DMIF, University of Udine

Data Management for Big Data

Big Data and Business Analytics

Andrea Brunello
andrea.brunello@uniud.it

May 2021

Outline

What is Big Data

Big data are beyond the usual limits of traditional databases, and are characterized by one or more of the properties:

- huge *Volume*
- high *Variety*
- acquired at high *Velocity*

Gartner analyst Doug Laney introduced the 3Vs concept in a 2001 MetaGroup research publication: "*3D data management: Controlling data volume, variety and velocity*"

Volume

Volume means basically the size of stored data, which may derive from human actions or can be machine-generated.

Sometimes the volume of data is so massive that they cannot be stored in their entirety, but have to be compressed/transformed online, as soon as they arrive (e.g., scientific sensor data).

Sometimes data can be stored using traditional RDBMS, while other times this choice may end up being too expensive in terms of cost or time $\rightsquigarrow$ NoSQL solutions, Hadoop.

Volume (Orders Of Magnitude)

IBM 350 disk storage (1956, 3.75 MB) → megabyte

Walmart's DW (1992, 1 TB) → terabyte

1 year of CERN's LHC data (15 PB) → petabyte

Quantities of bytes						
Common prefix				Binary prefix		
Name	Symbol	Decimal SI	Binary JEDEC	Name	Symbol	Binary IEC
kilobyte	KB/kB	10^3	2^{10}	kibibyte	KiB	2^{10}
megabyte	MB	10^6	2^{20}	mebibyte	MiB	2^{20}
gigabyte	GB	10^9	2^{30}	gibibyte	GiB	2^{30}
terabyte	TB	10^{12}	2^{40}	tebibyte	TiB	2^{40}
petabyte	PB	10^{15}	2^{50}	pebibyte	PiB	2^{50}
exabyte	EB	10^{18}	2^{60}	exbibyte	EiB	2^{60}
zettabyte	ZB	10^{21}	2^{70}	zebibyte	ZiB	2^{70}
yottabyte	YB	10^{24}	2^{80}	yobibyte	YiB	2^{80}

Differences between formats and the absence of a common structure are a typical characteristic of big data.

Data may come from different sources. Considering the web it may come from humans, like *user-generated content*, or it can be machine-generated, such as *logs, packet traces, etc.*

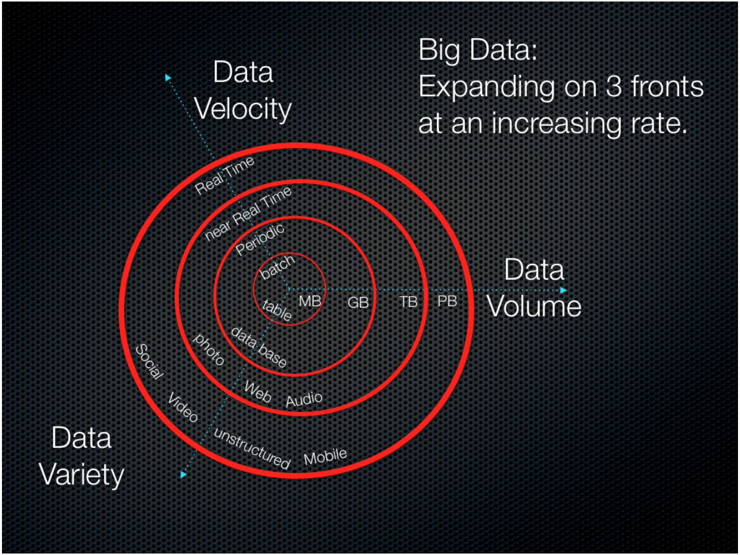
Heterogeneity of formats, structures and sources make it difficult to process and store such data using traditional tools.

Data acquired via sensors, or scientific instruments, may come at a high speed.

Some data have to be stored or analyzed as soon as they arrive, since they are transient (e.g., logs, data streams).

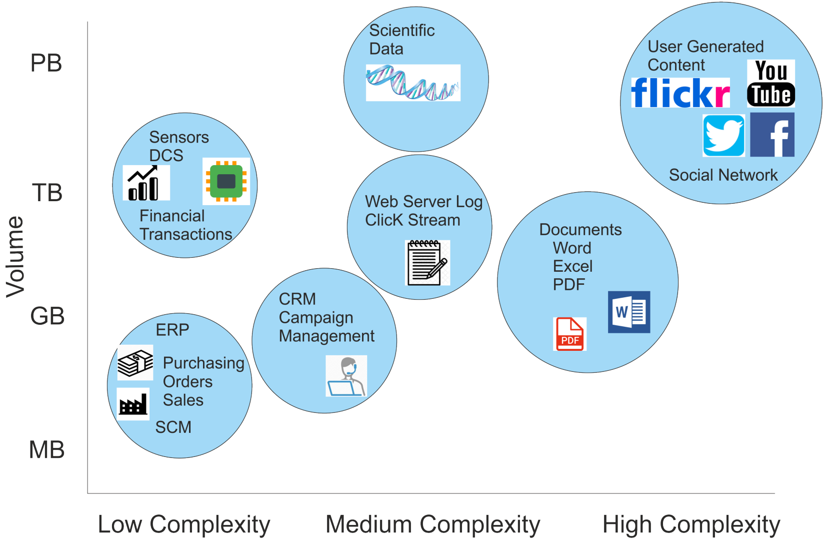
For companies that rely upon fast-generated data it is also important to even exploit/analyze such data as fast as possible.

"Just in its 1st phase, the SKA telescope will produce some 160 TB of raw data per second that the supercomputers will need to handle."
`https://www.skatelescope.org/frequently-asked-questions/`


Big Data:
Expanding on 3 fronts
at an increasing rate.
Data
Velocity
Real Time
near Real Time
Periodic
batch
Data
Volume
MB
GB
TB
PB
Data
Variety
table
data base
photo
Web
Audio
Social
Video
unstructured
Mobile

- **Volume**: scale of the data
- **Variety**: different forms of data
- **Velocity**: e.g., analysis of streaming data
- **Variability**: changes in the characteristics of the data
- **Value**: revenues, hypotheses that may arise from the data
- **Veracity**: trustworthiness, origin and reputation

Classification by Volume and Complexity



Big Data Use Cases

Radio Frequency Identification (RFID) is a technology for the automatic identification of objects, animals or people.

A *tag* uniquely identifies an object and it can be read remotely via radio frequency.

Several companies make use of this technology to control their processes (e.g., Walmart).

Typical applications include: automated inventory, object tracking, logistics, passports, anti-theft systems.

Considering that there are billions of tags all over the world, they are a good example of big data due to the huge volume of information they generate.

Data from the Web plays a big role in big data realm, and are characterized by volume, variety and velocity. Web data is about:

- HTML pages (in any language)
- Tweets
- Social network content (Facebook, LinkedIn, etc.)
- Forum comments and blog posts
- Documents in several formats: XML, PDF, Word, Excel, etc.

Industry 4.0 is a “hot” topic nowadays. Its a process that has its final goal in a factory (almost) completely automated and interconnected.

As an example, considering data generated by sensors:

- they can be used for real time monitoring, for instance to allow *predictive maintenance* to be performed
- tools of *stream analytics* are needed to deal with information flowing constantly and at a high rate (e.g., *Apache Flume*)

The Internet Of Things (IoT) is about (daily life) objects that are equipped with sensors and connectivity, acting as sources of data.

- this concept may involve both industry 4.0 and consumer products (like connected automobiles, kitchen appliances, etc.)
- data can be used for tasks such as surveillance, predictive maintenance, or performance enhancement in general
- if objects are attached to people even human behaviour and well-being can be analyzed

Big Data Challenges, Issues, and Opportunities

Big data come with challenges and opportunities:

- *business:* big data give companies the opportunity to develop new business models or to get advantages with respect to traditional business
- *technology:* size and complexity of big data require adequate solutions
- *financial aspects:* several use cases show that exploiting big data may lead to economic benefits. To this end, it is also important evaluate the costs involved with their management (e.g., cloud solutions)

Quality of big data is about a set of characteristics:

- *Completeness:* all data needed to describe an entity, a transaction, an event are present (e.g., missing fields for a contact entry)
- *Consistency:* absence of conflicting information inside the data (also considering *business rules*)
- *Accuracy:* the data conforms to the real values
- *Absence of duplication:* no redundancy of fields, records, or tables in the same or in different systems
- *Integrity*: with respect to RDBMS constraints: *data types, primary keys, foreign keys, check constraint*

Data may suffer from different kinds of error:

- errors due to manual data entry
- errors due to ill-designed databases
- errors due to the data handling software (e.g., issues within the ETL process)

The *data quality process* aims at determining which data offers an acceptable level of quality and which do not

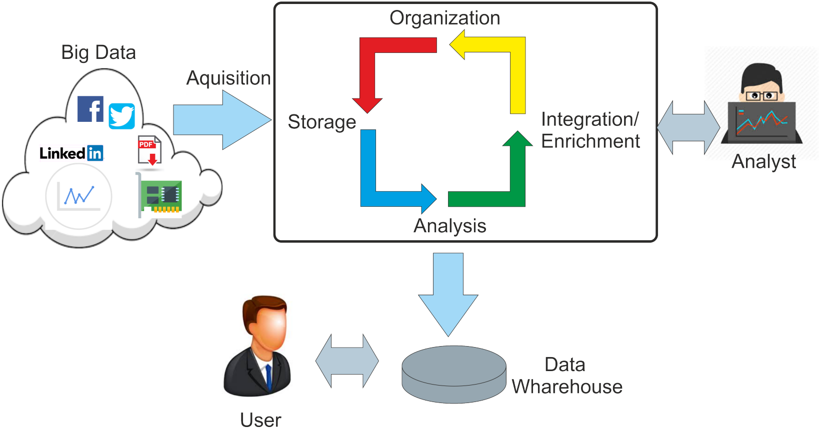
If the analyses, or the predictions, are based on low quality data, the results will probably be wrong or inaccurate (*garbage in = garbage out*)

Privacy and, especially, property are directly linked to the usage possibilities of the data:

- The Web, with tons of *user-generated content*, is a mine of personal behaviours, preferences and even thoughts. From social networks political, sexual or religious opinions can be extracted
- Confidential data, such as health issues, raise concerns about security: are they safe enough from possible hacks?
- It's impossible not to leave electronic traces of your *movement* via: phone calls, credit cards, GPS devices, geo-tagged photos

Big Data Technologies: Hadoop

Big data life cycle



What is Hadoop ?

Hadoop is an open-source platform designed to support distributed computation in a reliable and scalable way.

Hadoop was developed by Doug Cutting and Mike Cafarella in 2005 to address a scalability problem of an open-source crawler (Nutch).

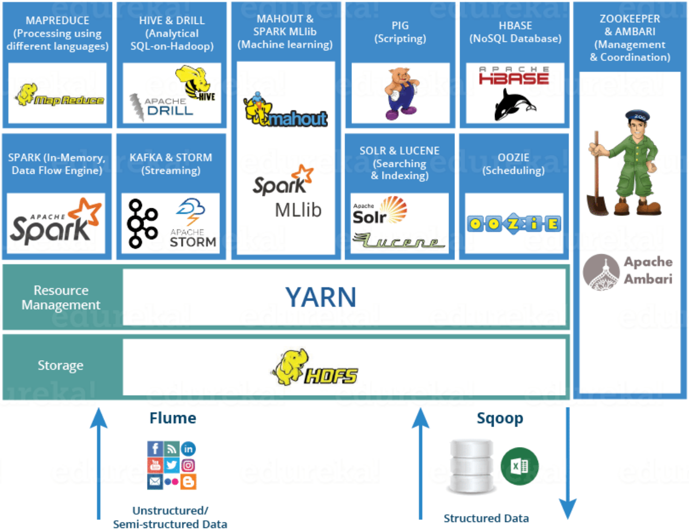
The first release in 2008 was an independent project of Apache. Nowadays, it is a collection of projects belonging to the same infrastructure for distributed computing.

Its main strength is the capability to use (cheap) commodity hardware to handle scalability.

Hadoop core components

- *Hadoop common:* software layer that acts as a support for the other modules, providing libraries and utilities.
- *HDFS:* distributed file system that stores data on commodity machines. It provides an effective way to access the data, guaranteeing redundancy to deal with failures. Any file format is supported, structured or not.
- *YARN:* (Yet Another Resource Negotiator) platform responsible for managing computing resources in clusters and using them to schedule users' applications.
- *MapReduce:* a parallel processing system for managing huge amounts of data, following the *divide et impera* strategy.

Hadoop Extended Ecosystem



Apache Sqoop

The name stands for SQL to Hadoop.

It is a straightforward command line tool.

Designed to transfer efficiently bulk data between Apache Hadoop and relational databases.

It supports the incremental reading of a relational table and the writing to HDFS, Hive or HBase.

They all allow to handle streaming data.

Flume is specifically designed to move unstructured or semi-structured data to Hadoop (HDFS, Hive, HBase), particularly log data.

Kafka is a more general-purpose tool. It adopts a distributed messanging system where publishers write data to topics and subscribers read from topics, providing a unified, low-latency, high-throughput platform for handling real-time data feeds. It should be used when the data destination is not (just) Hadoop.

Storm is a distributed real-time computation system not just for streaming, but it also includes other features such as real-time analytics, continuous computation, …

HBase is a non-relational distributed database modeled after Google's Bigtable and written in Java.

Hive is a data warehousing solution. It provides HiveQL, a language similar to SQL, that allows to run queries with MapReduce support in a transparent way.

Drill is a schema-free SQL query engine. It allows one to perform SQL queries against several NoSQL databases, and local files.

Mahout and Spark MLlib

They both provide scalable and distributed implementations of machine learning algorithms.

Mahout includes algorithms that support many tasks, such as classification, clustering, dimensionality reduction, and topic extraction. Originally based on MapReduce, today it is primarily focused on Spark.

Spark MLlib is a scalable machine learning library based on Spark. It includes all the most popular machine learning algorithms, such as random forests, gradient boosting trees, K-means, LDA, ...

Oozie, Solr and Lucene

Oozie is a server-based workflow scheduling system to manage Hadoop jobs. It combines multiple jobs sequentially into one logical unit of work.

Solr and *Lucene* provide a search engine software library. Major features include full-text search, real-time indexing, dynamic clustering, database integration, NoSQL features and rich document (e.g., Word, PDF) handling.

Spark is an open-source distributed general-purpose cluster-computing framework, like MapReduce.

Differently from MapReduce, which has to read from and write to a disk while performing processing tasks, Spark can do it in-memory.

As a result, developers claim that Spark is capable of running programs up to 100x faster than MapReduce, making it suitable for real-time computation.

Nevertheless, Spark requires a lot of memory to load the processes. On the contrary, leveraging the disk, MapReduce is able to work with far larger datasets than Spark.

Ambari and Zookeeper

Ambari and Zookeeper provide support to the Hadoop administrators.

Ambari allows the provisioning, management and monitoring of Hadoop clusters.

Zookeeper is essentially a service for distributed systems offering a hierarchical key-value store, which is used to provide a distributed configuration service, synchronization service, and naming registry for large distributed systems.

Leveraging Big Data: Business Intelligence and Analytics

Collecting data is important but, without analysis, there is no value from them.

Creating value from data is also referred to as *data monetization*.

Data analyses may be performed by means of suitably designed *Business Intelligence* and *Business Analytics* systems.

There are three main analysis types to extract value from data: *descriptive analytics*, *predictive analytics* and *prescriptive analytics*.

Not only analyses… sometimes data monetization pertains just selling data (e.g., by social networks, companies).

Business Intelligence (BI) can be defined as a set of tools and techniques for the transformation of raw data into meaningful and useful pieces of information for business analysis purposes.

BI entails the management of large amounts of data to help in identifying, improving, and possibly defining new strategic business opportunities.

In particular, it aims at providing *historical* and *current* views of business operations.

It is thus **descriptive**.

BI Related Decisions

Example, considering a customer churn analysis problem:

- How many customers left our company during last year?
- What are our most valuable customers?
- Does our pricing impact the churn rate?

Business analytics (BA) relies on data mining and machine learning to determine what the future will probably look like.

Also, by means of such tools, it may help to identify the main reasons behind specific phenomena (e.g., explanation models).

It is thus **predictive**.

Typical questions, churn example:

- How will a 10% increase of the price affect the churn rate?
- Which customers are more likely to leave our company?

Prescriptive analytics is a relatively recent development of Business Analytics.

It goes beyond predicting future outcomes, by also suggesting actions to benefit from the predictions and showing the implications of each decision option.

In the churn example:

- What are the variables that are most likely to affect the decision of the customer?

Several trends are currently driving the evolution of Business Intelligence and Analytics:

- competition between companies
- mobile BI and reporting
- self-service BI
- collaboration and information sharing
- cloud computing
- big data

References

A. Pavlo and M. Aslett, *What's Really New with NewSQL?*, SIGMOD Record, June 2016 (Vol. 45, No. 2)

A. Rezzani *Big data. Architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati*, Maggioli Ed. 2013

A. Rezzani *Big data Analytics*, Maggioli Ed. 2017