Big data: architectures and data analytics

Introduction to Big Data

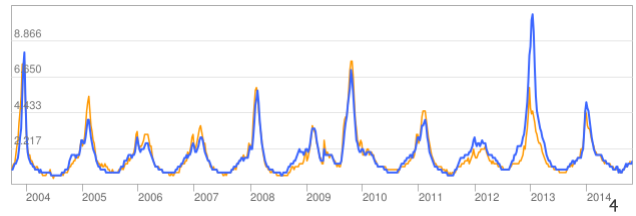
Based on the slides of Elena Baralis "Big Data: Hype or Hallelujah?"
http://dbdmg.polito.it/wordpress/wp-content/uploads/2010/12/BigData_2015_2x.pdf

Big data

3

Google Flu trends

- February 2010
  - Google detected flu outbreak two weeks ahead of CDC data (Centers for Disease Control and Prevention – U.S.A)
  - Based on the analysis of Google search queries

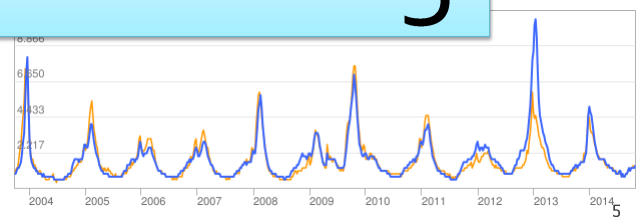


4

Google Flu trends

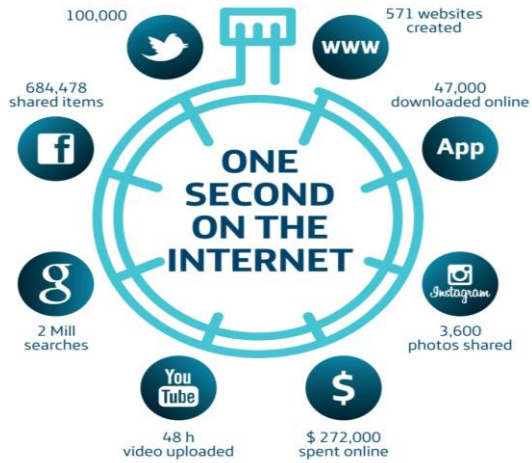
- February 2010
  - Google detected flu outbreak two weeks ahead of CDC data (Centers for Disease Control and .A)
  - lysis of ueries

Nowcasting



5

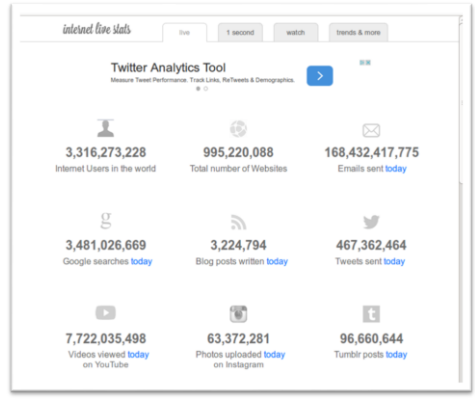
Data on the Internet…



6

Data on the Internet…

- Internet live stats
  - http://www.internetlivestats.com/

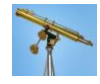


7

Who generates big data?

- User Generated Content (Web & Mobile)
  - E.g., Facebook, Instagram, Yelp, TripAdvisor, Twitter, YouTube

- Health and scientific computing

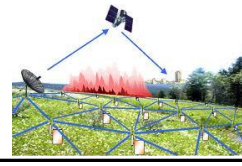


8

Who generates big data?

- Log files
  - Web server log files, machine system log files

- Internet Of Things (IoT)
  - Sensor networks, RFID, smart meters



9

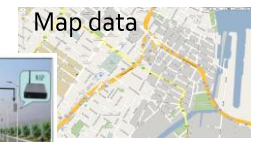
An example of Big data at work

- Crowdsourcing

Computing

Sensing

Map data



Real time traffic info

10

What is big data?

- Many different definitions
  - "Data whose scale, diversity and complexity require new architectures, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it"

11

What is big data?

- Many different definitions
  - "Data whose **scale**, **diversity** and **complexity** require new architectures, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it"

12

What is big data?

- Many different definitions
  - "Data whose scale, diversity and complexity require new **architectures**, **techniques**, **algorithms** and **analytics** to manage it and extract value and hidden knowledge from it"

13

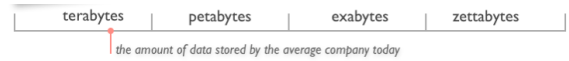
The Vs of big data

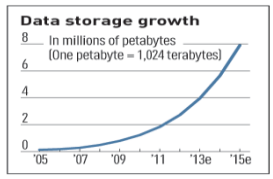
- The 3Vs of big data
  - **V**olume: scale of data
  - **V**ariety: different forms of data
  - **V**elocity: analysis of streaming data
- … but also
  - **V**eracity: uncertainty of data
  - **V**alue: exploit information provided by data

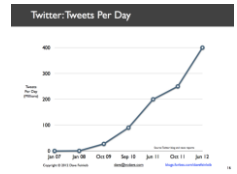
14

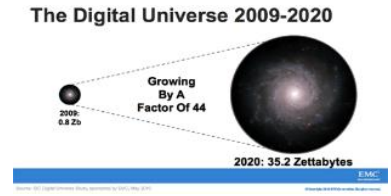
The Vs of big data

- Volume
  - Data volume increases exponentially over time
  - 44x increase from 2009 to 2020
    - Digital data 35 ZB in 2020





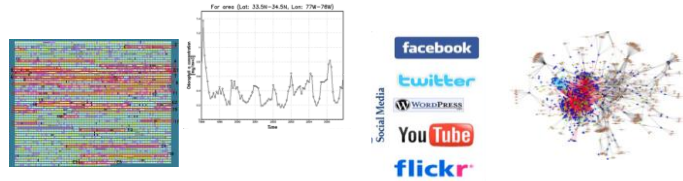




15

The Vs of big data

- Variety
  - Various formats, types and structures
    - Numerical data, image data, audio, video, text, time series



  - A single application may generate many different formats
    - Heterogeneous data
    - Complex data integration problem

16

The Vs of big data

- Velocity
  - Fast data generation rate
    - Streaming data
  - Very fast data processing to ensure timeliness

17

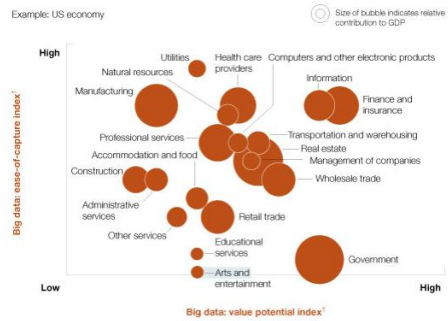
The Vs of big data

- Veracity
  - Data quality

18

The Vs of big data

- Value
  - Translate data into business advantage



[1]For detailed explication of metrics, see appendix in McKinsey Global Institute full report *Big data: The next frontier for innovation, competition, and productivity*, available free of charge online at mckinsey.com/mgi.
Source: US Bureau of Labor Statistics; McKinsey Global Institute analysis

19

Big data value chain

Generation → Acquisition → Storage → Analysis

- Generation
  - Passive recording
    - Typically structured data
    - Bank trading transactions, shopping records, government sector archives
  - Active generation
    - Semistructured or unstructured data
    - User-generated content, e.g., social networks
  - Automatic production
    - Location-aware, context-dependent, highly mobile data
    - Sensor-based Internet-enabled devices

20

Big data value chain

- Acquisition
  - Collection
    - Pull-based, e.g., web crawler
    - Push-based, e.g., video surveillance, click stream
  - Transmission
    - Transfer to data center over high capacity links
  - Preprocessing
    - Integration, cleaning, redundancy elimination

21

Big data value chain

- Storage
  - Storage infrastructure
    - Storage technology, e.g., HDD, SSD
    - Networking architecture, e.g., DAS, NAS, SAN
  - Data management
    - File systems (HDFS), key-value stores (Memcached), column-oriented databases (Cassandra), document databases (MongoDB)
  - Programming models
    - Map reduce, stream processing, graph processing

22

Big data value chain

- Analysis
  - Objectives
    - Descriptive analytics, predictive analytics, prescriptive analytics
  - Methods
    - Statistical analysis, data mining, text mining, network and graph data mining
    - Clustering, classification and regression, association analysis
  - Diverse domains call for customized techniques

23

Big data challenges

- Technology and infrastructure
  - New architectures, programming paradigms and techniques are needed
- Data management and analysis
  - New emphasis on "data"
  - ➡ **Data science**

24

Large scale data processing

- Traditional approach
  - Database and data warehousing systems
  - Well-defined structure
  - Small enough data
- Big data
  - Data sets not suitable for databases
    - E.g., Internet data crawled by Google, Yahoo!, Facebook, …
  - May need near real-time (streaming) analysis
    - Different from data warehousing
  - Different programming paradigm

25

Large scale data processing

- Traditional computation is **processor bound**
  - Small dataset
  - Complex processing
- How to increase performance?
  - New and faster processor
  - More RAM

26

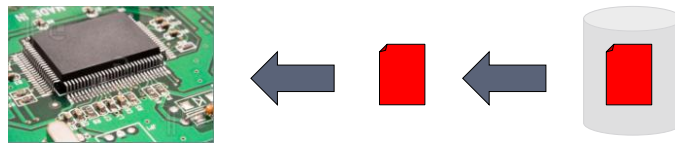
Large scale data processing

- Traditional data storage
  - On large SANs
  - Data transferred to processing nodes on demand at computing time
- Traditional distributed computing
  - Multiple machines, single job
  - Complex systems
    - E.g., MPI
  - Programmers need to manage data transfer synchronization, system failure, dependencies

27

The bottleneck

- Processors process data
- Hard drives store data
- We need to transfer data from the disk to the processor



28

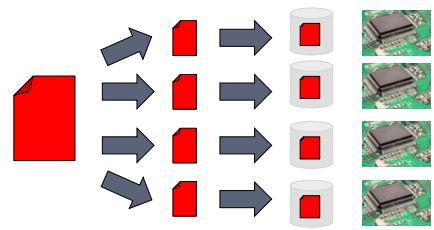
The bottleneck

- Hard drives evolution
  - Storage capacity increased fast in recent decades
    - E.g., from 1GB to 1TB
  - The transfer rate increased less
    - E.g., from 5MB/s to 100MB/s
- Transfer of disk content in memory
  - Few years ago: 3.33 min.
  - Now: 2.7 hours (if you have enough RAM)
- Problem: **data transfer from disk to processors**

29

The solution

- **Transfer the processing power to the data**
- Multiple distributed disks
  - Each one holding a portion of a large dataset
- Process in parallel different file portions from different disks



30

Issues

- Need to manage
  - Process synchronization
  - Hardware failures
  - Data loss
  - Joining data from different disks
  - Scalability
- Managed by new distributed architectures
  - E.g., Hadoop

31

Apache Hadoop

- Open source project by the Apache Foundation
- Based on 2 Google papers
  - Google File System (GFS), published in 2003
  - Map Reduce, published in 2004
- Reliable storage and processing system based on YARN (Yet Another Resource Negotiator)
  - Storage provided by HDFS
  - Different processing models
    - E.g., Map Reduce, Spark, Spark streaming, Hive, Giraph

32

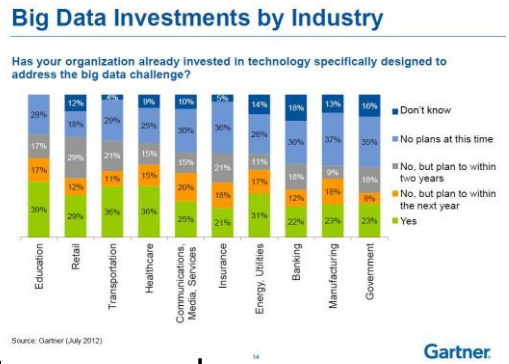
Hadoop scalable approach

- Data distibuted across nodes automatically
  - When loaded into the system
- Processing executed on local data
  - Whenever possible
- No need of data transfer to start the computation
- Data automatically replicated
  - For availability and reliability
- Developers only focus on the logic of the problem to solve

33

Conclusions

- Certainly not just hype



- … but not a panacea!

34