Big Data: Hype or Hallelujah?

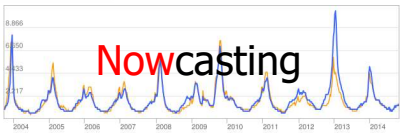
Elena Baralis
Politecnico di Torino

Big data hype?

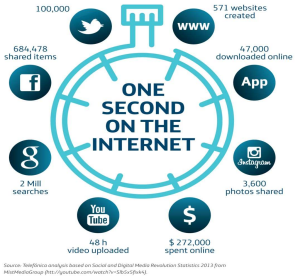
Google Flu trends

- February 2010
  - detected flu outbreak two weeks ahead of CDC data

Discontinued!



On the Internet…



- http://www.internetlivestats.com/

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"

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"

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"

What is big data?

- Many different definitions

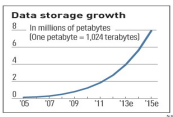
"Data whose scale, diversity and complexity require new architectures, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it"

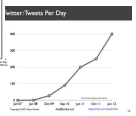
The Vs of big data

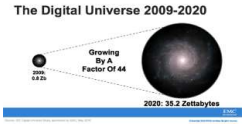
- The 3Vs of big data
  - Volume: scale of data
  - Variety: different forms of data
  - Velocity: analysis of streaming data
- … but also
  - Veracity: uncertainty of data
  - Value: exploit information provided by data

The Vs of big data

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







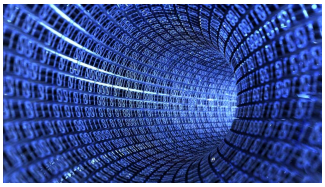
The Vs of big data

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

  - A single application may generate many different formats

The Vs of big data

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

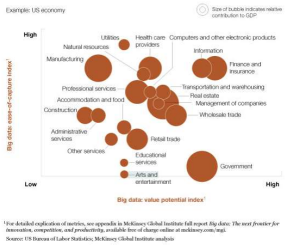


The Vs of big data

- Veracity
  - Data quality

The Vs of big data

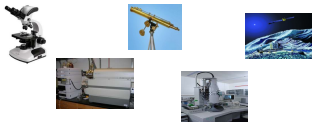
- Value
  - Translate data into business advantage



Who generates big data?

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

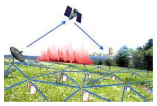
- Health and scientific computing



Who generates big data?

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

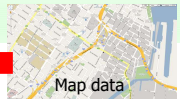
- Internet Of Things
  - Sensor networks, RFID, smart meters



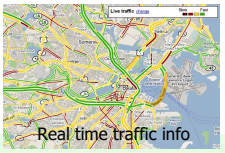
Big data at work

Crowdsourcing

Computing



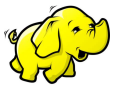
Sensing



Big data challenges

- Technology & infrastructure
  - New architectures, programing paradigms and techniques are needed

    Transfer the processing power to the data
  - Hadoop ecosystem

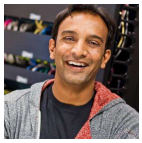


- Data management & analysis
  - New emphasys on "data"

    Data science

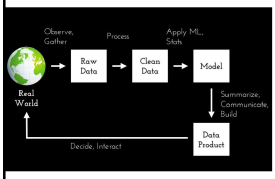
Data science

"Extracting meaning from very large quantities of data"



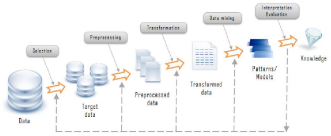
D.J. Patil coined the word *data scientist*

The data science process



AKA *KDD* process

Knowledge Discovery in Databases



Generation → Acquisition → Storage → Analysis

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

Big data value chain

Generation → Acquisition → Storage → Analysis

- 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

Big data value chain

Generation → Acquisition → Storage → Analysis

- 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

Big data value chain

Generation → Acquisition → Storage → Analysis

- 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

Large scale data processing

- Traditional approach
  - Database and data warehousing systems
  - Well-defined structure
  - Small enough data
- Big data
  - Datasets not suitable for databases
    - E.g. google crawls
  - May need near real-time (streaming) analysis
    - Different from data warehousing
  - Different programming paradigm

Large scale data processing

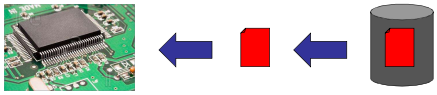
- Traditional computation is *processor bound*
  - Small dataset
  - Complex processing
- How to increase performance?
  - New and faster processor
  - More RAM

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

The bottleneck

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

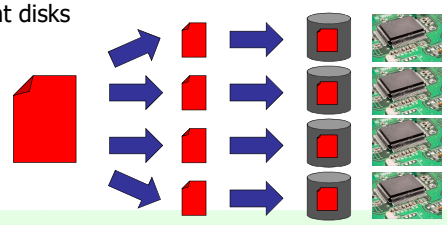


The bottleneck

- 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*

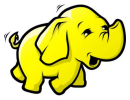
The solution

- *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



Issues

- Need to manage
  - Hardware failures
  - Network data transfer
  - Data loss
  - Consistency of results
  - Joining data from different disks
  - Scalability
- Managed by Hadoop



Apache Hadoop

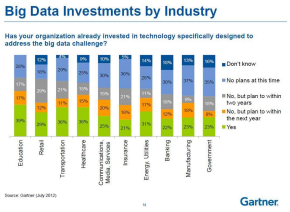
- Open source project by the Apache Foundation
- Based on 2 Google papers
  - Google File System (GFS), published in 2003
  - Map Reduce, published in 2004
- Reliable storage and processing system based on YARN (Yet Another Resource Negotiator)
  - Storage provided by HDFS
  - Different processing models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

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

Conclusions

- Certainly not just hype



- ... but not a panacea!