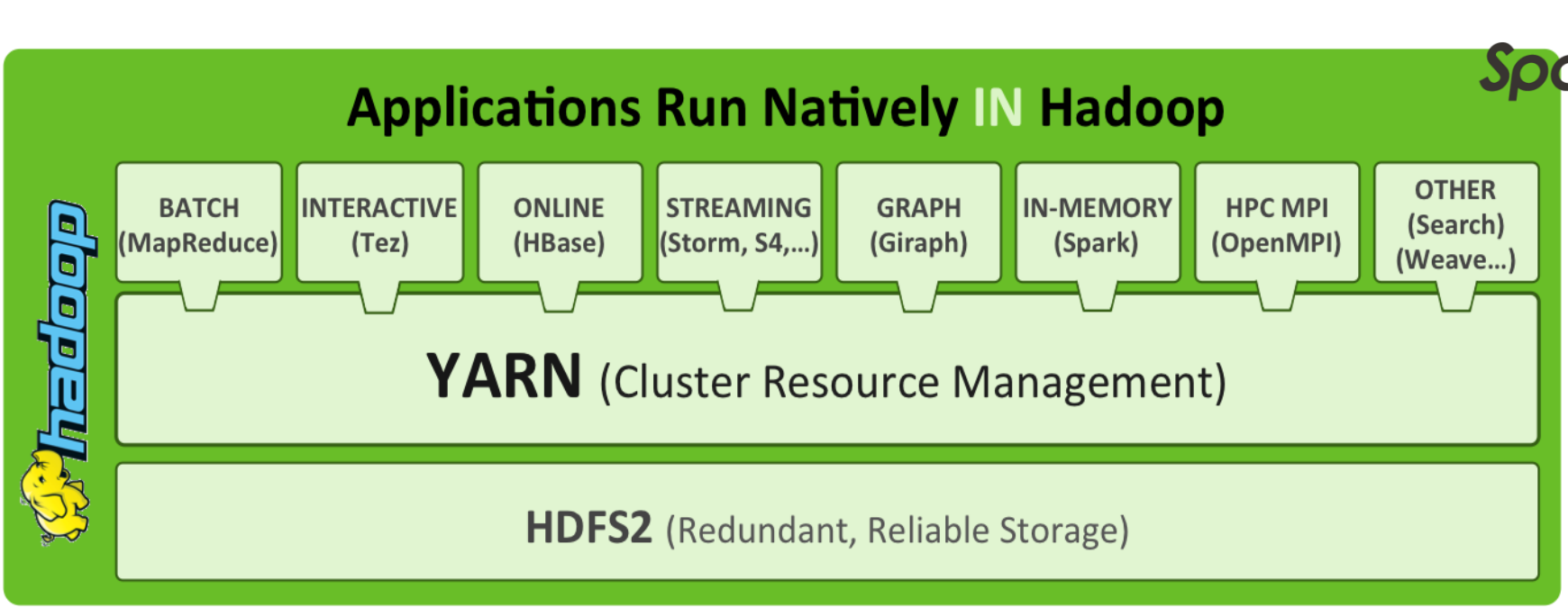
A resource management framework for **scheduling and handling resource requests** from distributed applications.



Source: http://en.wikipedia.org/wiki/Apache_Hadoop

Tools Running on YARN layer in Hadoop 2

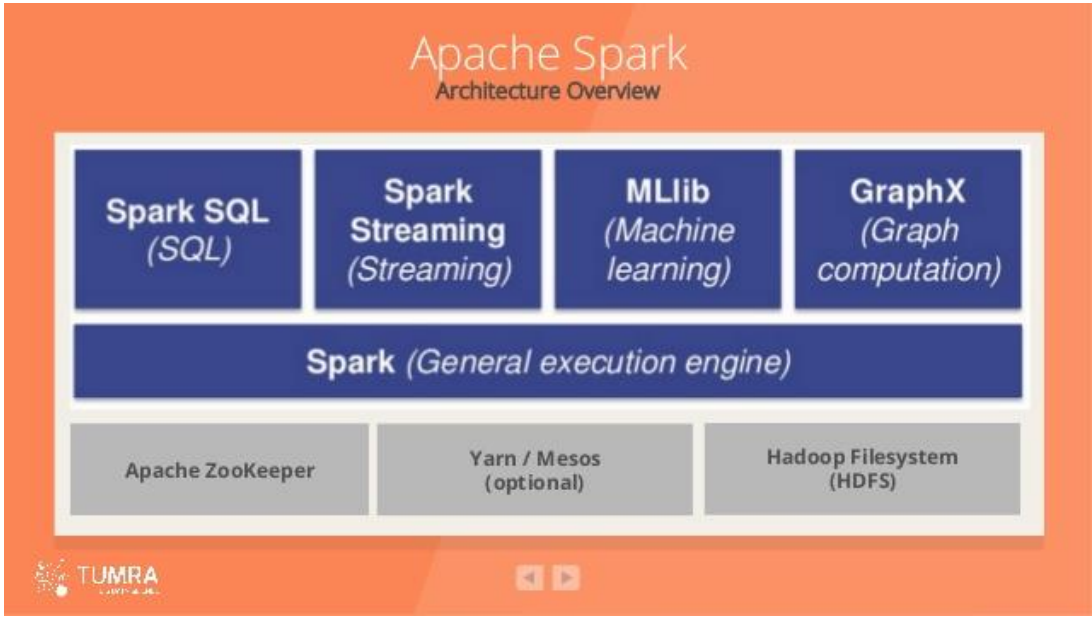
- Spark: An in-memory computing engine to provide real-time analytics.



Source: http://blog.andreamostosi.name/2014/03/hadoop-vs-berkeley/

Apache Spark

- A in-memory centric computing engine to Hadoop data that:
  - Runs up to 100X faster than Hadoop
  - Runs sophisticated algorithms
  - Released in early 2014



Source: Real-Time Big Data Analytics: Emerging Architecture by Mike Barlow, O'Reilly, 2013

Apache Spark

- **Why Spark is faster than Hadoop ?**
  - **Spark is geared toward in-memory processing**
    - Well suited for **iterative algorithms that require multiple passes** over in-memory data as in machine learning algorithms
    - Hadoop was designed to manage very large volumes of data in a batch process workflow

- Note: Spark can be used with Hadoop or without Hadoop.

Source: https://www.xplenty.com/blog/2014/11/apache-spark-vs-hadoop-mapreduce/

Apache Spark

- **Hadoop ecosystem vs. Spark ecosystem**

Hadoop ecosystem	Spark Ecosystem
Component	
HDFS	Tachyon
YARN	Mesos
Tools	
Pig	Spark native API
Hive	Spark SQL
Mahout	MLlib
Storm	Spark Streaming
Giraph	GraphX
HUE	Spark Notebook/ISpark

Source: http://adtmag.com/blogs/dev-watch/2015/03/hadoop-and-spark-friends-or-foes.aspx

Apache Spark Use Cases

- **Yahoo! Personalizes news pages for Web visitors**
- **Yahoo! runs analytics for advertising**
- **Conviva: the 2nd largest video streaming company**

Source: https://spark-summit.org/2014/talk/spark-use-case-at-telefonica-cbs

Apache Spark

- *"The goal of the Spark project is **not to threaten or replace Hadoop**, but rather integrate and interpolate well with a variety of systems (including Hadoop) to make it easier to build more powerful applications"*

  -Kavitha Mariappan, Databricks Inc.

- *"Today, Spark is **too immature** to be considered a replacement for Hadoop, but **it looks like the future**."*

  - Nick Heudecker, Gartner analyst

Source: http://adtmag.com/blogs/dev-watch/2015/03/hadoop-and-spark-friends-or-foes.aspx

Table of Contents

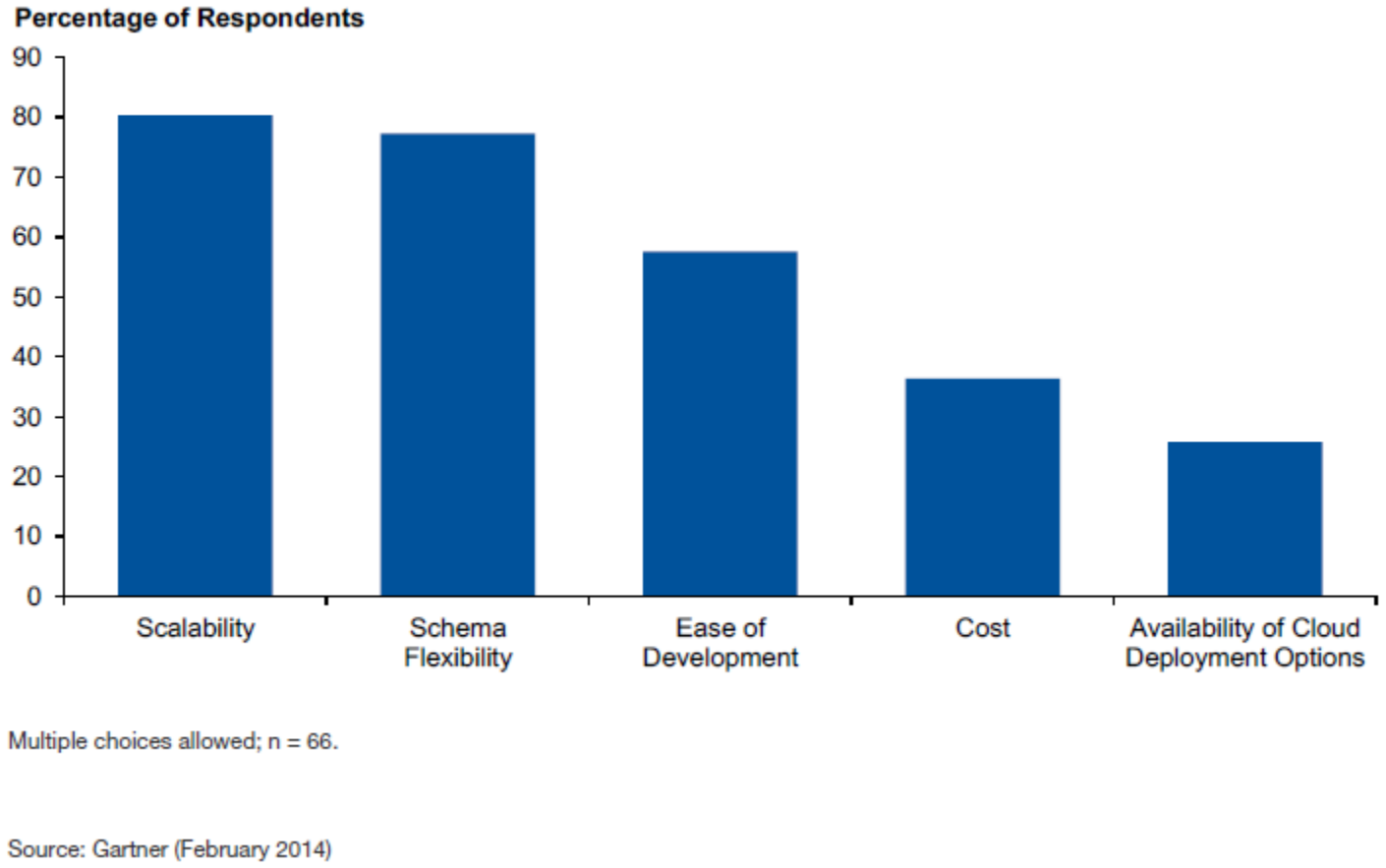
- What is Big Data?
- Data-driven paradigm (DDP)
- Big Data Technologies
  - Hadoop Ecosystems
  - **NoSQL databases; NewSQL databases**
  - In-memory databases
  - Cloud computing
  - Big data warehousing (ETL, ELT, and Data virtualization, EDW, LDW, Data Integration)
- Values and Use cases
- Research issues
- Conclusions

NoSQL Databases

- Designed with needs of **big data**, **big users**, and **cloud** in mind.
  - NoSQL means "Not only SQL"
  - Support unstructured or **non-relational data types**
  - **Schema-less**: no rigid schema enforced by the DBMS
  - **Scale out massively** at low cost and fast retrieval (elastic scaling)
  - **Low cost operational** data management for large #users
  - Support **scalability, performance, fault-tolerance**
  - May not guarantee full **ACID properties**
  - Designed for **real-time, non-uniform big data**, but it's **operational** rather than analytical

Motivation For Using NoSQL

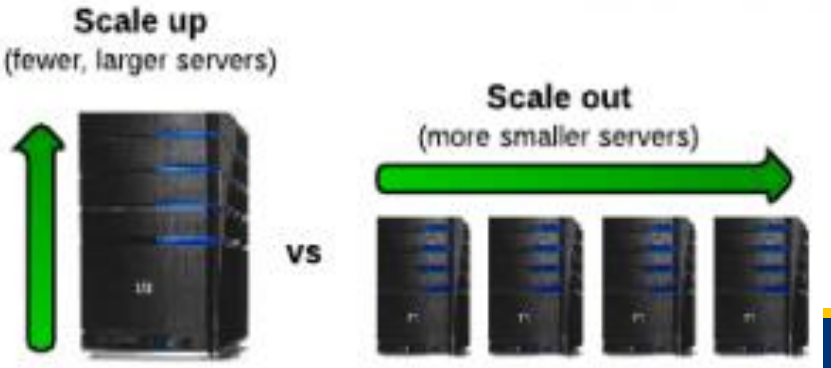
- Multiple choices allowed, sample size = 68



Source: Gartner, A Tour of NoSQL in Eight use Cases

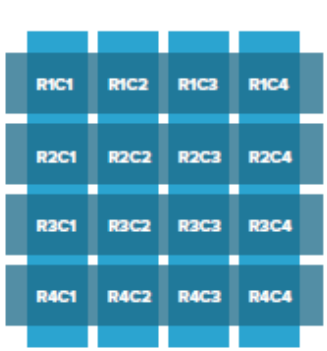
NoSQL Databases

- Making the shift from relational to NoSQL
  - The transition has been spurred by the need for flexibility both in the **scaling model** and the **data model**.
  - Scaling Model
    - **RDB: scale up** (i.e., To add capacity, one need a bigger server)
    - **NoSQL: scale out** (i.e., Instead of buying a bigger server, one add more commodity servers)



NoSQL Databases

- **Data Model**
  - RDB: Define a schema before use; changing the schema is difficult.
  - **NoSQL: No need for schema definition** prior to inserting data nor a schema change when data capture and management needs evolve.



Relational data model
Highly-structured table organization with rigidly-defined data formats and record structure.

Document data model
Collection of complex documents with arbitrary, nested data formats and varying "record" format.

NoSQL Databases

- Most of them lack full ACID compliance for guaranteeing transactional integrity and data consistency.

- ***Eventual consistency*** limits mission-critical transactional applications.

- RDBs and NoSQL DBs will co-exist for many years to come.
  - RDBs: Transaction-based systems
  - NoSQL: Search engines, web-based systems, real-time, cloud, mobile applications, low-cost initial engagement, IoT

NoSQL Databases

- NoSQL models
  - **Key-value Store (KVS)**
    - Redis, Memcached, Riak, DynamoDB, FoundationDB, etc.



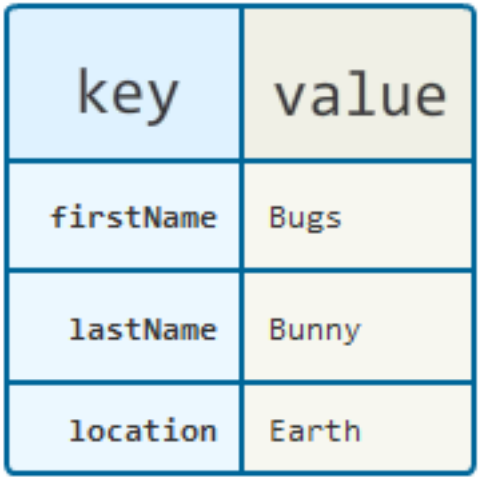
  - **Column Store**
    - Cassandra, HBase, Accumulo, Hypertable, etc.

  - **Document Store**
    - MongoDB, CouchDB, Couchbase, MarkLogic, Cloudant, etc.

  - **Graph Database**
    - Neo4j, Titan, OrientDB, Sparksee, Sesame, InfiniteGraph, etc.


NoSQL Models

- **Key-value Store (KVS)**
  - Key is the main referent inside the data. Value can be any type of data.
  - The number of items in each row can be different.
  - Ex: Redis, Memcached, Riak, DynamoDB, etc.

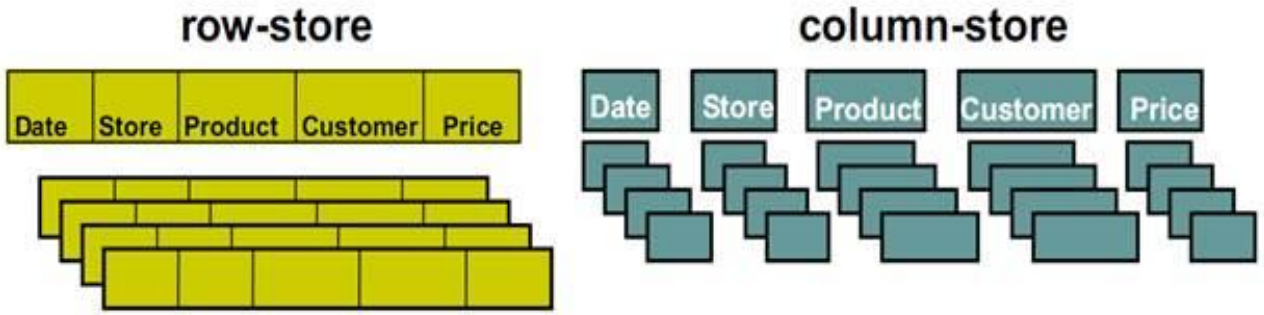


key	value
firstName	Bugs
lastName	Bunny
location	Earth

Source: http://stackoverflow.com/questions/25955749/what-is-a-key-value-pair

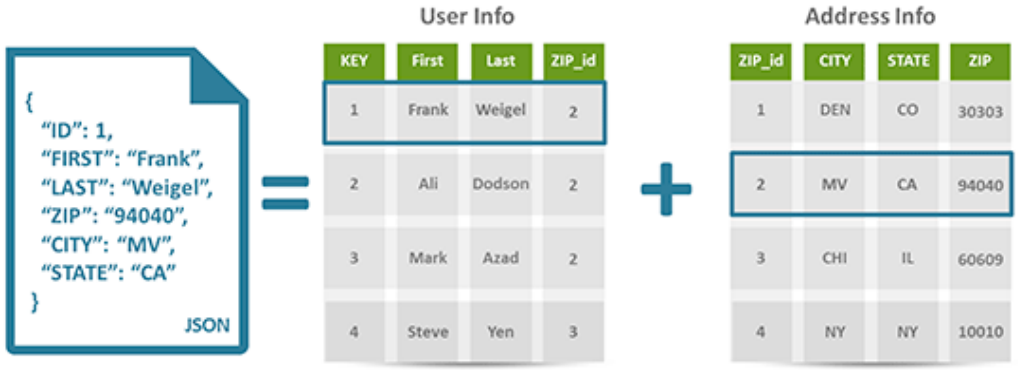
NoSQL Models

- **Column Store**
  - "Wide column model", data are stored in column cells instead of rows.
  - For scenarios where writes are rare and you frequently read several columns of many rows at once, it is better to store data in groups of columns.
  - Cassandra, HBase, Accumulo, Hypertable, etc



NoSQL Models

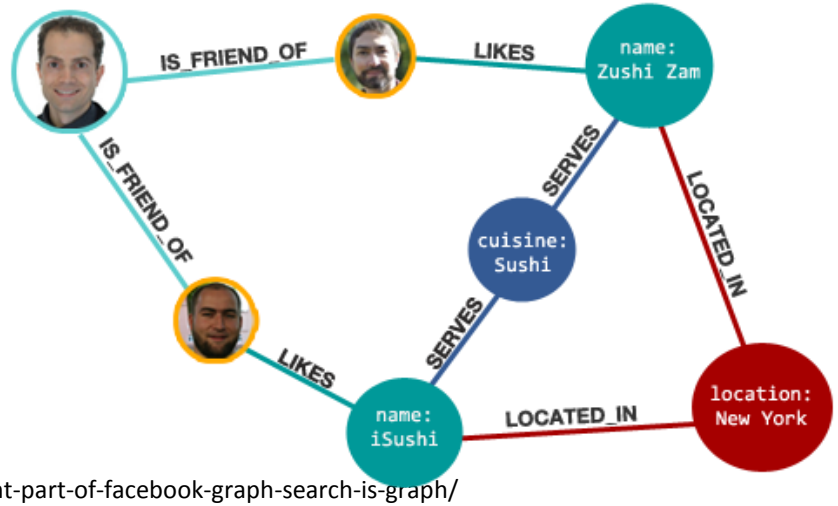
- **Document Store**
  - Similar to Key-value store, but the value is a complete document, such as XML, JSON, etc.
  - Documents are self-describing, hierarchical tree data structures which can consist of maps, collections, and scalar values.
  - MongoDB, CouchDB, MarkLogic, Cloudant, etc



Source: https://developer.ibm.com/bluer

NoSQL Models

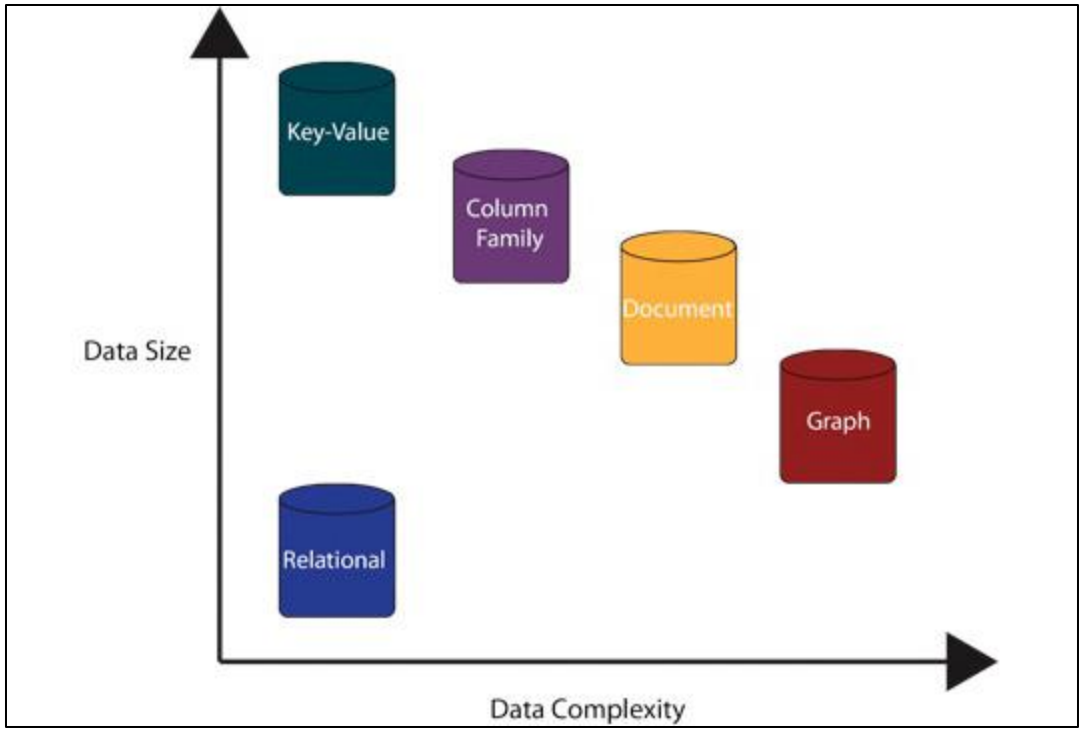
- **Graph Database**
  - Useful for inter-connected data such as communication patterns, social networks, bio interactions.
  - Displaying data in graphic format allows for index-free connections.
  - Allows us to ask deeper and more complex questions
  - Difficult to distribute components of a graph among a network of servers as graphs become larger



Source: http://neo4j.com/blog/why-the-most-important-part-of-facebook-graph-search-is-graph/

NoSQL Databases

- NoSQLs in terms of data size and data complexity



Source: http://www.infoworld.com/article/2617876/database/which-freaking-database-should-i-use-.html

Rankings of DB Systems by Popularity

- NoSQL is moving up.
  - DB-Engines Ranking *by popularity* (April, 2015)
    - From Websites, Google trends, social networks, job Ads, etc.

Rank			DBMS	Database Model	Score		
Apr 2015	Mar 2015	Apr 2014			Apr 2015	Mar 2015	Apr 2014
1.	1.	1.	Oracle	Relational DBMS	1446.13	-22.96	-67.95
2.	2.	2.	MySQL	Relational DBMS	1284.58	+23.49	-8.09
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1149.11	-15.68	-61.31
4.	4.	↑ 5.	MongoDB	Document store	278.59	+3.58	+64.25
5.	5.	↓ 4.	PostgreSQL	Relational DBMS	268.31	+3.88	+38.08
6.	6.	6.	DB2	Relational DBMS	197.65	-1.20	+13.06
7.	7.	7.	Microsoft Access	Relational DBMS	142.19	+0.50	-0.57
8.	8.	↑ 9.	Cassandra	Wide column store	104.89	-2.42	+26.17
9.	9.	↓ 8.	SQLite	Relational DBMS	102.30	+0.59	+12.13
10.	10.	↑ 13.	Redis	Key-value store	94.55	-2.49	+36.09

Source: DB-Engines, http://db-engines.com/en/ranking

Use Cases of NoSQL Databases

- Mobile applications, IoT, Real-time big data (D, KV, C)
- Multisourced structured & Unstructured document management systems (D)
- Content management systems , Catalog (D)
- Patient health and prescription management (KV)
- Real-time gaming (KV)
- Data management from social networks (C)
- Massive marketing services (C)
- Master data management and version management(G)
- Complex relationship management in drugs (G)

NoSQL Databases: Problems

- Not good for OLTP applications due to **no ACID properties**.
  - Eventual consistency is acceptable in some applications
  - A few recent systems claim they support ACID
    - FoundationDB and OrientDB
- A **low-level query language** using APIs
- **No standards**

THE CAP THEORM

- **Brewer's theorem:**

A networked shared-data system can have at most **two of three** desirable properties

- **C**onsistency: every user see the same data at the same time
- **A**vailability: every request receives a response of success or fail
- **P**artitions tolerance : system works even if a partition fails

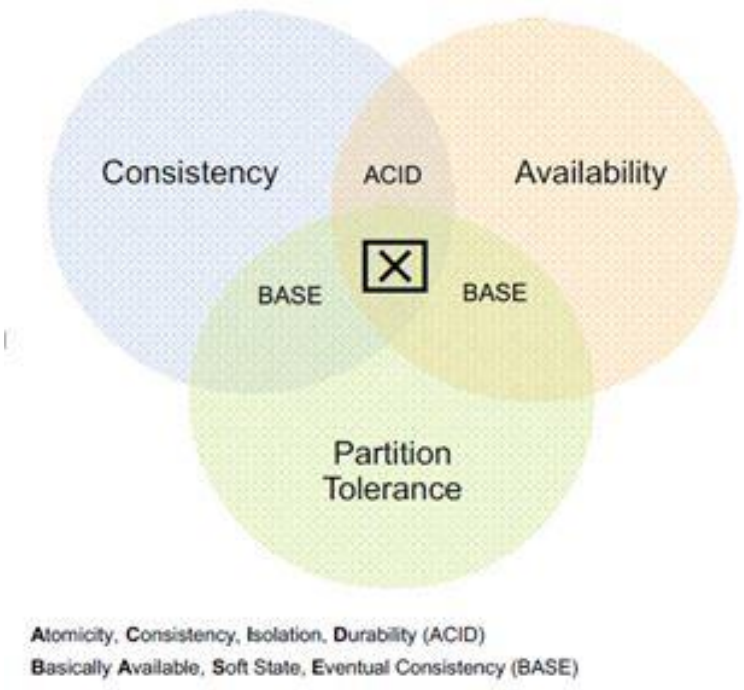
ACID, BASE, and CAP

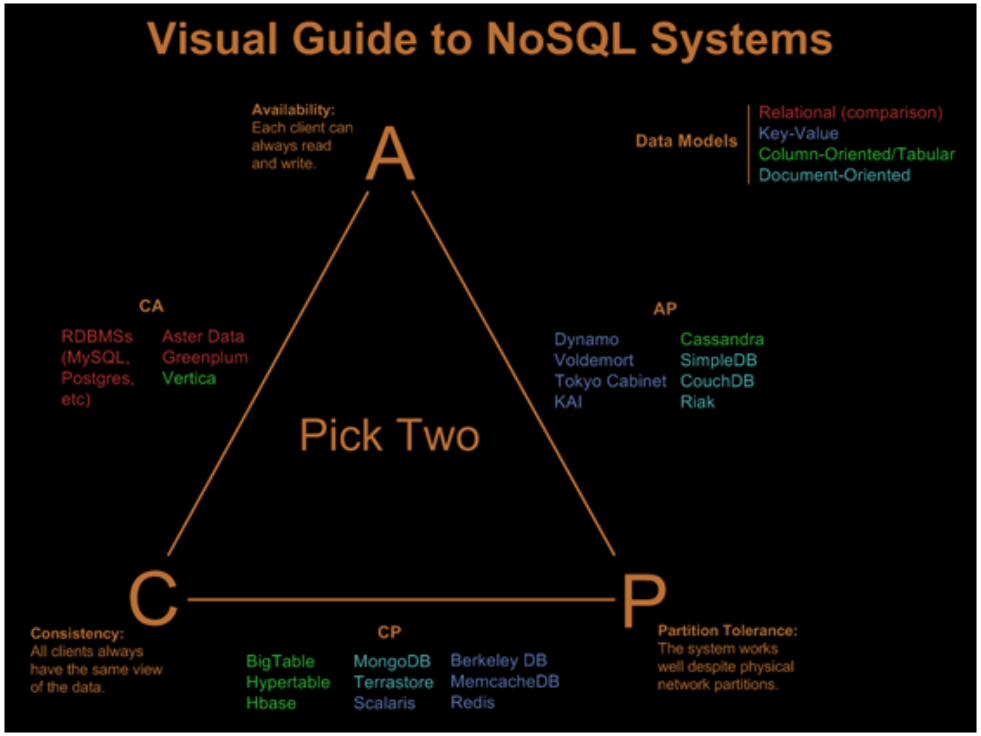
- ACID and BASE represent two design philosophies at opposite ends of the **consistency-availability spectrum**.
- The **ACID** focus on **consistency** (traditional approach of DBs).
- The **BASE** represents design approaches for **high availability**.
  - **Basically Available, Soft state**, **Eventually consistent**
- Modern larger-scale systems use a **mix of both approaches**.

- A system has to choose between **consistency** and **availability** when partitions are present
  - **RDBs:** emphasize **CA**
  - **NoSQL**: sacrifice **C** with preference of **A** and **P**

Source: Eric Brewer, InfoQ, http://www.infoq.com/articles/cap-twelve-years-later-how-the-rules-have-changed

ACID, BASE, and CAP THEORM





Source: Eric Brewer, InfoQ, http://www.infoq.com/articles/cap-twelve-years-later-how-the-rules-have-changed
Source: http://stackoverflow.com/questions/11292215/where-does-mongodb-stand-in-the-cap-theorem

NewSQL

- NewSQL is a class of *new relational database* with *scale-out scalability* and *performance*:
  - provide the **scalable performance** comparable to NoSQL systems for OLTP workloads
  - Support the **ACID** properties and SQL
  - Ex: Google spanner, VoltDB, MemSQL, NuoDB, Clustrix

Source: Aslett, Matthew (2011). "How Will The Database Incumbents Respond To NoSQL And NewSQL?". 451 Group (published 2011-04-04)

NewSQL

- Two common features of various NewSQL systems
  - They all support the relational data model (manage both **structured** and **unstructured**)
  - They all use **SQL** as their primary interface.
- Weakness: Limited support for “variety” due to the need of schema

	Old SQL	NoSQL	NewSQL
Relational	Yes	No	Yes
SQL	Yes	No	Yes
ACID transactions	Yes	No	Yes
Horizontal scalability	No	Yes	Yes
Performance / big volume	No	Yes	Yes
Schema-less	No	Yes	No

Source: http://labs.sogeti.com/newsql-whats/

NewSQL

Source: hemang Tailor, Sushant Choudhary, Vinary Jain, Rise of NewSQL

Use Cases of NewSQL Databases

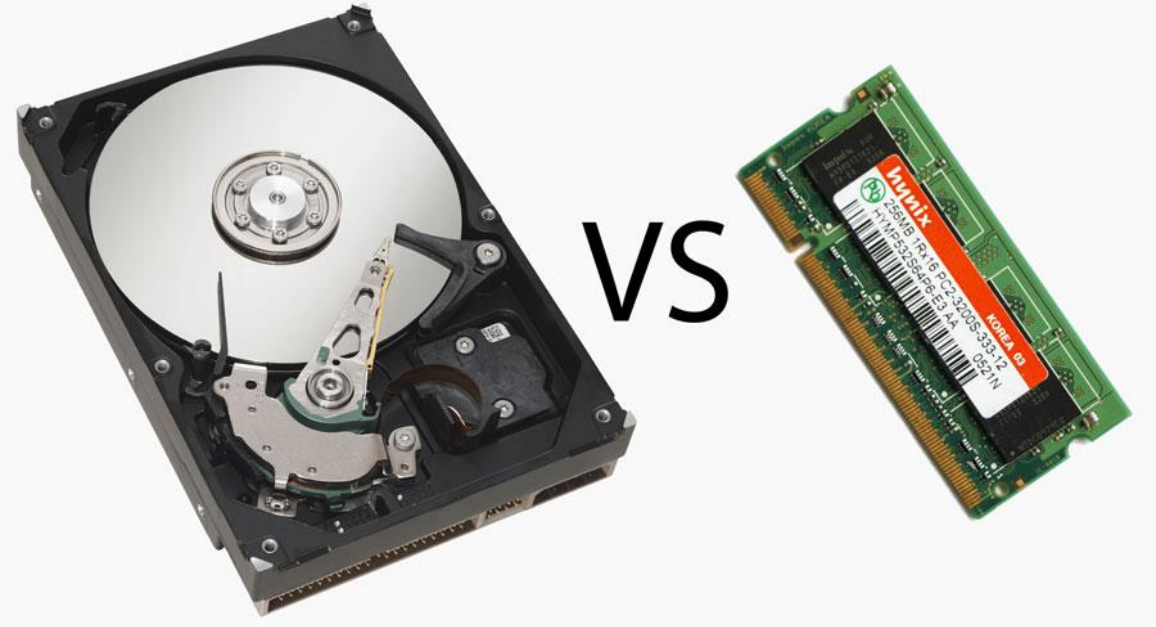
- Real-time fraud detection
- On-line gaming
- Financial trades
- Digital advertising
- Real-time pricing and billing
- Teleco streams
- Retail loyalty & reward systems
- IoT

Table of Contents

- What is Big Data?
- Data-driven paradigm (DDP)
- Big Data Technologies
  - Hadoop Ecosystems
  - NoSQL databases; NewSQL databases
  - **In-memory databases**
  - Cloud computing
  - Big data warehousing (ETL, ELT, and Data virtualization, EDW, LDW, Data Integration)
- Values and Use cases
- Research issues
- Conclusions

In-Memory Computing

- Background: Improvement in network bandwidth, multicore processors, but not disk I/O speed.
  - DRAM costs are dropping about **32%** every 12 months.
  - RAM uses **99% less electrical power** than spinning disks.

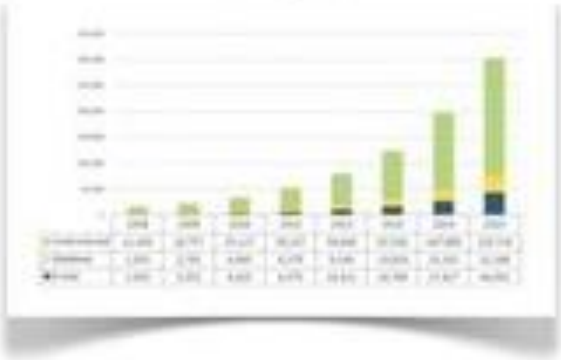


Source: Gartner, Retrieved from http://timoelliott.com/blog/2013/04/why-in-memory-computing-is-cheaper-and-changes-everything.html

In-Memory Computing

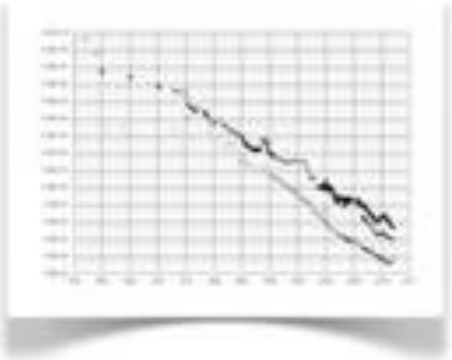
- Why in-memory computing now?

Data Growth in Petabytes



Data growth is exponential!

RAM Cost over time



Cost drops 30% every 18 months

BigData Technologies Planned

70% will use in-memory or appliance

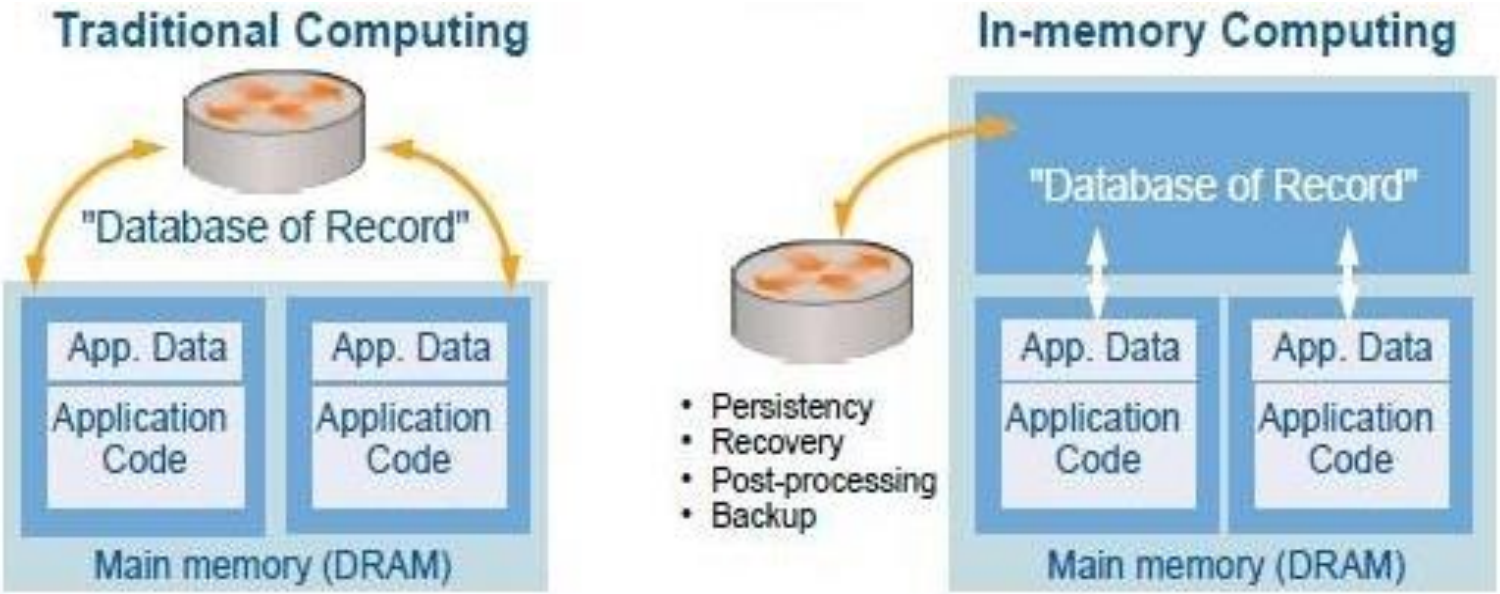
Gartner

In-Memory Computing will have an industry impact comparable to web and cloud.
Ram is the new disk, and disk is the new tape

Source: http://vimeo.com/72919320

In-Memory Computing

- Computing with all data in RAM rather than in disk.
  - No disk buffer
- IMDBs are ACID-compliant relational databases offering SQL.
  - Durability are supported by snapshots, transaction logging, etc



Source: Gartner, Retrieved from http://timoelliott.com/blog/2013/04/why-in-memory-computing-is-cheaper-and-changes-everything.html

In-Memory Computing

- **Example**
  - GridGain (In-Memory Computing) software demonstrated **one billion business transactions per second** on **10 commodity servers** with the total of **1TB of RAM**. The total cost of these 10 commodity servers is less than $25K.

Source: http://www.gridgain.com/in-memory-computing-in-plain-english/

Key Benefits of IMC

- ***Speed***: Processing of in-memory data is much faster than the processing of "on disk" data (at least 5000X faster)
- ***Scale***: Partitioning of in-memory data across multiple physical servers enables the implementation of high-scale applications.
  - Easy integration of data
- ***Processing power***: Fast analytical processing for computing-intensive tasks; easy to generate insights
  - **Fast analytics**

Source: Gartner, Taxonomy, Definitions and Vendor Landscape for In-Memory Technologies

In-Memory Data Management Options

- **In-Memory Database (IMDB)**
  - RDBMS for storing data in memory with SQL support
- **In-Memory Data Grid (IMDG):**
  - IMDGs store data in RAMs of servers distributed over a cluster.
  - Greater ability to scale-out than the IMDB
  - Less sophisticated SQL support than the IMDB

Source: Gartner, Taxonomy, Definitions and Vendor Landscape for In-Memory Technologies

IMDB vs. IMDG

	In-Memory Data Grid	In-Memory Database
Existing Application	Changed	Unchanged
Existing RDBMS	Unchanged	Changed or Replaced
Speed	Yes	Yes
Max. Scalability	Yes	No

Source: http://www.gridgain.com/in-memory-database-vs-in-memory-data-grid-revisited/

Types of IMDBs

- **By partition:**
  - In-memory **column-store DBMSs**
  - In-memory **row-store DBMS**

- **By processing:**
  - **Operational IMDB:**
  - **Analytical IMDBMS**
  - **Hybrid Transactional and Analytical Processing (HTAP)**

Types of IMDBs

- **Operational IMDB:**
  - to manage transactions
  - Speed up transactions by **100 to 1,000 times**
  - Relational, NoSQL, NewSQL
  - SAP Hana, Microsoft SQL Server 2014, Oracle TimesTen, Aerospike, MemSQL, VoltDB, Couchbase.

Types of IMDBs

- **Analytical IMDBMS**
  - Addressing analytical needs
  - IBM, Microsoft and Teradata have expanded their DBMSs to include in-memory columnar capabilities.
  - SAP Hana, Microsoft SQL Server 2014, Oracle Database 12c in-memory option, IBM DB2 with BLU Acceleration, Teradata Intelligent Memory

Types of IMDBs

- **HTAP IMDBMS**
  - Supporting both **analytical** and **transactional** workloads.
  - **Removing the latency of moving data** from operational databases to data warehouses and data marts for **analytical processing**
  - Enables real-time analytics and situation awareness on **live transaction data** as opposed to after-the-fact analysis on **stale data** (as in traditional approaches).
  - Vendors
    - SAP Hana, Microsoft SQL Server 2014, MemSQL, VoltDB.

Source: Gartner, Market Guide for In-Memory DBMS

Use Cases of In-Memory DBMS

- **Characteristics of IMDB use cases**
  - **Speed** of query and calculation is critical.
  - **Often-changing requirements** make it difficult to create dedicated optimizations such as aggregates or indexes.
  - **Low-latency analysis** requires that calculations over data be done in real time and in the finest grain of detail.

- **Examples of Use Cases with above characteristics**
  - *Real-time pricing calculations* (low-latency and speed of query)
  - *Profitability analysis at any point in time* (speed of query and often changing requirements).
  - Real-time applications (pricing calculations , fraud detection)
  - Personalized Product Offering
  - CEP

Source: Gartner, Taxonomy, Definitions and Vendor Landscape for In-Memory Technologies

In-Memory Computing

- **In-Memory architectures are widely used in Relational, NoSQL (Aerospike, Couchbase), and NewSQL (VoltDB, MemSQL).**

- **Systems**
  - Commercial: SAP HANA, Oracle TimesTen, VoltDB, MemSQL

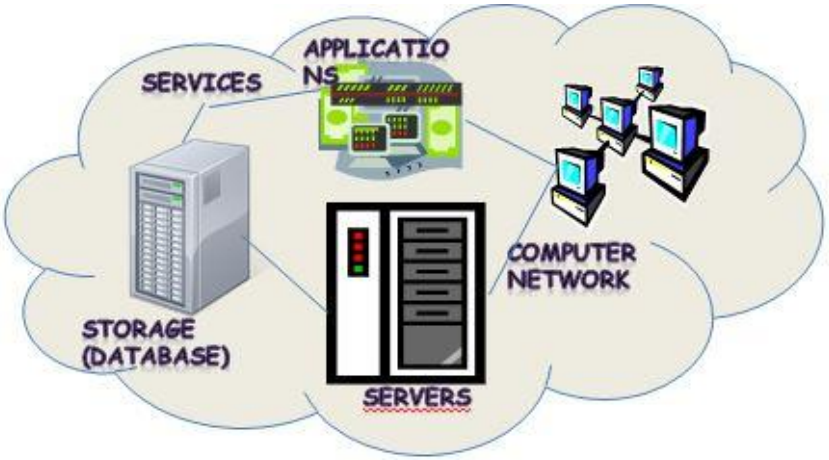
    IBM SolidDB, Sybase ASE, etc
  - Open source; CSQL, MonetDB

Table of Contents

- What is Big Data?
- Data-driven paradigm (DDP)
- Big Data Technologies
  - Hadoop Ecosystems
  - NoSQL databases; NewSQL databases
  - In-memory databases
  - **Cloud computing**
  - Big data warehousing (ETL, ELT, and Data virtualization, EDW, LDW, Data Integration)
- Values and Use cases
- Research issues
- Conclusions

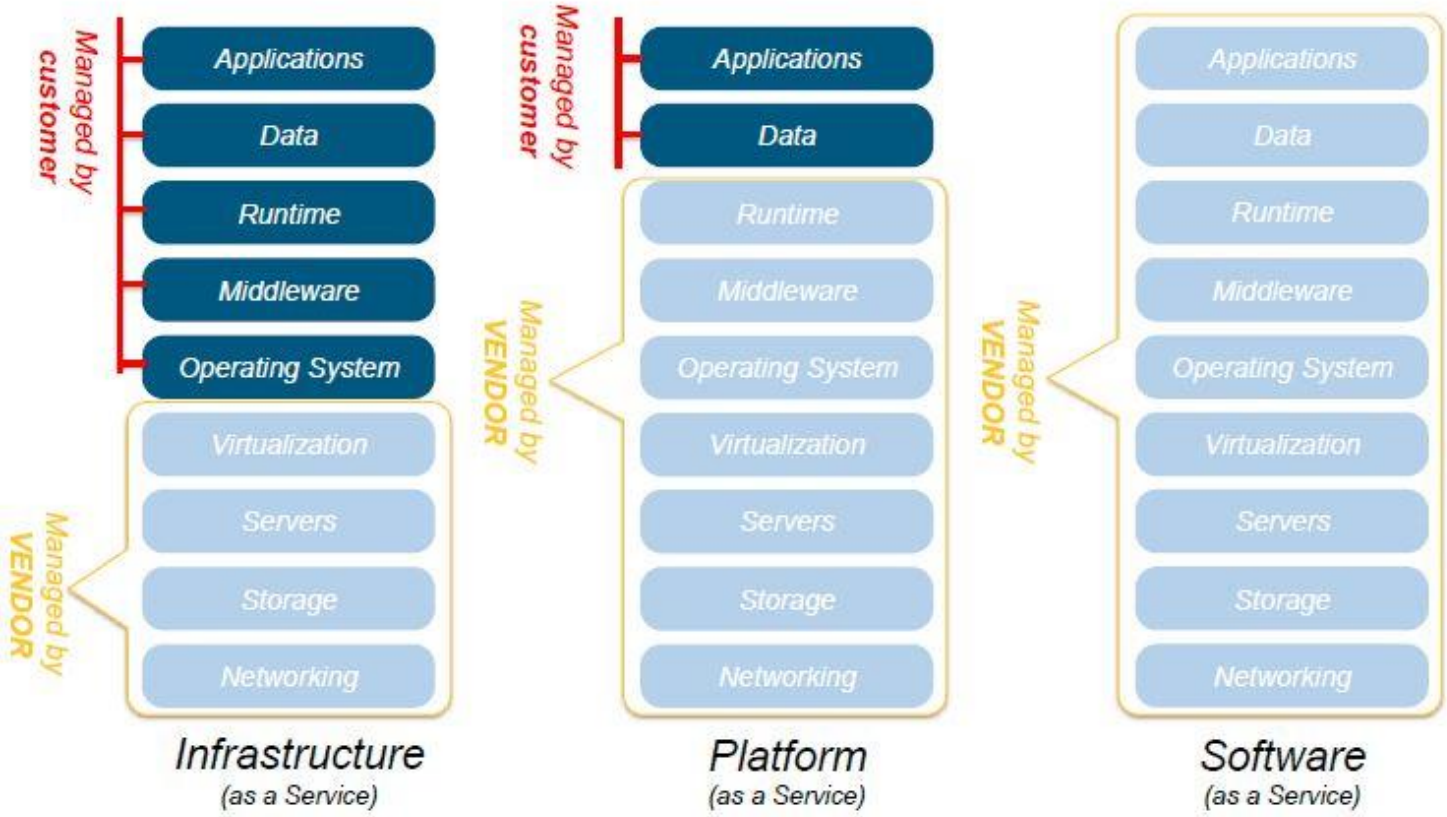
Cloud Computing

- Cloud Computing is the delivery of computing as a service through a Web browser
- Computing as a utility: "**pay as you use and need**", "**access anytime and anywhere**" via Web
- Major Benefits
  - **Reduce costs for IT infrastructure**
    /maintenance
  - **Quick engagement**
    E.g., Use AWS for Hadoop framewo
    - Debug at Lab system, and use AWS for main analysis



Cloud Computing

Cloud Architecture Patterns



Cloud Computing

- **IaaS**
  - Saves issue of HW requirements, scalability, server maintenance
  - Creates issues on network latency and response time
  - Might stumble on unexpected difficulties due to lack of control
- **PaaS**
  - Might have a steep learning curve due to cloud-specific environment
  - Allow users to focus on semantic issues such as correctness of the model, optimal design, not storage, but security issues
  - DB as a service
- **SaaS**
  - Mapping between our domain model and the model provided by SaaS
  - integration is by far the most challenging part of the cloud

Vertical Cloud

- Optimization of cloud computing and cloud services for a particular vertical or specific application use.
- Verticals:

  - E.g., Cognitive Scale announced vertical industry clouds for use in the travel, healthcare, retail and financial services arena.

Source:: http://www.webopedia.com/TERM/V/vertical_cloud_computing.html

Hybrid Cloud Computing

- Store ***private/sensitive/critical data*** *in* ***on-premise*** and **sharable data** in a **public cloud**
- Valuable for **dynamic** or **highly changeable** workloads.

Source: http://searchcloudcomputing.techtarget.com/definition/hybrid-cloud
Source: http://angloafrican.com/hybrid-cloud-look-out-to-the-next-big-thing/

Hybrid Cloud Computing

- **Example Use Cases (Healthcare)**
  - Use a public cloud for non-sensitive shareable patient data between healthcare providers and insurance companies
    - Compliance with HIPAA is a regulatory hurdle.

- **Example Use Cases (Finance)**
  - Process trade orders through the private cloud and running analytics on trades from the public cloud

Source: http://www.zdnet.com/article/hybrid-cloud-what-it-is-why-it-matters/

Data Integration in Hybrid Clouds

- Integration Platform as a service (iPaaS) component as a **generalized cloud gateway**

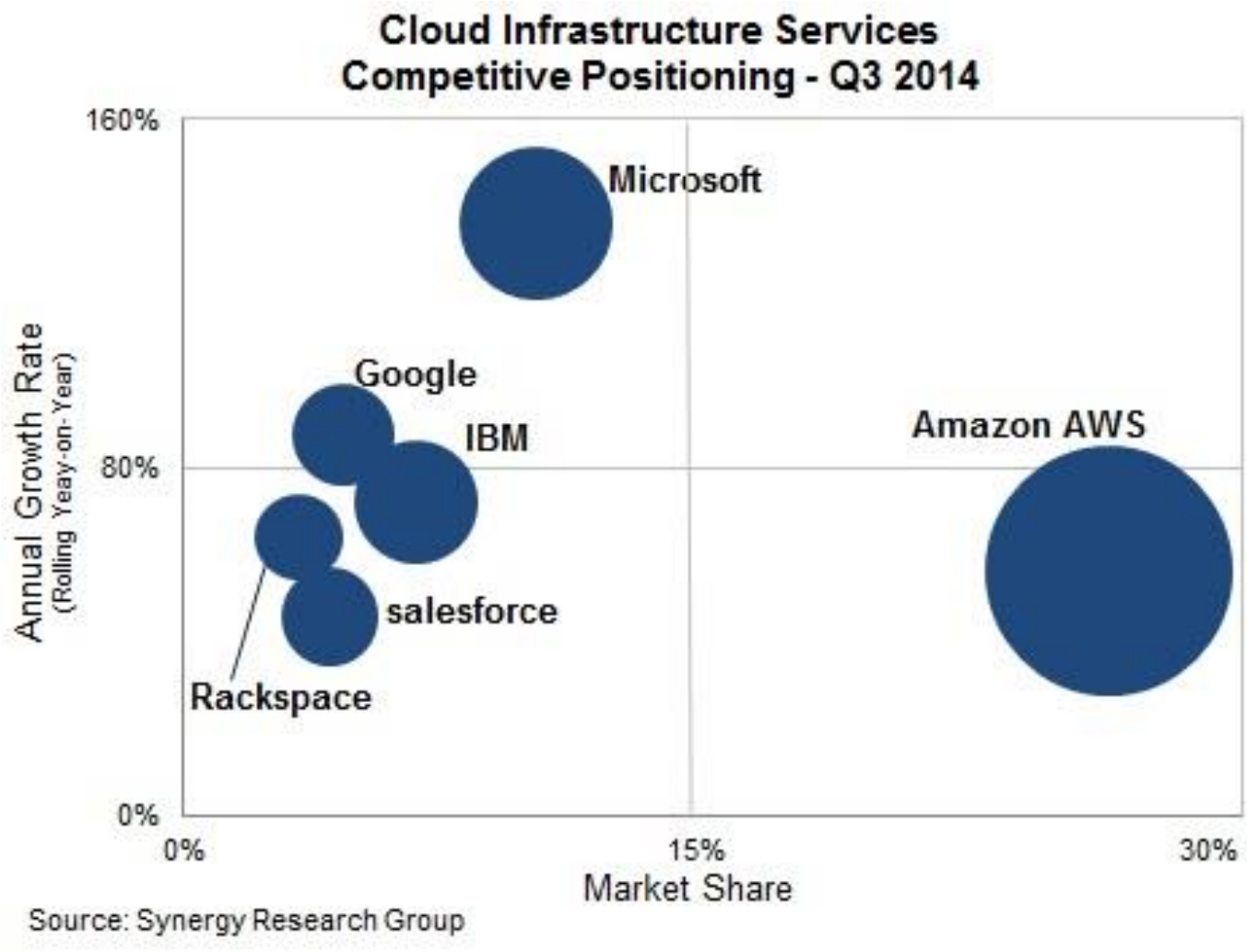
Source: Gartner, Selecting the Most Sutiable Platform for Cloud Service Integration

Cloud-Based Integration Platforms

- **Emerging Hybrid integration platform**
  - Is an integration platform as a service (**iPaaS**)
  - that integrates Web, mobile, cloud, and on-premise data
  - Interoperability is still challenging

Source: Gartner, Why There's a Move Toward Hybrid Integration Platforms

Market share & Revenue growth (Q3 2014)



amazon web services™

Windows Azure

IBM SmartCloud Enterprise

Google Compute Engine

http://http://www.cloudcomputing-news.net/news/2014/oct/29/microsoft-takes-clear-lead-iaas-second-place-race-aws-still-way-out-front/

Cloud Computing

- **Concerns**:
  - **Performance** is dependent on others
  - **Reliability**
  - **Control of data**
  - **No standard API**
  - **Cost** of data migration, loading, and integration
  - **Security**:
  - **Privacy**

Cloud Computing

- **Cloud is a strategy**
  - Provides worry-free maintenance; rapid scalability; incremental cost as you pay as you use
  - Cloud itself will not solve data integration and management issues.
  - Just introduces efficiency to the process
  - Hence, cloud needs good data governance, high quality metadata, and well-understood data integration process

Table of Contents

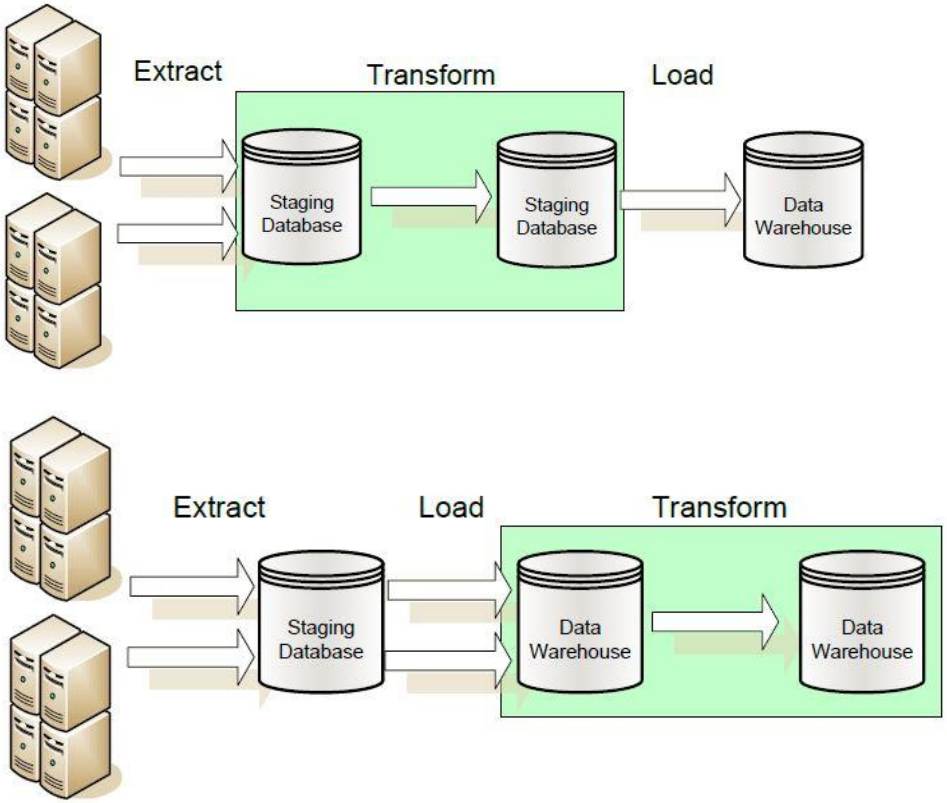
- What is Big Data?
- Data-driven paradigm (DDP)
- Big Data Technologies
  - Hadoop Ecosystems
  - NoSQL databases; NewSQL databases
  - In-memory databases
  - Cloud computing
  - **Big data warehousing (ETL, ELT, and Data virtualization, EDW, LDW, Data Integration)**
- Values and Use cases
- Research issues
- Conclusions

Recent Advances in DW Technologies

- ETL vs ELT
- Data Virtualization
  - Virtual data mart
  - Integration of DWs with external data
  - Federation of DWs

ETL vs. ELT

- ELT: **Transformations** take place in the target database after data loading.
  - **ETL:**
  - **ELT:**



- Makes data available much faster than ETL
- More appropriate for Big Data applications in which large volumes of data are rapidly generated
- **In real-time systems**
- **To apply analytics to raw data**

Data Virtualization

- Provides a single layer of abstraction on the top of multiple data sources, **without having to create and store new copies of the information**.

Traditional Data Warehouse

- Traditional DWs are repository-oriented
- Most analytic use cases still need DWs
- Needs to generate regulatory reports from a single version of truth

Data Warehouse in Big Data Environment

- Big data changes data management practice
  - Traditional DWs have difficulty of coping with volume, velocity, and variety from big data such as clickstream data, social media data , and sensor data
    - Scalability issues
    - Not economical
    - Performance issues

    ➔Hadoop can cope with them.

- 90% of data warehouses will not be replaced (Gartner, 2014)

Big Data and Data Warehouse

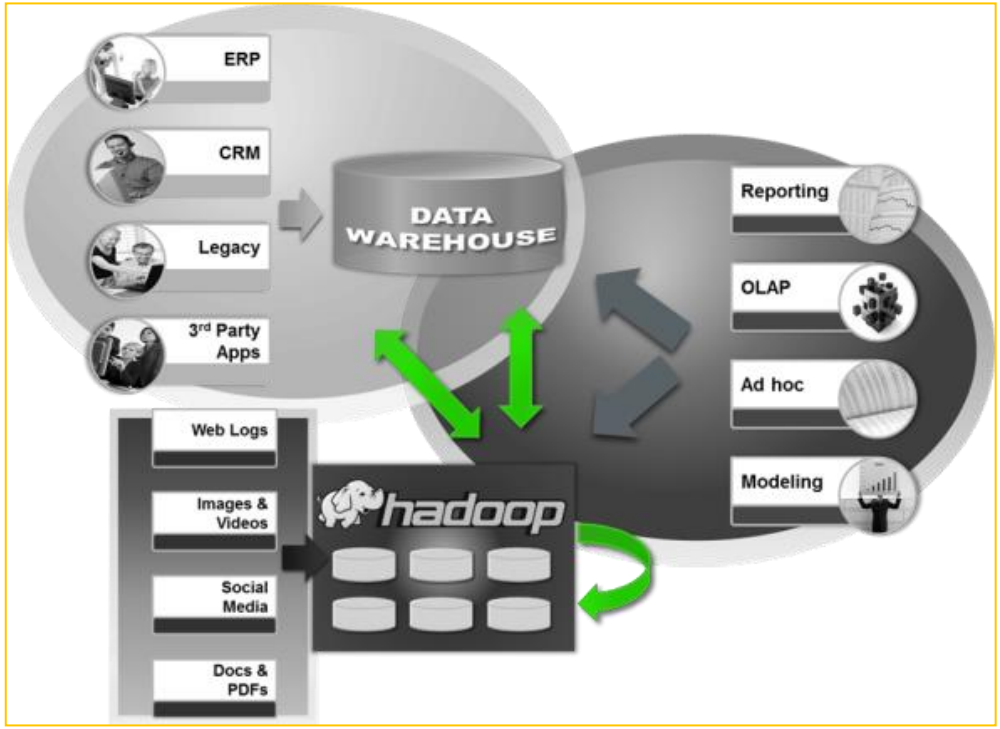
- Big Data and Data Warehouse technologies are optimized for different purposes.
- The goal is to use these solutions for what they were designed to do. In other words: *Use the best tool for the job*.

BUSINESS REQUIREMENT	BIG DATA	DATA WAREHOUSE
Discovery of unexplored business questions	●	●
Clean, consistent, high-quality data	◒	●
Low latency, interactive reports, OLAP	◒	●
Raw, unstructured data	●	
Analysis of preliminary data	●	

Source: http://tamaradull.com/2013/03/20/the-5-ws-when-should-we-use-big-data-vs-data-warehousing-technologies/

Big Data and Data Warehouse coexistence

- EDW and Big Data technologies can co-exist
- Each does what it does best



http://tamaradull.com/2013/03/20/the-5-ws-when-should-we-use-big-data-vs-data-warehousing-technologies/

Using Hadoop in DW Environment (I)

- Use cases of Hadoop in DWs:
  - **To replace/support ETL /ELT processing as a front-end to the DW**
  - **To store the data in HDFS and to directly process raw data**

Source: 5 Steps to Offload your Data Warehouse with Hadoop, Syncsort

Using Hadoop in DW Environment (II)

- Use cases of Hadoop in DWs:
  - **To offload (archive) cold data from the DW (cloud) in the back-end**

Source: http://www.novetta.com/2015/02/why-you-should-offload-your-data-warehouse-to-hadoop/

Using Hadoop in DW Environment (III)

- Use cases of Hadoop in DWs:
  - **To extend EDW as an analytic platform**

Source: **http://jameskaskade.com/?p=2343**

Evolving Data Warehouse in Big Data Environment

- DWs needs to evolve into a **fully enabled data management and information-processing platform**

- To accommodate big data, **Logical DWs** (LDWs) would be common:
  - Deals with multiple data sources, types, and structures
  - Architecture includes: data virtualization, distributed processing, IMDBs
  - Master data management, meta data management

Logical Data Warehouse

- **An integrated, unified data management platform that:**
  - Presents a single interface to all the resources in a data eco-systems
  - Manages data variety and volume of data for both structured and other data types such as machine data, text documents, images and videos.
  - Uses a data virtualization layer for federation of data sources
  - Supports distributed processing with Hadoop

Source: tamaradull, Retrieved from http://tamaradull.com/2013/03/20/the-5-ws-when-should-we-use-big-data-vs-data-warehousing-technologies/

Cloud Data Warehouse

- Implementing a DW system in a cloud
- Major advantages:
  - Pay-as-you-go model; incremental cost($1K per year for 1TB)
  - Implementation time is lower than with on-premises implementation
  - Need less admin skills
- Concerns:
  - Data integration for both cloud and on-premises is challenging
  - Data quality and governance requirement
- Amazon Redshift is at the front
- Adoption will grow as Microsoft, Teradata, and SAP enter the market

Source: tamaradull, Retrieved from http://tamaradull.com/2013/03/20/the-5-ws-when-should-we-use-big-data-vs-data-warehousing-technologies/

Data Integration Challenges

- **Real-world data are dirty**
- **Diversity**, complexity and variability of organizational data (Web, Social, mobile, cloud, on-premise, sensors, hybrid integration, real-time integration)
- Tracking **data provenance**, from data generation through data preparation
- **Capabilities**
  - **Connectivity capabilities (data source and target support)**
  - **Data transformation capabilities**
  - **Data governance support capabilities (quality, metadata)**
- Too many choices of data integration tools, techniques and architecture
  - Difficult to create a clean roadmap for smooth evolution

Source: Gartner, Research Library: Fundamentals for Data Integration Initiatives

Real-Time Data Integration

- Continuous sourcing of data with Always-On listening
- Variable data capture modes
  - time-based, event-based, different granular levels
- In-Line data transformation and cleansing
- Identifying and acting on changes and events as they cccur
- Operation and management of a real-time data integration environment
- Copying with the need for real-time analysis and decision-making

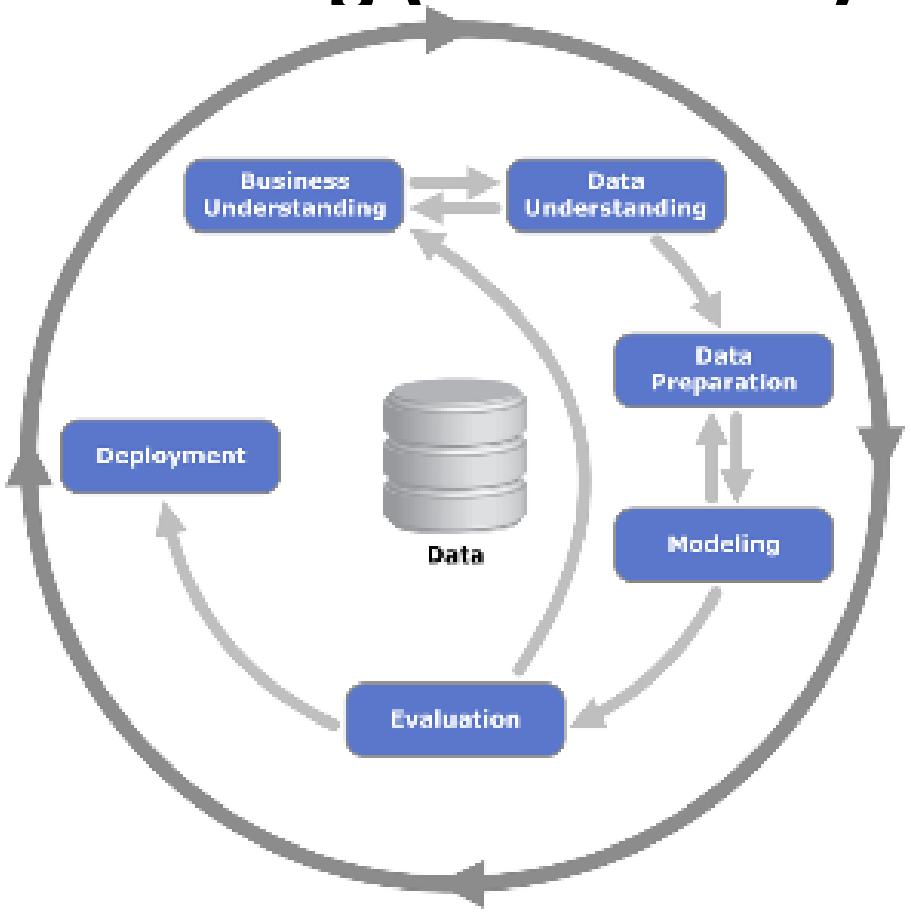
Table of Contents

- What is Big Data?
- Data-driven paradigm (DDP)
- Big Data Technologies
  - Hadoop Ecosystems
  - NoSQL databases; NewSQL databases
  - In-memory databases
  - Cloud computing
  - Big data warehousing (ETL, ELT, and Data virtualization, EDW, LDW, Data Integration)
- **Values and Use cases**
- Research Issues
- Conclusions

Getting Values in Big Data Projects

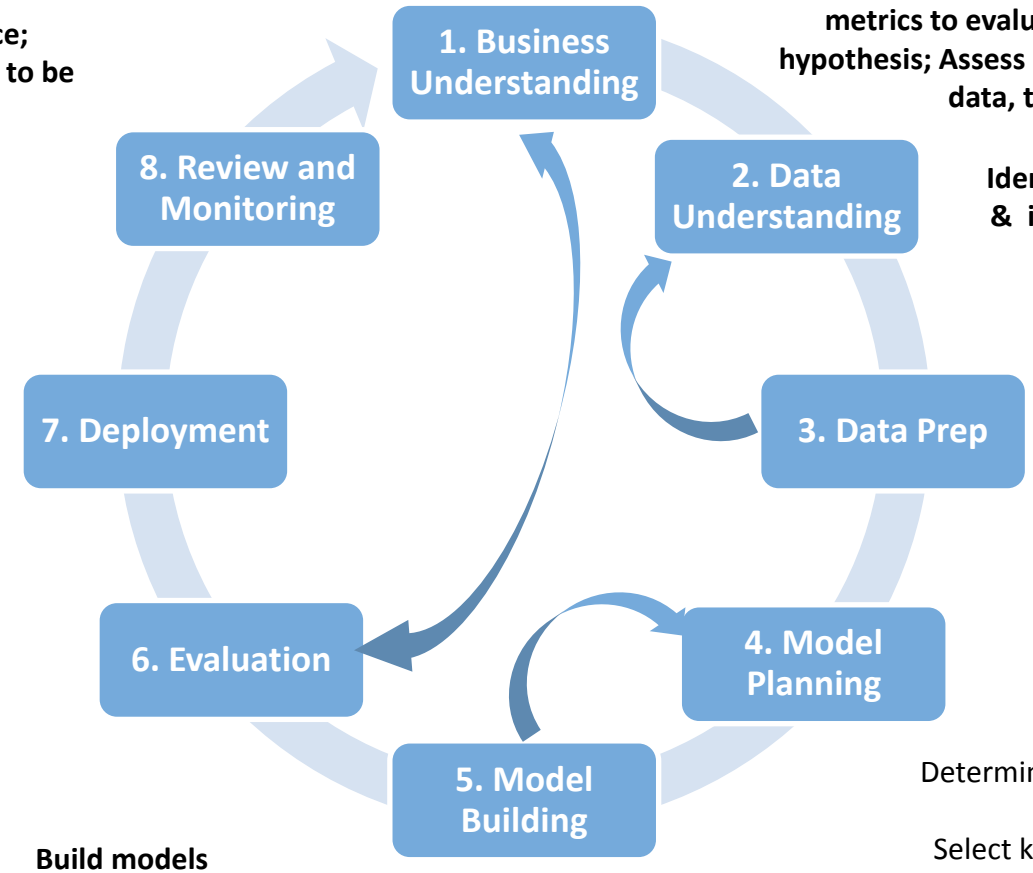
1. Data-driven culture
   - Look at the data dispassionately, without replying on intuition
   - Ex: Apple, Google, Wal-mart, with their supply-chain management
2. Analytical talent around data science
   - Innovations can come from doing diagnostics, predictive, and prescriptive analytics
3. Understanding on big data solutions
   - Hadoop ecosystems, NoSQL
4. Security
5. Leadership
   - Data analytic life cycle: from business strategy, questions, data, solution, evaluation, monitoring): CDO disciplines

Cross Industry Standard Process for Data Mining (CRISP-DM)



Source: Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Data Analytics Lifecycle



What is the question to solve and metrics to evaluate? Generate hypothesis; Assess resources (people, data, tools).

Identify data resources; data reuse & integration plan, datatification, decision on tools

Acquire data; Perform data profiling,, cleanse, and transform,; Explore data and verify quality

Determine the methods, techniques, and workflow**?**
Select key variables and determine correlation between them

Build models
Perform analysis and iterate

Perform evaluation against metrics;
Communicate results and recommendations

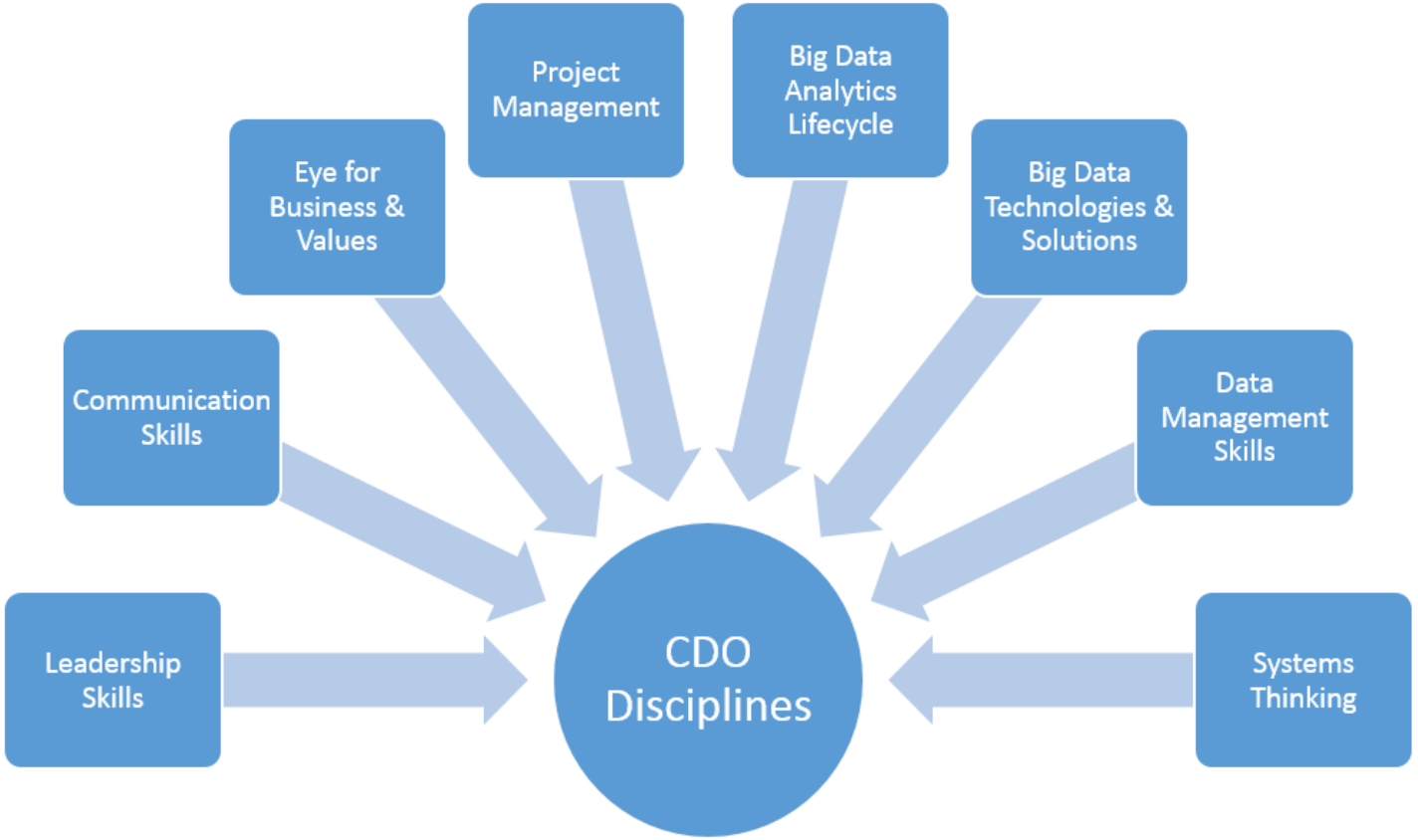
Integrate analytics procedures into management dashboards and operational systems.

Monitor performance;
Identify parts that need to be improved

Chief Data Officer (CDO)

- A CDO provides **vision** and **strategy** for all data management initiatives:
  - Is a champion for **global data policies, standards, governance, quality**, data source management, education, and vendor relationships across the enterprise
  - Identifies **business questions** and **metrics** in business context
  - **Oversees big data project roadmap and workflows** from conceptual analysis to deployment

Who is a Chief Data Officer (CDO)?



Values in Big Data Projects

- **Automate decision-making**
- **Generate deeper business insights**
- **Optimize**
- **Design new processes**

-- (Gartner, 2014)

Use Cases in Big Data Projects

- **Marketing and sales growth**
- **Operational and financial performance improvement**
- **Risk and compliance management**
- **New product and service innovation**
- **Direct/Indirect data monetization**

-- (Gartner, 2015)

Big Data Use Cases

- Healthcare
  - Early detection of a disease (e.g., Alzheimer)
  - Customized drugs based on patient's history
  - Early detection of epidemics with crowdsourcing
  - Smart health projects with care networks for older people
  - Integrating genomic analysis with healthcare
  - Disease prevention, flu forecast and prevention
  - Detecting abnormal situations in ICU
  - IBM Watson (Seton Health Care Family use Watson to learn 2M patient data annually)

Big Data Use Cases

- **Customer analysis**
  - Personalized coupon
  - Fraud detection for IRS, social security claims
  - Churn analysis
  - Better user profiling, more targeted marketing
  - More customized (optimized) pricing and automated bidding on a number of exchanges
  - Geo-marketing via cell phone (restaurants, retail)
  - Expand your existing customer analytics with social media data
    - They influence each other

Big Data Use Cases

- **Web**
  - Better taxonomies to classify text (news, user reports, published articles, blog posts, yellow pages etc.)
  - Detection of duplicate and fake accounts on social networks
- **Education and Academic**
  - Customized, on-demand, online education with automated grading
  - Better detection of fake reviews for systems based on collaborative filtering (in short, superior collaborative filtering technology)
- **Crime prevention**
  - Criminal protection by predicting likely locations of criminal activities

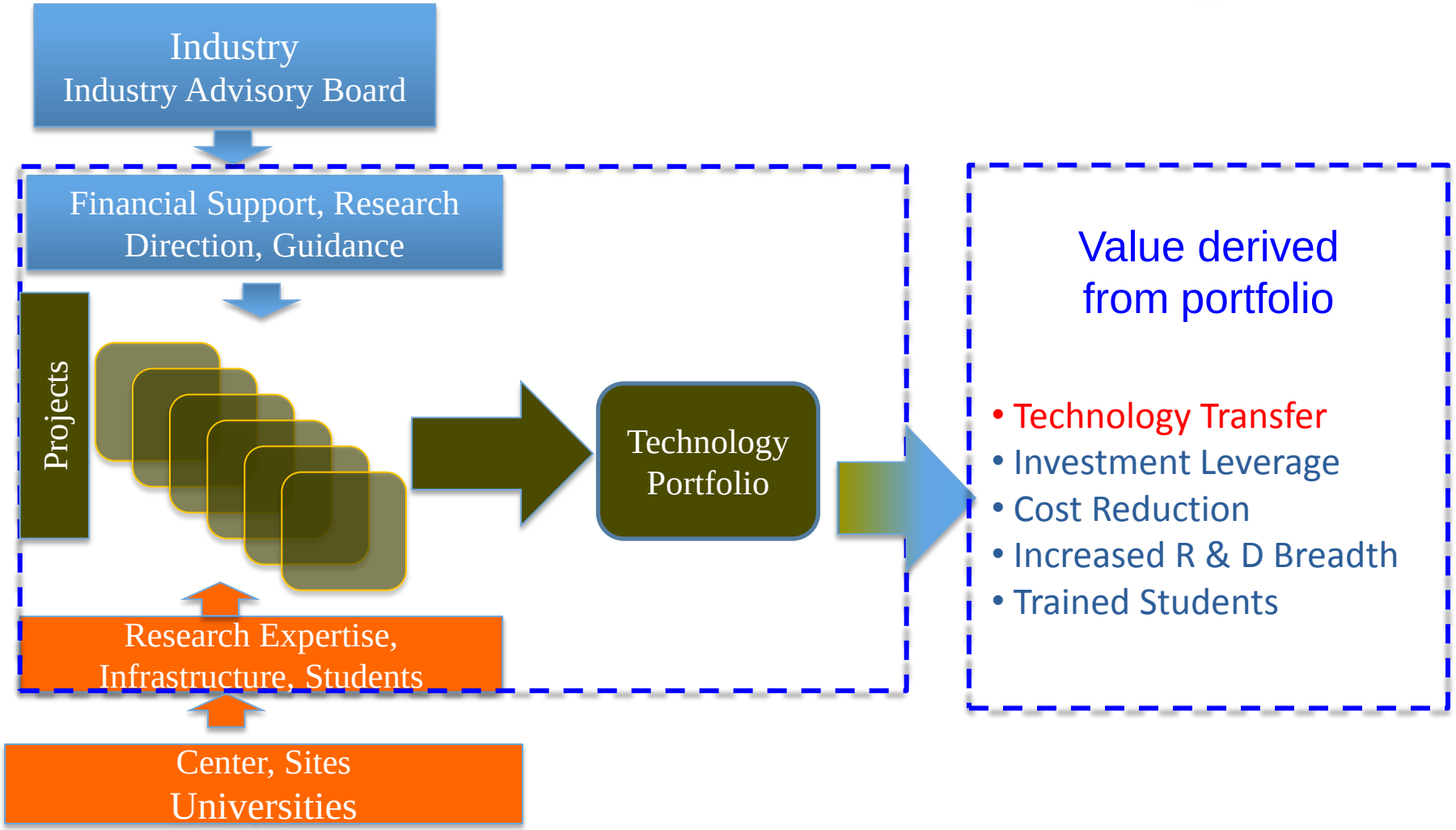
Big Data Use Cases

- **Disaster/risk prevention/detection**
  - Fire prevention based on geodata, household data, lifestyle data
  - Preparation against Tsunami, taipoon
  - Detection of earthquakes, solar flares - including intensity forecasting
- **Public service improvement**
  - Optimizing electricity trading and delivery
  - Smart utility meters with sensors
  - Drafting new laws from complaints and social phenomenon
  - Personalized labor support system (Germany, saving 10B euro saving)
  - Compliance detection using its event management solution (HP)
- **Terror prevention**
- **Defense application**

Table of Contents

- What is Big Data?
- Data-driven paradigm (DDP)
- Big Data Technologies
  - Hadoop Ecosystems
  - NoSQL databases; NewSQL databases
  - In-memory databases
  - Cloud computing
  - Big data warehousing (ETL, ELT, and Data virtualization, EDW, LDW, Data Integration)
- Values and Use cases
- **Research issues**
- Conclusions

NSF I/UCRC Concept & CVDI



NSF Center for Visual & Decision Informatics

- CVDI Vision Statement
  - Drive continuous innovation through knowledge sharing among partners leading to invention and commercialization of decision-support technologies.

- Key Capacities of CVDI
  - Big data management
  - Data mining, Data warehousing, OLAP
  - Data integration, ontology
  - Predictive analytics
  - Gap analysis
  - Information visualization
  - Visual analytics
  - Automated data analysis
  - Social media analysis
  - Bioinformatics & healthcare informatics

CVDI 2012 Projects

- **Social Media**

  1. Social Media for Decision Informatics with Application to Emerging Events.

  2. Social Media for Decision Informatics with Application to Healthcare Management.

- **Data Mining and Data Integration**

  3. Multi-Industry Semantic Discovery Tool Sets for Data Integration, Data Warehousing, and e-Science.

- **Visualization and Visual Analytics**

  4. Visualization of Multivariate Spatiotemporal Data.

  5. Real-Time Analysis and Visual Exploration of Multi-Dimensional Sensor Data.

CVDI 2013 Projects

- **Semantic Information Extraction, Integration, and Visualization for Big Data Analytics**
- **Large-scale Social Media Analytical Tools with Application to Detecting Emerging Events**
- **Visual Analytic Approaches to Mining Large-scale Time-evolving Graphs**
- **A Spatio-Temporal Data Mining Approach for Fraud Detection**
- **Scalable Visualization, Gap Analytics and Link Prediction for Multiple Big Data Industry Sectors**

CVDI 2014 Projects

- **Novel Methods for Hidden Relation and Error Detection from Structured and Unstructured Data**
- **Multi-Level and Multi-Source Visual Analytics of Evidence-Based Knowledge Diffusion Processes**
- **Visual Analytic Approaches for Mining Large-Scale Dynamic Graphs**
- **A Predictive Analytics Framework for Spatio-Temporal Hotspots**
- **Analyzing, Modeling, and Summarizing Social Media and Linked Data Sets**

Healthcare Research Projects:
U.S. Healthcare Expenditures

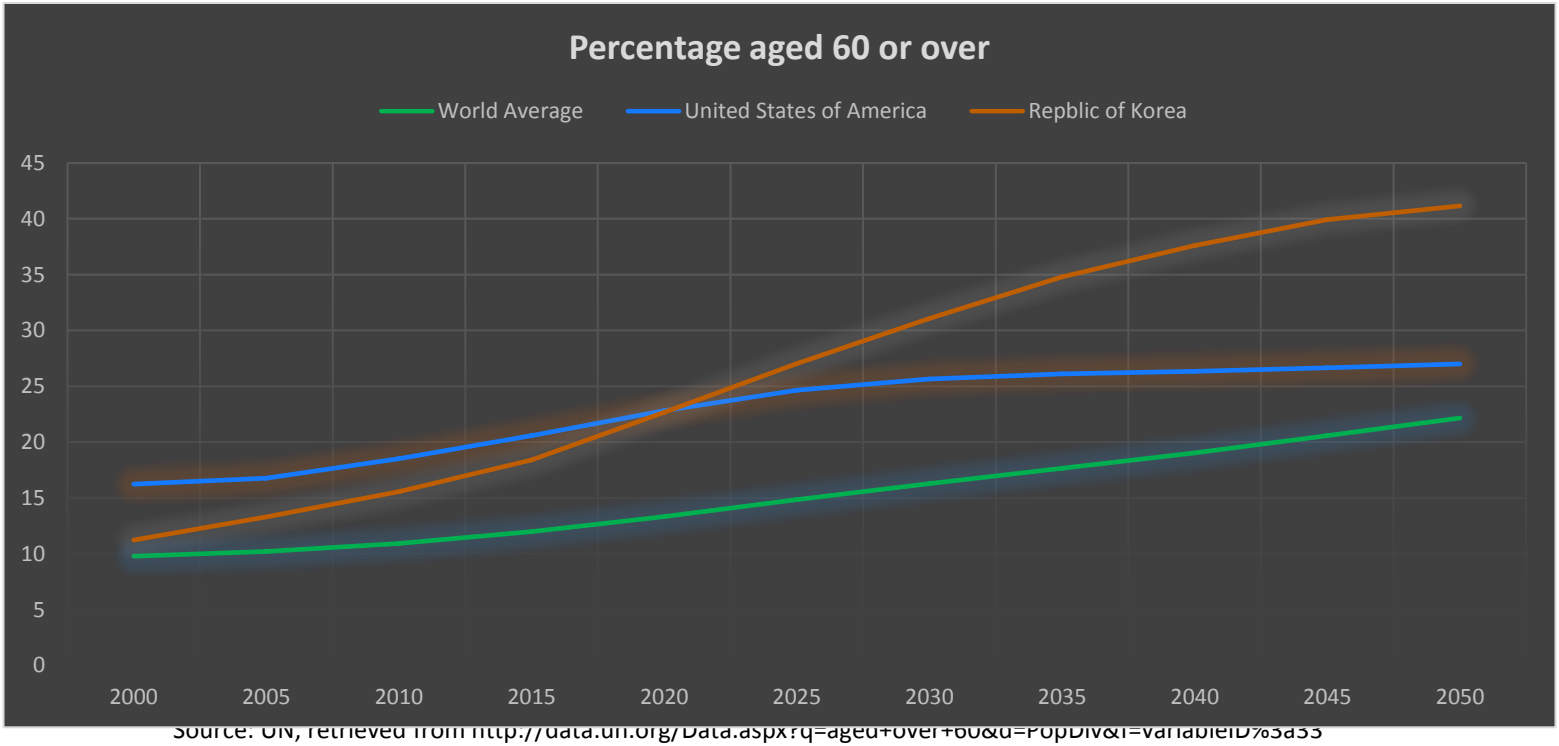
- U.S. healthcare costs exceed 17% of GDP (Harvard Business Review, 2011).

- If US healthcare were to use big data creatively and effectively to drive efficiency and quality, the sector could create more than $300 billion in value every year (McKinsey, 2011).

Source: [1]Harvard Business Review, https://hbr.org/2011/09/how-to-solve-the-cost-crisis-in-health-care/ar/1
[2] McKinsey, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Aging Data

- Population aged 60 or over rapidly increasing!
  - In Korea, 25% by 2022; 33% by 2034
  - In US, 25% by 2027
  - World average, 15% by 2025

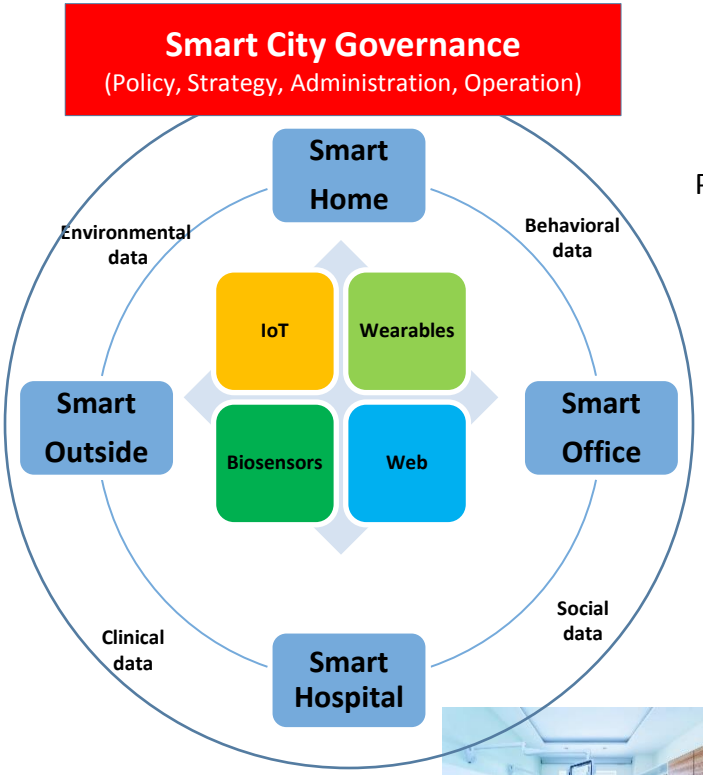


Source: UN, retrieved from http://data.un.org/Data.aspx?q=aged+over+60&d=PopDiv&f=variableID%3a33

Smart Health Enabled Smart City

Activity Tracking
Remote Patient Monitoring
Personalized Health Assistance
Food Consumption Monitoring

Activity Tracking
Medical Facilities Hotspotting
Real-Time Traffic Monitoring
Air Quality Monitoring



Activity Tracking
Prevention of Being Sedentary
Psycho-social Mental Health Improving
Peer-Interaction System

Real-Time Vital Monitoring
Adaptive Treatment Planning
Prescription Adherence Improving
Wait-Time Management
Readmissions Management
Genomic Analysis

Components of Smart Aging Technologies

Wearable devices

Mobile applications

Monitoring

EHR and Healthcare data

Big Data Generation Layer

Alerting

Drexel Data Science Institute

BIG DATA ANALYTICS

INSTITUTE FOR HEALTHCARE INFORMATICS

Drexel UNIVERSITY

hadoop

R

Neo4j the world's leading graph database

human factors

SOCIAL SCIENCE DEPARTMENT

Big Data Analysis Layer

Table of Contents

- What is Big Data?
- Data-driven paradigm (DDP)
- Big Data Technologies
  - Hadoop Ecosystems
  - NoSQL databases; NewSQL databases
  - In-memory databases
  - Cloud computing
  - Big data warehousing (ETL, ELT, and Data virtualization, EDW, LDW, Data Integration)
- Values and Use cases
- Research issues
- **Conclusions**

Conclusions (I)

- Big data technologies are too complex
  - ✓ Needs to evolve
- Big data technologies address issues on Web, Mobile, Social, Cloud, and Big Data Analytics
- Hadoop ecosystems will evolve
  - ✓ Lots of hypes, but slow acceptance
  - ✓ Still less than 1% of big data market
- Hadoop and EDW will co-exist
  - ✓ The EDW is evolving into an integrated DM platform
- Big data driving force: IoT, Smart City

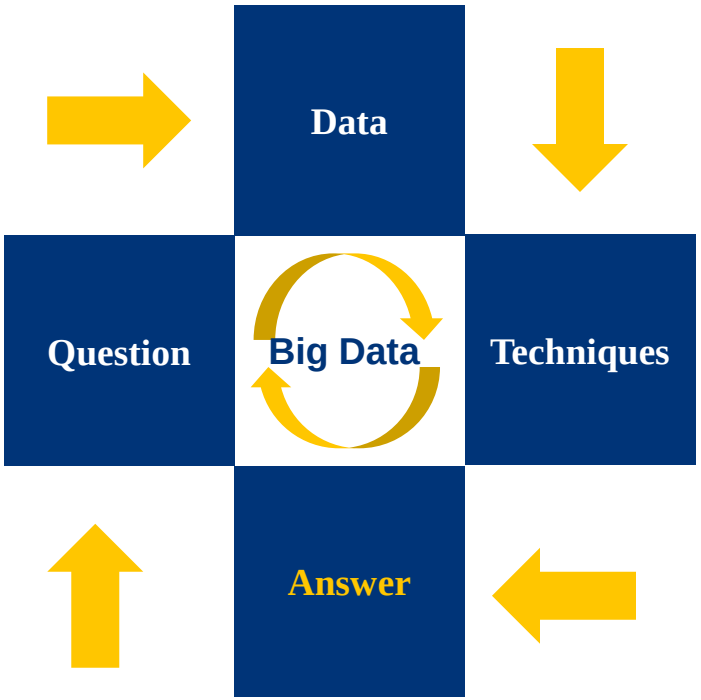
Conclusions (II)

- Discovering value is challenging
  - ✓ Real-world data are dirty and projects are complex!
  - ✓ Analytics: Need data scientists!
  - ✓ Analytic life cycle: Chief data officers (CDOs)

- Answering Big Data questions is an **interactive and explorative process**
  - ✓ **Begin big data projects as experiments**
  - ✓ Prove feasibility and value first;
  - ✓ Operationalize only after proven value from business, legal compliance, economical, and technological standpoints

Problem-Solving in Big Data

Big data problems are also CS problems



- Create a business case
- Look at the input data
  - Merge/combine multiple data sources
- Specify the desired output data,
- and think hard about whether and how you can compute the desired result

- Interpret the results
- Look for new angles in knowledge discovery