

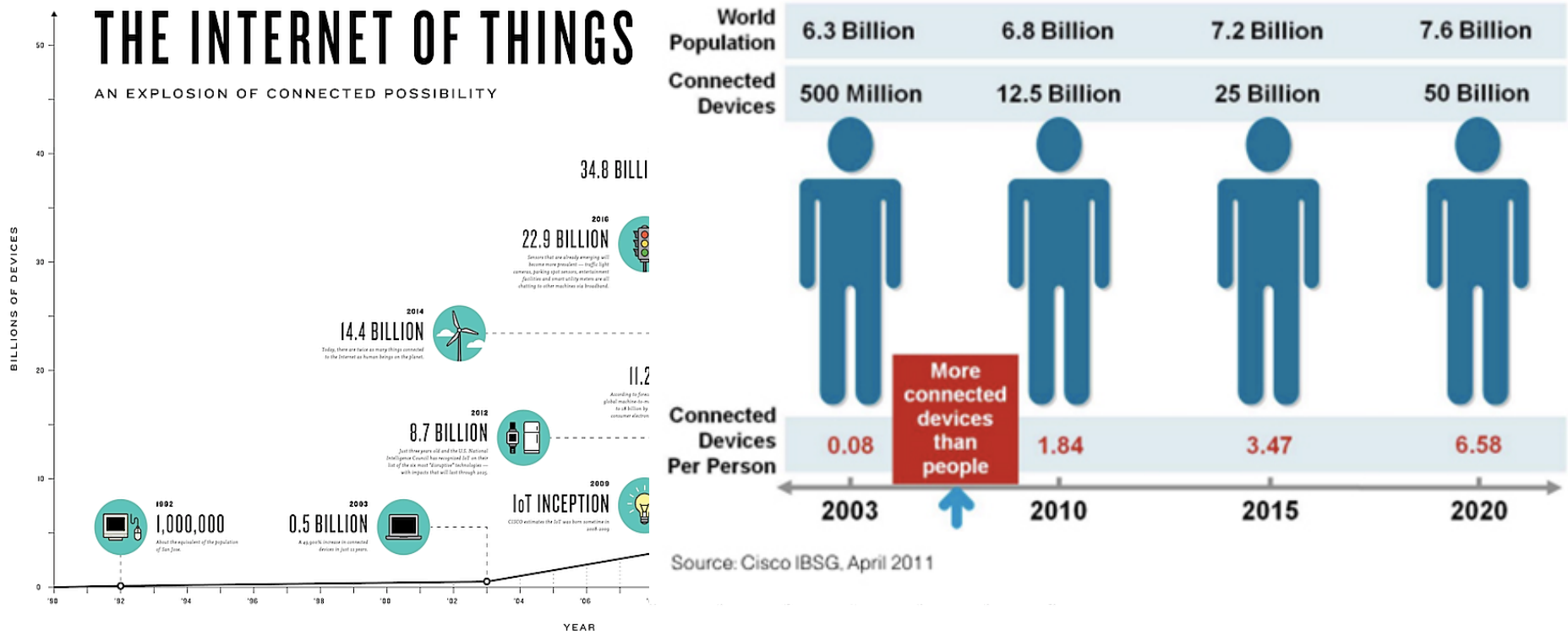
Big Data and IoT

Baris Aksanli

02/10/2016



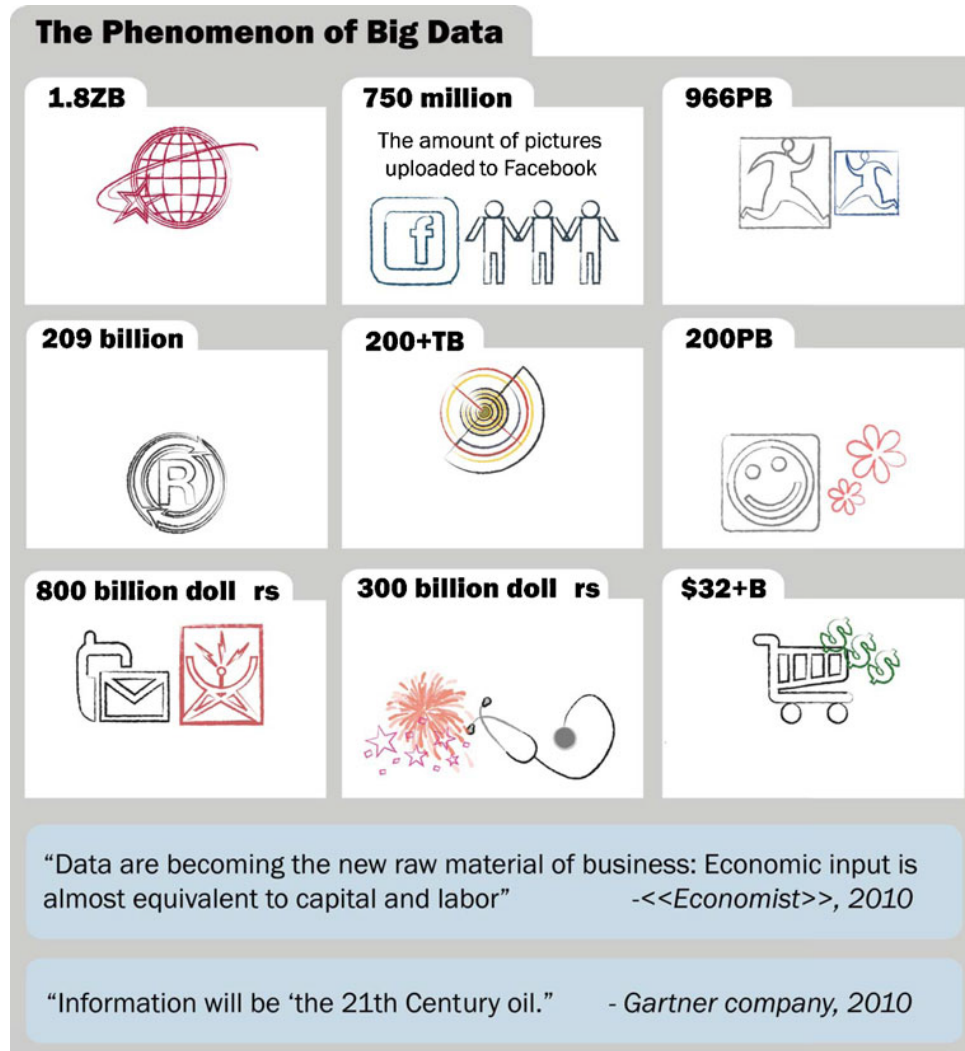
Why is there big data?



- Number of devices increasing exponentially
 - They continuously generate data
 - For example, on average, 72 hours of videos are uploaded to YouTube in every minute.

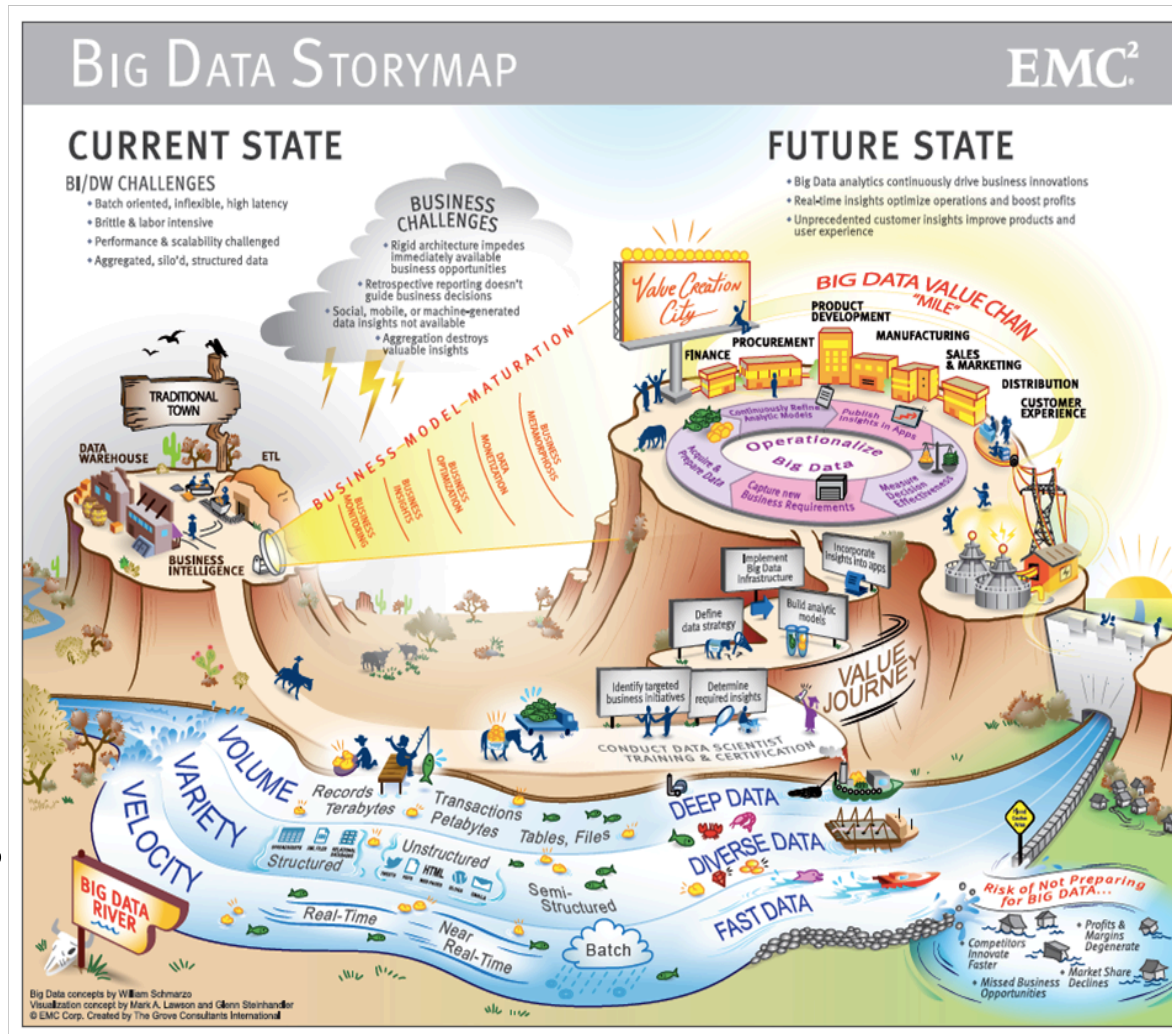
How much data is big?

- 2010: Apache Hadoop: “datasets which **could not be** captured, managed, and processed by **general computers** within an **acceptable scope**.”
- **3V** model: Volume, Velocity, Variety [META]
— +1V: Value [IDC]



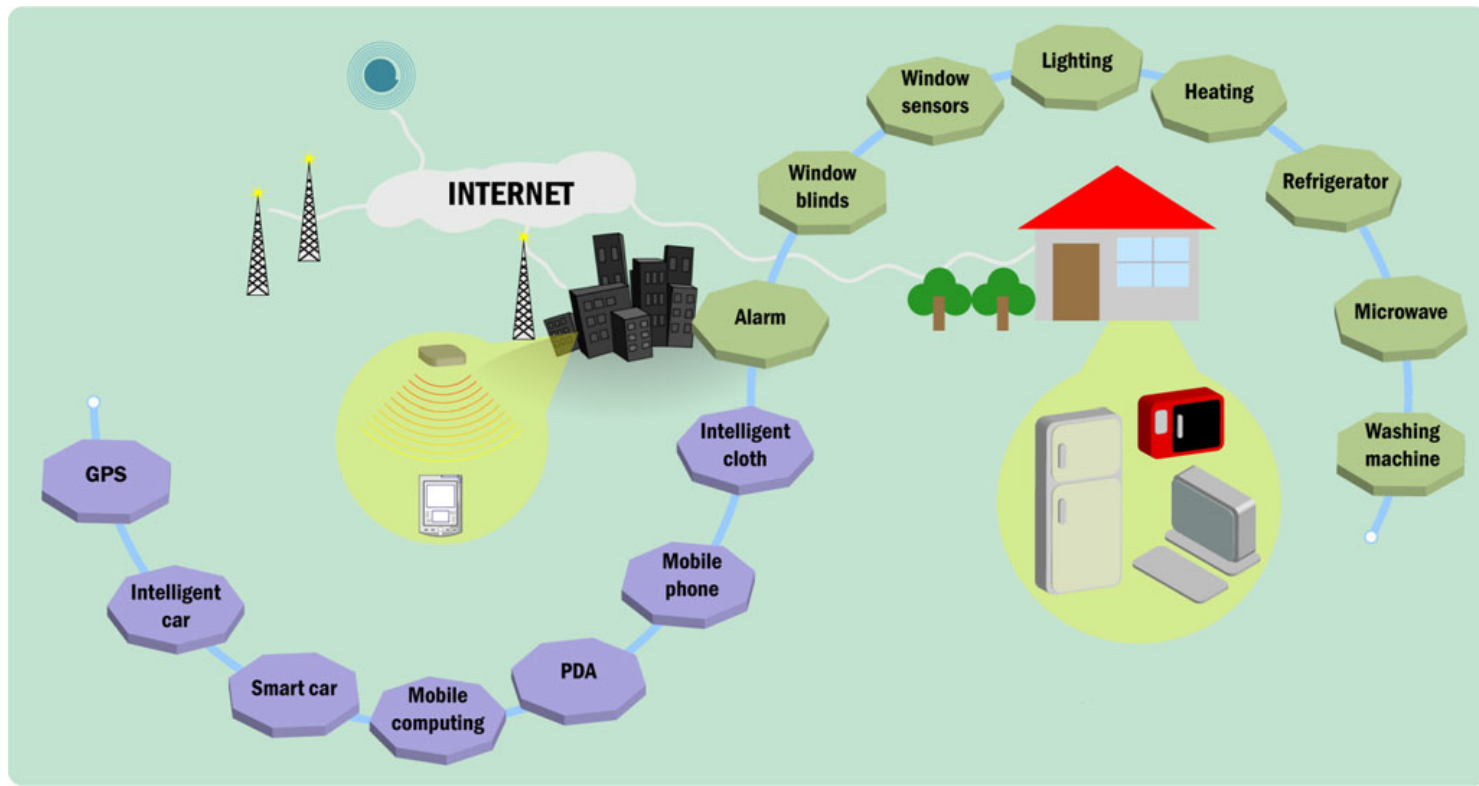
Value of Big Data

- New business and efficiency opportunities
- \$300B in US medical industry
- Increased efficiency of government operations
- Search engines personalized for users
- Personalized ads, products, etc.

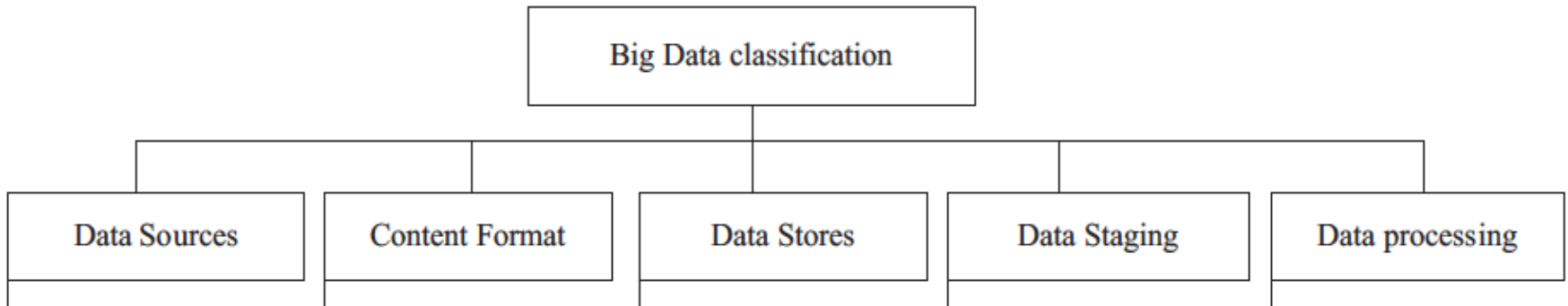


IoT and Big Data

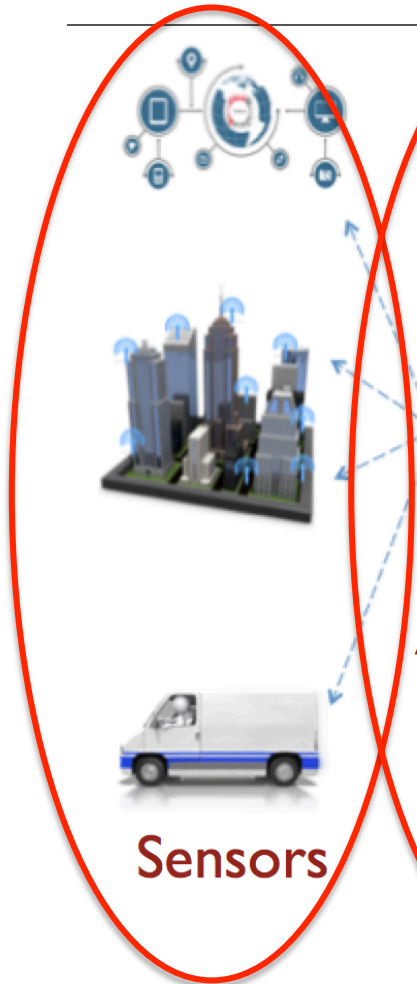
- IoT applications continuously generate data
 - Even the smallest device generates data
- **The problem: data processing capacity is lower than data generation speed**



Big Data Classification



Path of the Data



Data collection
& acquisition

Data transfer

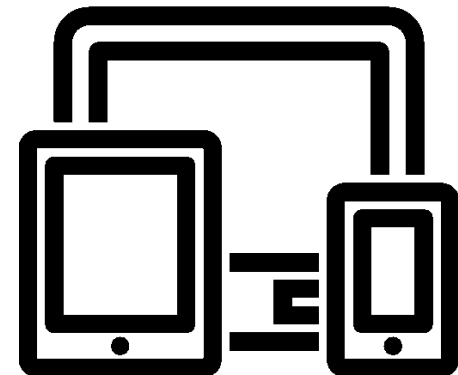
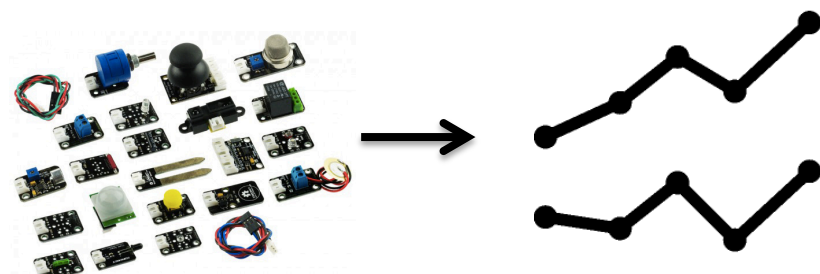
& analysis

Data Generation

- Enterprise data: big companies, e.g. Facebook, Amazon
 - Business data is expected to double every 1.2 years
 - Walmart processes 1M customer trades/hour
 - Akamai advertises 75M events/day
- IoT data: pervasive applications, clinical medical care-R&D
 - Large scale, heterogeneous and strongly correlated data
 - 30 billion RFID tags and 4.6 billion camera phones are used around the world today
 - If Wal-Mart operates RFID on item level, it is expected to generate 7 terabytes (TB) of data every day
- Bio-medical data: human gene sequencing
 - One sequencing of human gene may generate 100 sequences of 600GB raw data
- Other areas: physics, bio-informatics, etc.
 - Astronomy: Sloan Digital Sky Survey (SDSS), the data volume generated per night surpasses 20TB

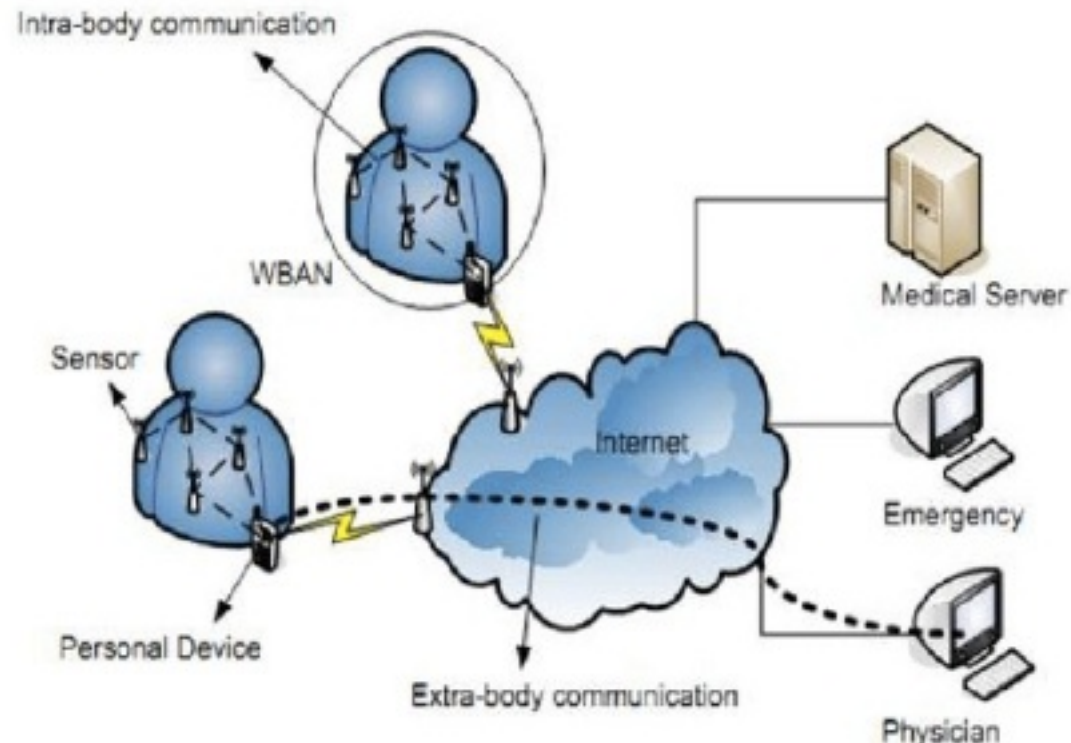
Data Acquisition

- Log files: almost all digital devices provide logging capability
 - Web activity recording, financial applications, network monitoring
- Sensing: physical quantities into readable digital signals
 - Sound wave, voice, vibration, automobile, chemical, current, weather, pressure, temperature, etc.
 - Localization
- Mobile platforms: similar to sensing
 - More personalized, specific to a user



Data Transportation

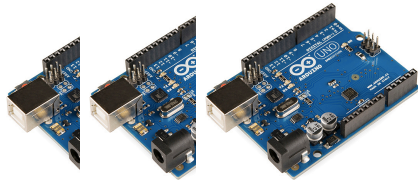
- Data transfer to a storage infrastructure for processing and analysis
- Inter data center network (DCN) transmissions:
 - Source to data center
 - Using WAN: 40-100Gbps
- Intra DCN transmissions:
 - Data center interconnect
 - Top-of-the-rack vs. aggregator switches
 - 1-10-100 Gbps



Data Preprocessing

- Eliminate or reduce redundancy, noise, meaningless data
 - Increase storage efficiency, data analysis speed
- Integration: combining data from different sources
 - Data warehouse: ETL (Extract, Transform and Load)
 - Data federation
 - Mostly used by search engines
- Cleaning: how can data be cleaned?
 - Define error types -> identify errors -> correct errors -> document errors -> modify infrastructure to prevent errors
- Redundancy elimination
 - Redundancy detection, data filtering, data compression
 - Areas: Images, videos
 - One solution: Compression!

Preprocessing Capabilities



Arduino
16 MHz
32KB flash



Network
speed: 1Gpbs



Raspberry Pi 2
600 MHz
1GB Ram



Network
speed: 10Gpbs



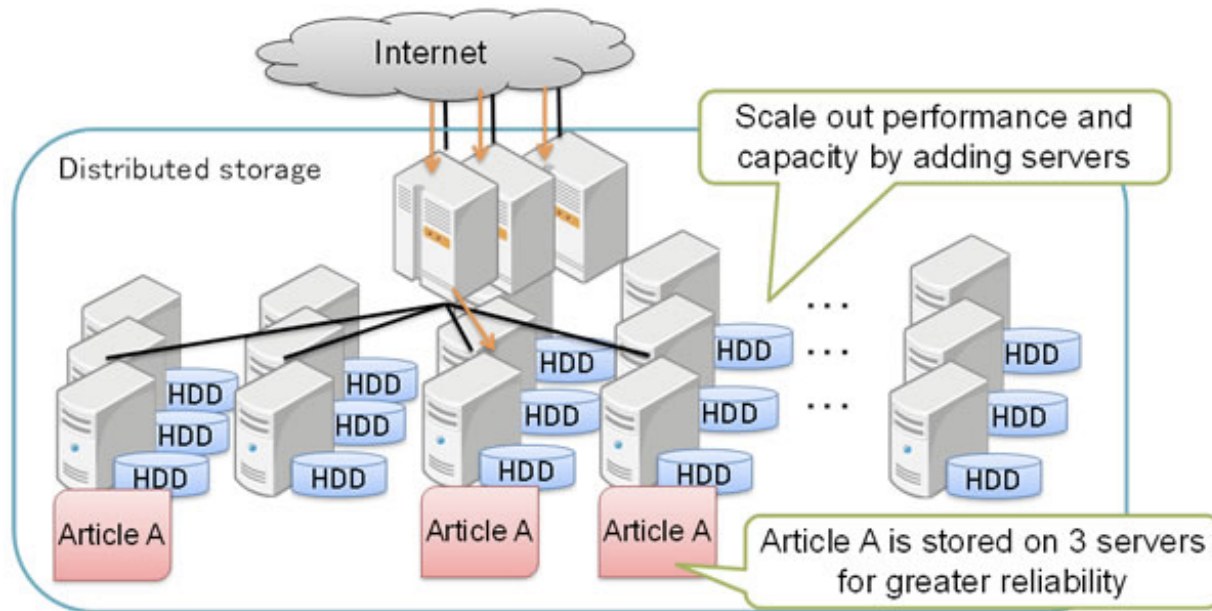
Commodity server
3 GHz
32 GB Ram

- Assume there is a job with 1TB total size
- 100K Arduino, 1K Raspberry Pi 2, 100 servers
- Time spent in computation vs. networking
 - Arduino level
 - Raspberry Pi 2 level
 - Server level

Big Data Storage

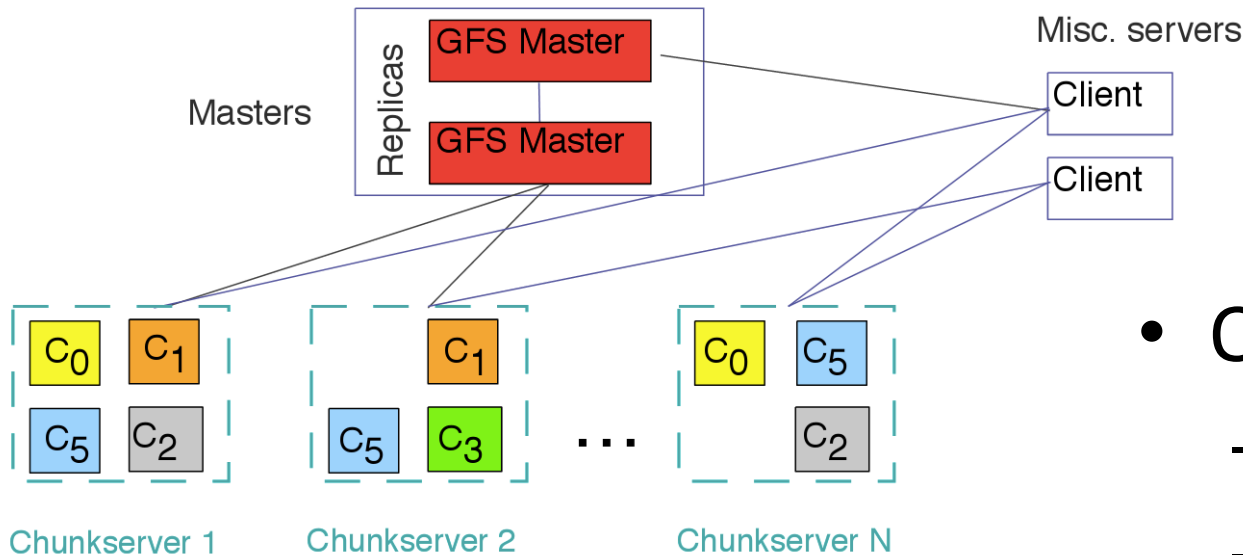
- Storage and management of **large-scale** data sets while achieving **reliability** and **availability** of data accessing
 - Traditionally on servers with structured RDBMSs.
- Existing storage systems for massive data
 - Direct attached storage (DAS)
 - Several hard disks directly connected with servers
 - Only suitable to interconnect servers with a small scale
 - Network attached storage (NAS)
 - NAS utilizes network to provide a union interface for data access and sharing
 - The I/O burden is reduced extensively since the server accesses a storage device indirectly through a network
 - Storage area network (SAN)
 - Designed for data storage with a scalable and bandwidth intensive network
 - Data storage management is relatively independent within a storage local area network

Distributed Storage System



- CAP: Consistency – Availability – Partition tolerance
 - At most two of the three requirements can be satisfied simultaneously
- CA vs. CP vs. AP systems
 - CA: for single servers
 - CP: useful for moderate load [BigTable and Hbase]
 - AP: useful when no high demand on accuracy [Dynamo and Cassandra]

File systems for Big Data



- Google file system (GFS)
 - File broken into chunks (typically 64MB)
 - Master manages metadata
 - Data transfers happen directly between clients and chunkservers

- Other examples:
 - HDFS and Kosmos
 - Extensions to GFS
 - Cosmos from MS
 - Haystack from FB

Database Technology

- Key-value databases: data is stored corresponding to unique key-values -> shorter query response time
 - Provide expandability by distributing key words into nodes
 - Dynamo [Amazon] and Voldemort [LinkedIn]
- Column-oriented databases: store and process data according to columns rather than rows
 - Both columns and rows are segmented in multiple nodes to realize expandability
 - BigTable [Google] and Cassandra [Facebook]
- Document databases: can support more complex data forms and key-value pairs can still be saved
 - Structured data storage with objects
 - MongoDB [Binary JSON objects], SimpleDB [Amazon] and CouchDB [Apache]

Programming Models

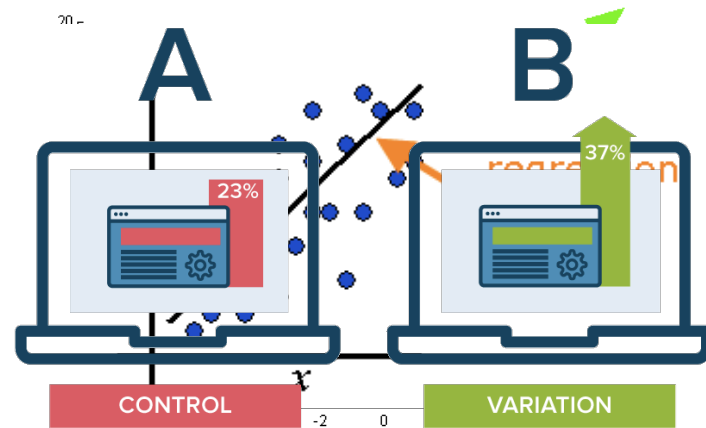
- Traditional parallel models do not perform well
 - Scalability issues: big data are generally stored in hundreds and even thousands of commercial servers

Table 1 Comparison of MPI, MapReduce and Dryad

	MPI	MapReduce	Dryad
Deployment	Computing node and data storage arranged separately (Data should be moved computing node)	Computing and data storage arranged at the same node (Computing should be close to data)	Computing and data storage arranged at the same node (Computing should be close to data)
Resource management/ scheduling	–	Workqueue(google) HOD(Yahoo)	Not clear
Low level programming	MPI API	MapReduce API	Dryad API
High level programming	–	Pig, Hive, Jaql, ...	Scope, DryadLINQ
Data storage	The local file system, NFS, ...	GFS(google) HDFS(Hadoop), KFS Amazon S3, ...	NTFS, Cosmos DFS
Task partitioning	User manually partition the tasks	Automation	Automation
Communication	Messaging, Remote memory access	Files(Local FS, DFS)	Files, TCP Pipes, Shared-memory FIFOs
Fault-tolerant	Checkpoint	Task re-execute	Task re-execute

Data Analysis

- Goal is to extract useful values, w/suggestions or decisions
- Traditional data analysis
 - Cluster analysis: grouping objects
 - Factor analysis: describe the relation among many elements with a few factors
 - Correlation analysis: dependence among variables
 - Regression analysis: dependence relationships among variables hidden by randomness
 - A/B testing: improve target variables by comparing the tested group
 - Statistical analysis: summarize and describe data sets



Big Data Analytics

- Bloom filter: using hash functions to conduct lossy compression storage of data
 - High space efficiency and high query speed
- Hashing: transforms data into shorter fixed-length numerical values or index values
 - Rapid reading but hard to find a good hash function
- Index: fast data retrieval and modification
 - Additional cost for storing index files which should be maintained dynamically when data is updated
- Trier: trie tree, a variant of hash tree
 - Fast string operations
 - Leverage common prefixes of character strings to reduce comparison on character strings

Tools for Big Data Analysis



- The top five most widely used software, according to a survey of “What Analytics, Data mining, Big Data software that you used in the past 12 months for a real project?” of 798 professionals made by KDNuggets in 2012
- R [30.7%]
- Excel [29.8%]
- Rapid-I Rapidminer [26.7%]
- KNMINE [21.8%]
- Weka/Pentaho [14.8%]

Summary

- Big data is different than traditional massive data
 - Cannot be processed by general computers within acceptable time
 - Why big data is an inevitable result of the IoT
- The basics of big data and analytics
 - Data generation/acquisition
 - Data storage
 - Data analytics
- Many systems built to address a different aspect of big data