Introduction to Data Analysis

Amin Davoudi
Seminar in Big Data
University of Helsinki

More and more data available

The amount of data on the web is measured in Exabyte (1018) and zettabytes (1021)

Communications

Digital sensors

IoT -devices

Logs

Using of Big Data

- Pricing
- Out of home advertising
- Retail Habits
- Politics
- Weather
- Heart Disease
- Infectious diseases
- Doctor performance
- Optimizing Business Processes
- Performance Optimization
- Optimizing Machine and Device Performance
- Financial Trading

Defining Big Data

Several definitions

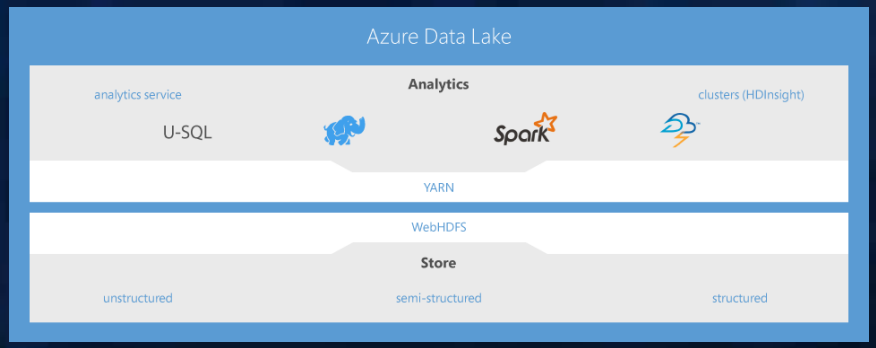
- Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.
- Big data is data too big to be handled and analyzed by traditional database protocols such as SQL.
- The data is too big, moves too fast, or doesn't fit the strictures of your database architectures
- The size is not the only feature of Big Data:
- Many authors explicitly use the Three V's (**Volume, Variety** and **Velocity**) to characterize Big Data:
  - **Volume** (Data in rest): The benefit gained from the ability to process large amounts of information is the main attraction of big data analytics.
  - **Variety** (Data in many forms): These data do not have a fixed structure and rarely present themselves in a perfectly ordered form and ready for processing (e.g. traditional relational database vs. NoSQL –database).
  - **Velocity** (Data in motion): Velocity involves streams of data, structured records creation, and availability for access and delivery.

Defining Big Data

The mechanism for storage and retrieval of data depends on the nature of data:

- Highly structured, Semi-structured and unstructured form on data.
- Social networks data, health care data, financial data, biochemistry and genetic data, astronomical data.
- Web logs, social media feeds, raw feed directly from a sensor source, email and etc.
- Videos, still images, audio, clicks.
- A **data lake** is a collection of storage instances of various data assets additional to the originating data sources.

- Gartner



Big Data management

- Data processing is seen as the gathering, processing, management of data for producing "**new**" information for end users.
- Key challenges are related to **storage**, **transportation** and **processing** of high throughput data.

Big Data analysis is splits into four steps (**4 A's**):

- **Acquisition:** Data from a variety of sources (web, DBMS(OLTP), NoSQL, HDFS) and dealing with diverse access protocols.
- **Organization:** Deal with various data formats (texts formats, compressed files, variously delimited, etc.) and extract the actual information like named entities, relation between them, etc.
- **Analyze:** Queries, modeling, and building algorithms to find new insights.
- **Decision:** Being able to take valuable decisions. It is very important for the user to "**understand and verify**" outputs.

Privacy in Big Data

- Privacy can cause problems at the creation of data for the analysis by no revealing the information needed.
- Privacy can also cause inconsistencies at the purging of database.
  - If we delete all individuals data we can get incoherence with data.
- General Data Protection Regulation (GDPR):
  - Take a stand on: **Data portability, Lead supervisory authorities, Data protection officers, Consent, Transparency, Profiling, High risk processing, Certification, Administrative fines, Breach notification, Data transfers, Contracts and liability, Consent**.
- Access, parse, normalize, standardize, integrate, cleanse, extract, match, classify, mask, and deliver data represents 80% of a Big Data project.

Big Data Technologies

- Various tools which can be used in Big Data management from data acquisition to data analysis.
- Most of tools are parts of Apache projects and are constructed around the Hadoop.
  - The ability to cheaply process large amounts of data, regardless of its structure.
- Hadoop = Hadoop Distributed File System (HDFS) + MapReduce.
  - HDFS is a distributed file system based on Google File System (GFS)
- In HDFS applications, files are written once and accessed many times; consequently data coherency is ensured and data are accessed in high performance.
- HDFS file system metadata are stored in a dedicated server, the NameNode, and the application data in other servers called DataNodes.
- Files are divided into blocks. Each block is replicated on a number of datanodes; all the datanodes containing a replica of a block are not located in the same rack.

Big Data Technologies

MapReduce: The name "MapReduce" expresses the fact that users specify an algorithm using two kernel functions: "Map" and "Reduce". It is a main programming model and associated implementation for processing and generating large datasets.

- Suitable for semi structured or unstructured data.
- The MapReduce's output is a set of <key, value> pairs.
- The Map function is applied on the input data and produces a list of intermediate <key, value> pairs.
- Reduce function merges all intermediate values associated with the same intermediate key.
- In a Hadoop cluster, a job (MapReduce program) is executed by breaking it down into pieces called tasks. When a node in Hadoop cluster receives a job, it is able to divide it, and run it in parallel over other nodes.

Big Data Technologies

MapReduce: Think as an Anonymous function (lambda abstraction) .

Function definition that is not bound to an identifier.

Map: Square (1,2,3,4,5,6) = 2,4,9,16,25,36

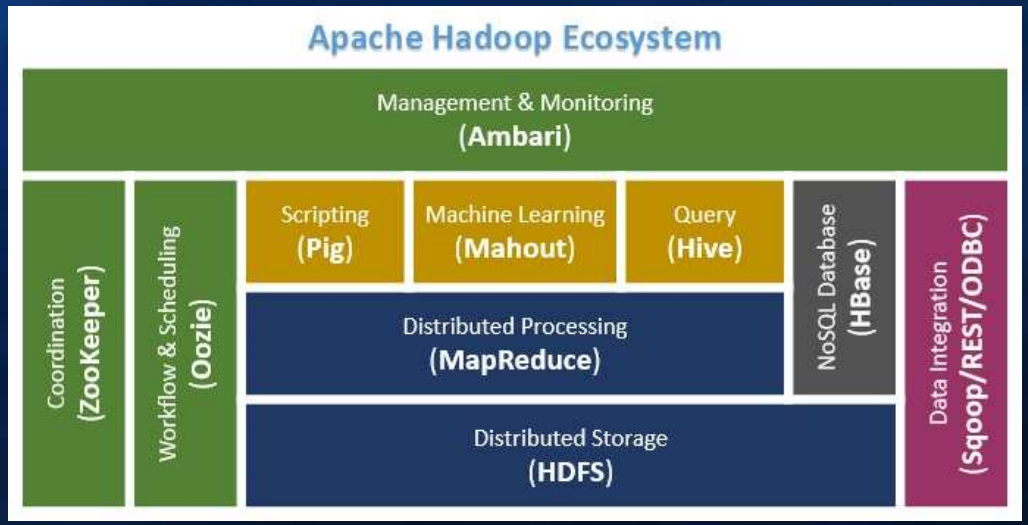
+

Reduce: Sum(2,4,9,16,25,36) = 98

Other projects

Around HDFS and MapReduce there are tens of projects. Those projects can be classified according to their capabilities:

- Storage and Management Capability.
- Database Capability.
- Processing Capability.
- Data Integration Capability.



Big Data visualization techniques

The ultimate goal of Big Data analysis and the achievement of this goal requires good visualization of Big Data content.

- Techniques and technologies used for creating images, diagrams, or animations to communicate, understand, and improve the results of big data analyses:
  - Tag Cloud
  - Clustergram
  - History Flow
  - Spatial information

Data analytics

Data analytics (DA) is an advanced analytic technique on big data in order to draw conclusions about the information they contain.

Due to the characteristics of big data, mainly variety, there are many techniques used for analytics on big data:

- Association rule learning
- Machine learning
- Data mining
- Cluster analysis
- Crowdsourcing
- Text analytics

Some of techniques raise criticism related to the memory consumption, scalability and reliability.

Analysis models

- Descriptive analysis
  - Data mining, business performance
- Diagnostic analysis
  - What happened and why
- Predictive analysis
  - What is (likely) going to happen
- Prescriptive analysis
  - Develop and analyze alternatives
- Decision-supporting analysis
  - Information visualization

DATA -> INFORMATION -> KNOWLEDGE -> INTELLIGENCE

Machine learning

Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

- sas.com

Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). The program is provided feedback in terms of rewards and punishments as it navigates its problem space.

- Wikipedia

Adding Big Data capability to an existing information system

- Data acquisition: Since traditional databases have to deal with structured data, existing ecosystem needs to be extended across all of the data types and domains.
- Data integration capability needs to deal with velocity and frequency. The challenge here is also about ever growing.

There is not a commonly agreed solution!

Big data quality, the next semantic challenge

Companies and governments are interested in two types of data in a big data context.

- First, they consider data generated by human, mainly those disseminated through web tools (social networks, cookies, emails...).
- Secondly they want to merge data generated from connected objects.
- This perspective raises new questions about the quality of the data.

Big data quality, the next semantic challenge

- Big Data is **big** and **messy**, challenges can be classified into engineering tasks (managing data at an large scale) and semantics (*finding and meaningfully combining information that is relevant to your needs*) have identified each a relevant challenge for Big Data:
- The meaningful data integration challenge which can be seen as a five-step challenge:
  - Define the problem to solve.
  - Identify relevant pieces of data in Big Data.
  - Organize it into appropriate formats and store it for processing.

Identifying relevant pieces of information in messy data

- Cut out irrelevant data.
  - mostly done by a "bag of words"
- Select relevant one according to a threshold.
- Built indices on "schema paths" (concepts whose instances have to be joined to answer a given query) to identify the sources which may contain the information needed.

Ethics and privacy

- More pieces of valuable information can be identified or inferred than it was possible before.
- Feasible analysis in real time and thus a continuous refining of users' profiles.
- Make users traceable.
- Allows data owners to build more complex and rich profiles of users.
- Variety leads to a diversification of business plans making big data more attractive at a bigger level.
- The pivotal point is about the balance between **benefits** and **drawbacks** of snooping around people's big data.

Mayer-Schönberger and Cukier four principles:

- Privacy should be seen as a set of rules encompassing flows of information in ethical ways but not the ability to keep data secret.
- Shared information can still be confidential.
- Big data mining requires transparency.
- Big data can threaten privacy.

Conclusion

- Define and characterize the concept of Big Data.
- Semantics (reasoning, coreference resolution, entity linking, information extraction, consolidation, paraphrase resolution, ontology alignment) with a zoom on "V's".
- Volume is the most tackled aspect and many works leverage Hadoop MapReduce to deal with volume.
- Unlike velocity, web and social media informality and uncertainty are addressed by scientists.
- Uncertainty can be handled manually or automatically (identification and/or isolation of inconsistencies).
- About velocity, knowledge bases must be continually updated and data processed periodically.
- In variety, we must deal with various data formats (tweets in and natural language texts and distributed data.
- Big Data must be addressed jointly and on each axis to make significant improvement in its management.

Conclusion

Efficient distribute processing system is important!

- **In order to scale to big data, you need a have some kind of processing distributed system built on top**
  - Hadoop, Spark, Storm, Others …
- Choosing the right tools and algorithms
  - Random forest, SVM, Neural networks, Deep belief networks etc.
- Choose the right features (more important than the models that you are using)
- Understand what the you are doing!

Real-time Demo and Links for Open Data

- Data.gov (Search through 194,832 USA data sets about topics ranging from education to Agriculture.)
- US Census Bureau latest population, behavior and economic data in the USA.
- European Union Open Data Portal thousands of datasets about a broad range of topics in the European Union.
- Google Public data explorer search through already mentioned and lesser known open data repositories.