DMIF, University of Udine

Introduction To Big Data

Andrea Brunello
andrea.brunello@uniud.it

April, 2020

Introduction

From Data to Decision Support Systems

Beginning from the 60's, companies started to collect data about their business processes and to store them on magnetic storage. Simple descriptive analyses over aggregated data were performed.

From the 80's, with the advent of the relational model and SQL, data analyses became a dynamic and interesting field. More flexible analyses could be performed, at different levels of detail.

Complexity arises from dealing with different software, technologies and models, provided by several vendors.

On Line Transaction Processing

Analyses were performed over *On Line Transaction Processing* (OLTP) systems, which are highly normalized.

Normalization is good to support transactional operations (insert, delete, update).

Nevertheless, many JOIN operations are required in order to denormalize data and present it the desired tabular form: this may lead to poor performances with massive amounts of data.

In OLTP databases there is usually a lack of historical depth about data.

Data Warehousing

To overcome limitations of OLTP systems, in the 90's the era of *Data Warehousing* begins with the aim of integrating several sources of operational data and supporting analyses.

Data gathered from all the business process in a company are conveyed into a Data Warehouse in an integrated, consistent and certified way.

This step is crucial to unleash the power of *Business Intelligence* in companies of any size.

Business Intelligence

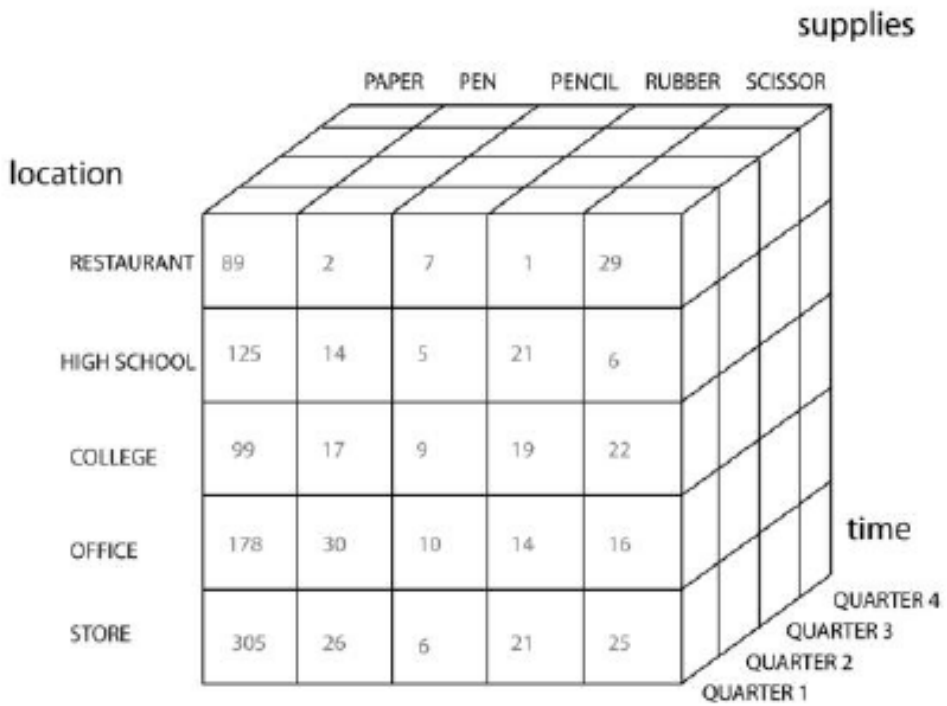
Business Intelligence (BI) can be defined as a set of tools and techniques for the transformation of raw data into meaningful and useful pieces of information for business analysis purposes.

BI makes use of computations, analyses and aggregations to transform, store and present information in an effective way to support strategical, tactical and operational decisions (descriptive analytics).

BI moved from analyses performed over data warehouses with SQL to multidimensional OLAP databases.

On Line Analytical Processing

On Line Analytical Processing (OLAP) databases rely on multidimensional data structures named cubes (or hypercubes, as there are usually more than 3 dimensions).



Data Mining

From the 2000's companies felt the need of techniques to analyze data in order to anticipate events and to obtain forecasts about both problems and opportunities (predictive analytics).

Such techniques and analyses are named *data mining* because of their ability to "mine" data to extract hidden, previously unknown, and highly valuable information.

Some Problems Related To Data Mining

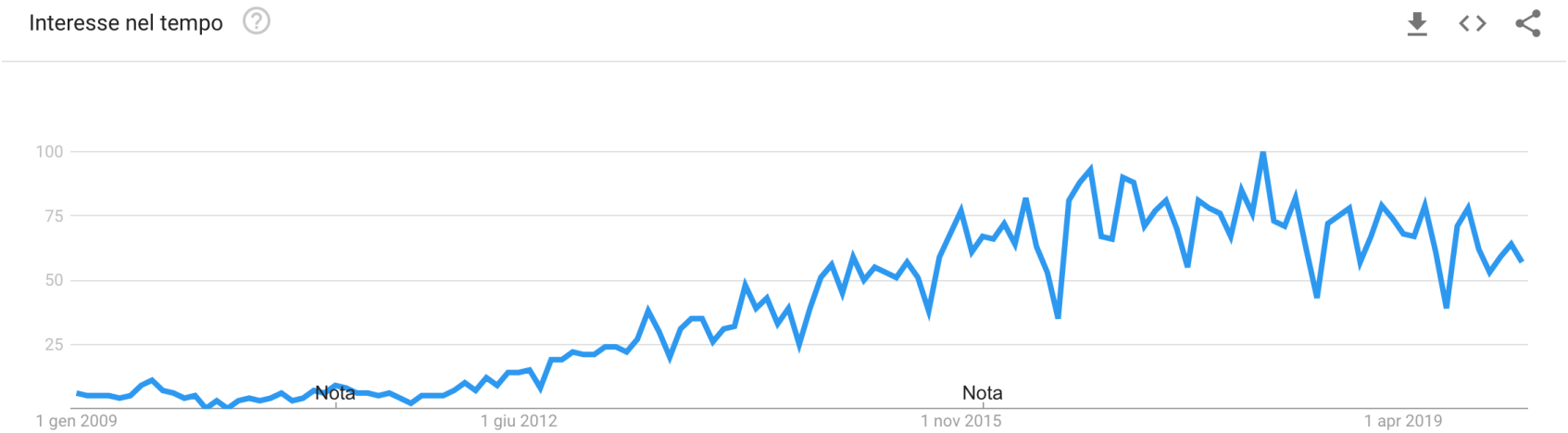
- *Anomaly detection.* Data mining can be employed to detect credit card fraud or equipment failure
- *Churn analysis.* Detect customers that are likely to move to the competitors
- *Forecasts.* Determine future trends about the sales or, in general, about time series
- *Targeted marketing.* Use specific marketing strategies to target the most easily influenced customers
- *Market basket analysis.* Suggest certain items to a specific customer on the basis of his past behaviour

Trends In Business Analytics

Starting from 2010 new trends are driving the evolution of BI and data analyses:

- growth of business analytics
- mobile BI and reporting
- self-service BI
- collaboration and information sharing
- cloud computing
- big data

Google Trend - Big Data



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 (Order Of Magnitude)

IBM 350 disk storage (1956, 3.75 MB)

Walmart's DW (1992, 1 TB)

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

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}

Variety

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.

Velocity

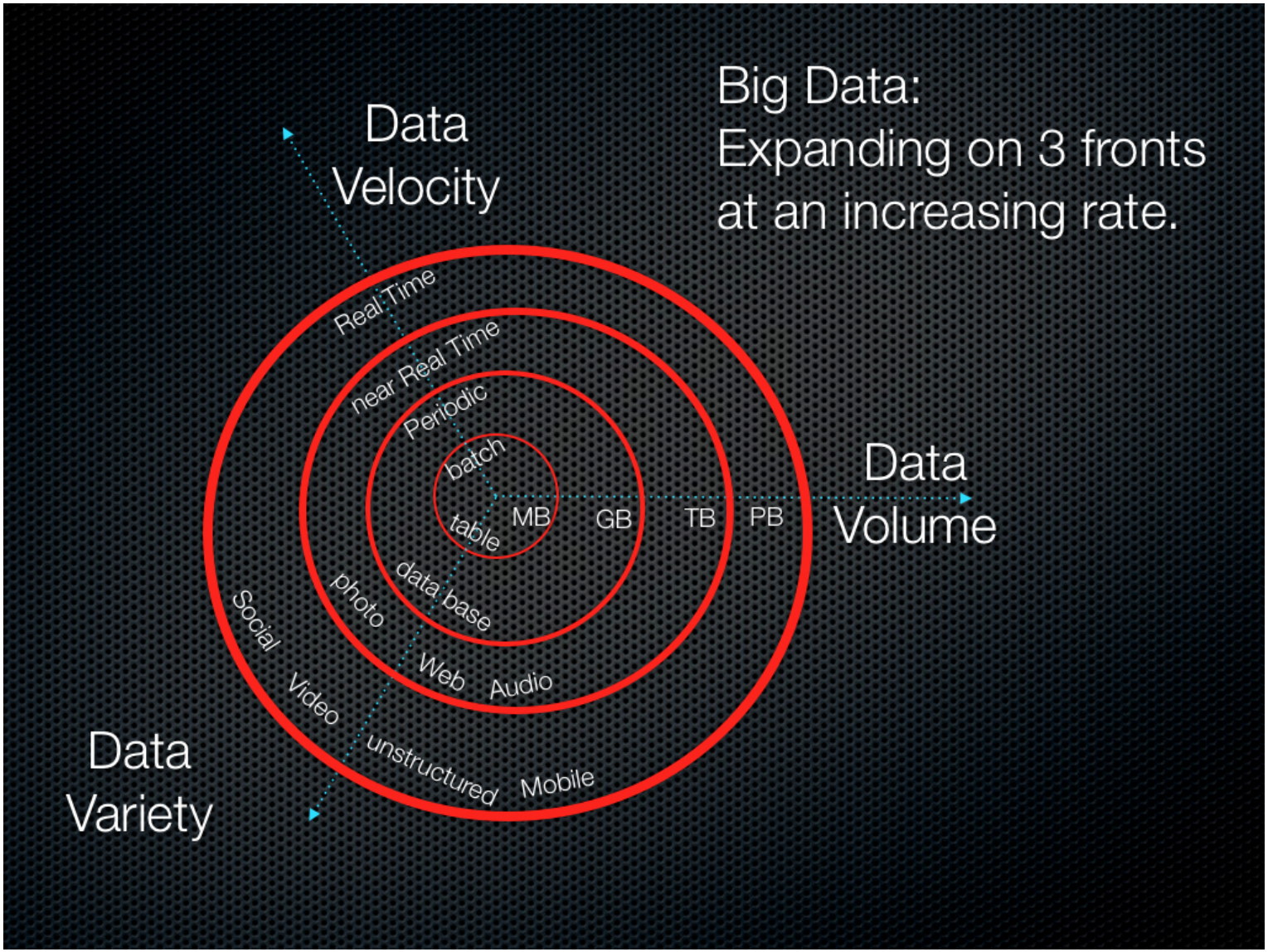
Data acquired via sensors, or scientific instruments, may come at a high speed.

Some data have to be stored 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/`

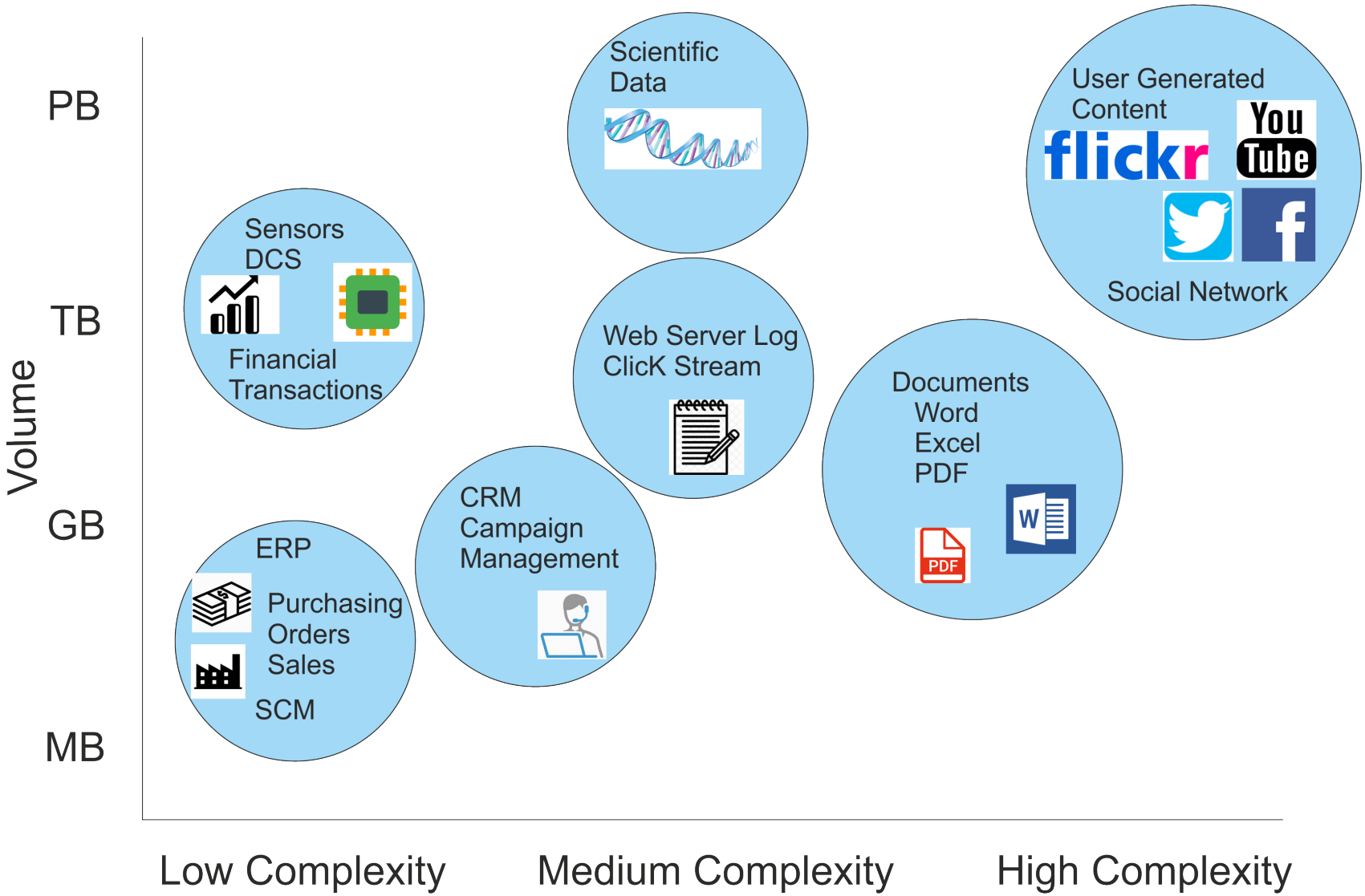
The 3 original Vs



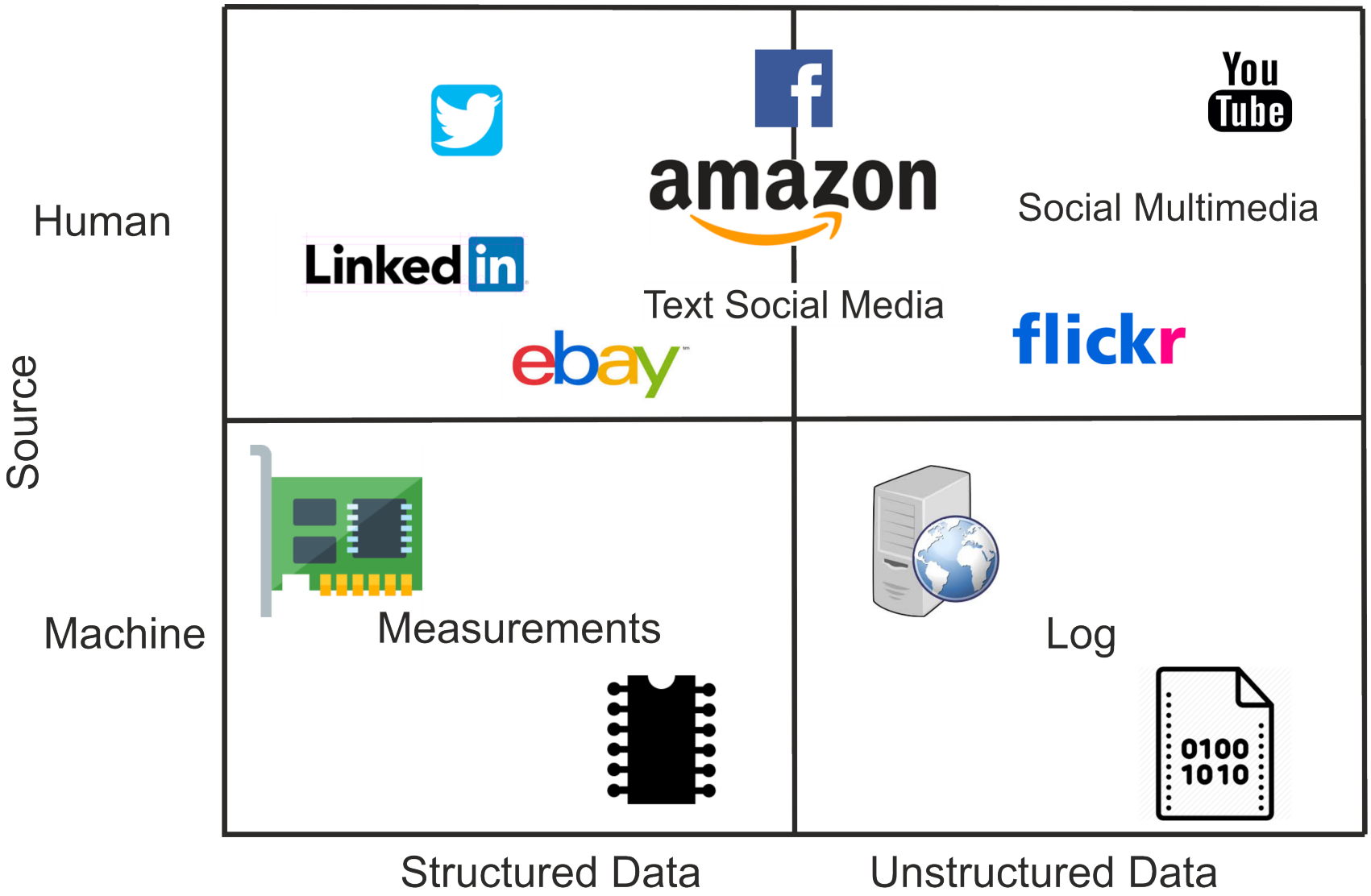
Big Data
Additional Vs

- **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

Big Data
Classification by volume and complexity



Big Data
Classification by source and level of structuring



Big data comes 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)

Big Data Use Cases

Big Data Use Cases
RFID

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.

Big Data Use Cases
Data from the Web

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.

Big Data Use Cases
GIS data

Data from *Geographical Information Systems* (GIS) are often used to enrich the big data analyses:

- geographical context may help to find patterns based on spatial distribution
- merging social network messages with their location may help to detect emergency situations and act accordingly
- public administrations can use GIS data to predict and prevent diseases, crimes, etc.
- private companies can get a deeper insight on their customers behaviour

Big Data Use Cases
Banking

In a medium sized bank there can easily be several millions of transactions per day. Working only on aggregated data (e.g., on monthly basis) can lead to a potential loss of useful information:

- size of the raw data is a limit to the application of predictive models in a reasonable time
- using big data technologies (Hadoop MapReduce, Spark) one may develop models regarding *churn analysis, customer clustering, or the set up of targeted marketing campaigns*

Big Data Use Cases
Industry 4.0

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*)

Big Data Use Cases
Internet Of Things

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

Extracting Value From Data

Data Monetization

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

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

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

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

Descriptive Analytics

It is used in almost every company.

It is focused on historical or current data, and makes use of data warehouses and OLAP engines.

Operations such as drill down, roll-up, slicing and dicing and pivoting are performed on data cubes.

Typical questions:

- how many cars did we sell in France last year?
- how big is our inventory at the moment?

Predictive Analytics

Predictive analytics aims at developing a vision of the future making use of past data.

It relies on statistical analysis and machine learning, in addition to proper data modellinq, preprocessing and querying.

Typical questions:

- how will our overall sales figures look like in the next semester?
- to which customers should we grant a loan (= which customers are likely to pay us back in the future)?

Prescriptive Analytics

Prescriptive analytics is an innovative concept which has its roots in predictive analytics, but it goes even further.

It can suggest to *decision makers* the actions to take in order to reach a desired goal.

Typical example:

- what are the factors the mostly influence the probability to sell a car?
- how will future sales be affected if I work on them?
- considering this, act on such factors accordingly, to reach the desired sales goal.

Big Data Issues

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*

Big Data Issues
Error sources

Data may suffer from different kinds of error:

- errors due to manual data entry
- error 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

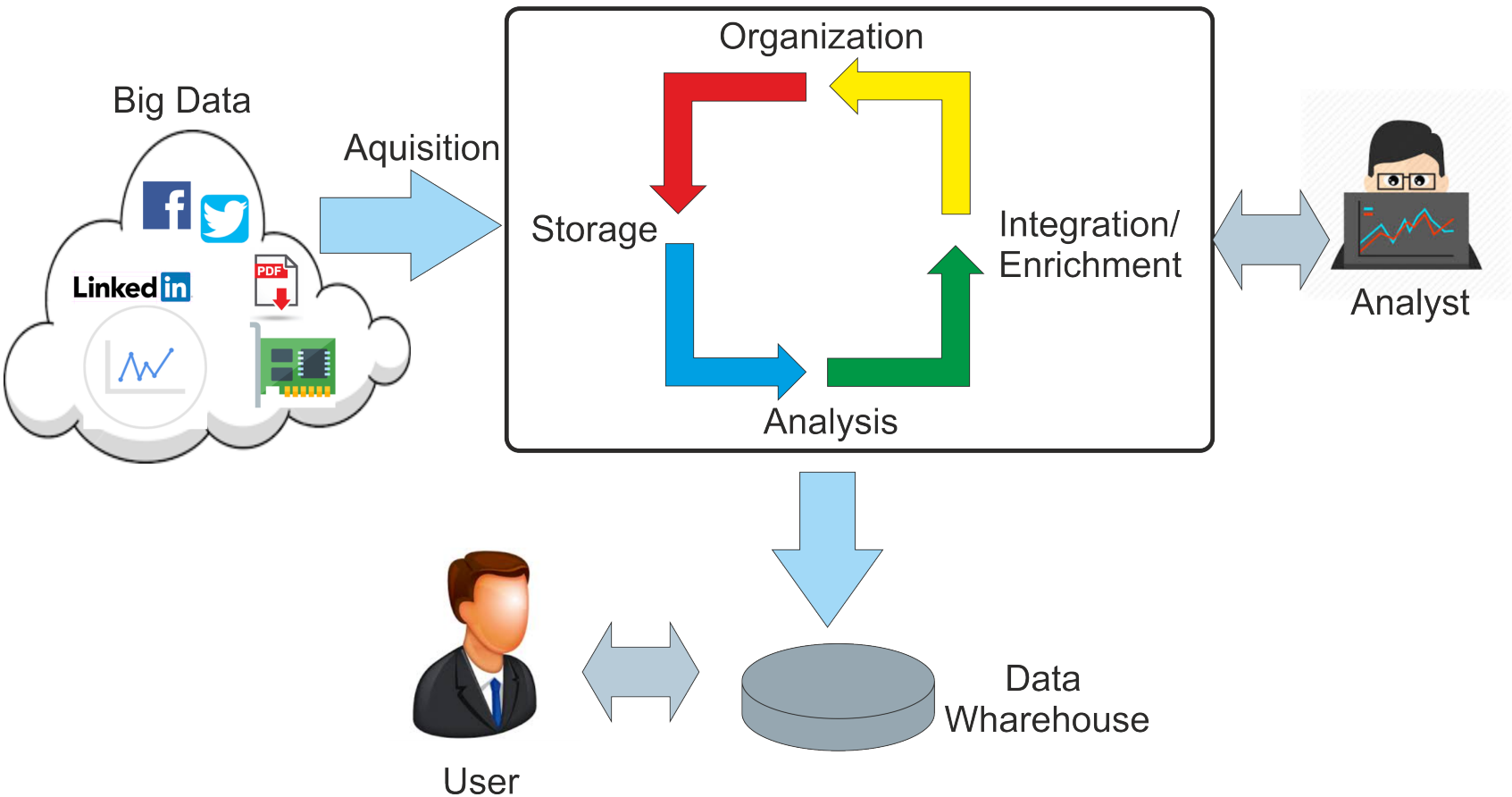
If the analyses, or the predictions, are based on low quality data, the results will probably be wrong or inaccurate (*garbage in = garbage out*)

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

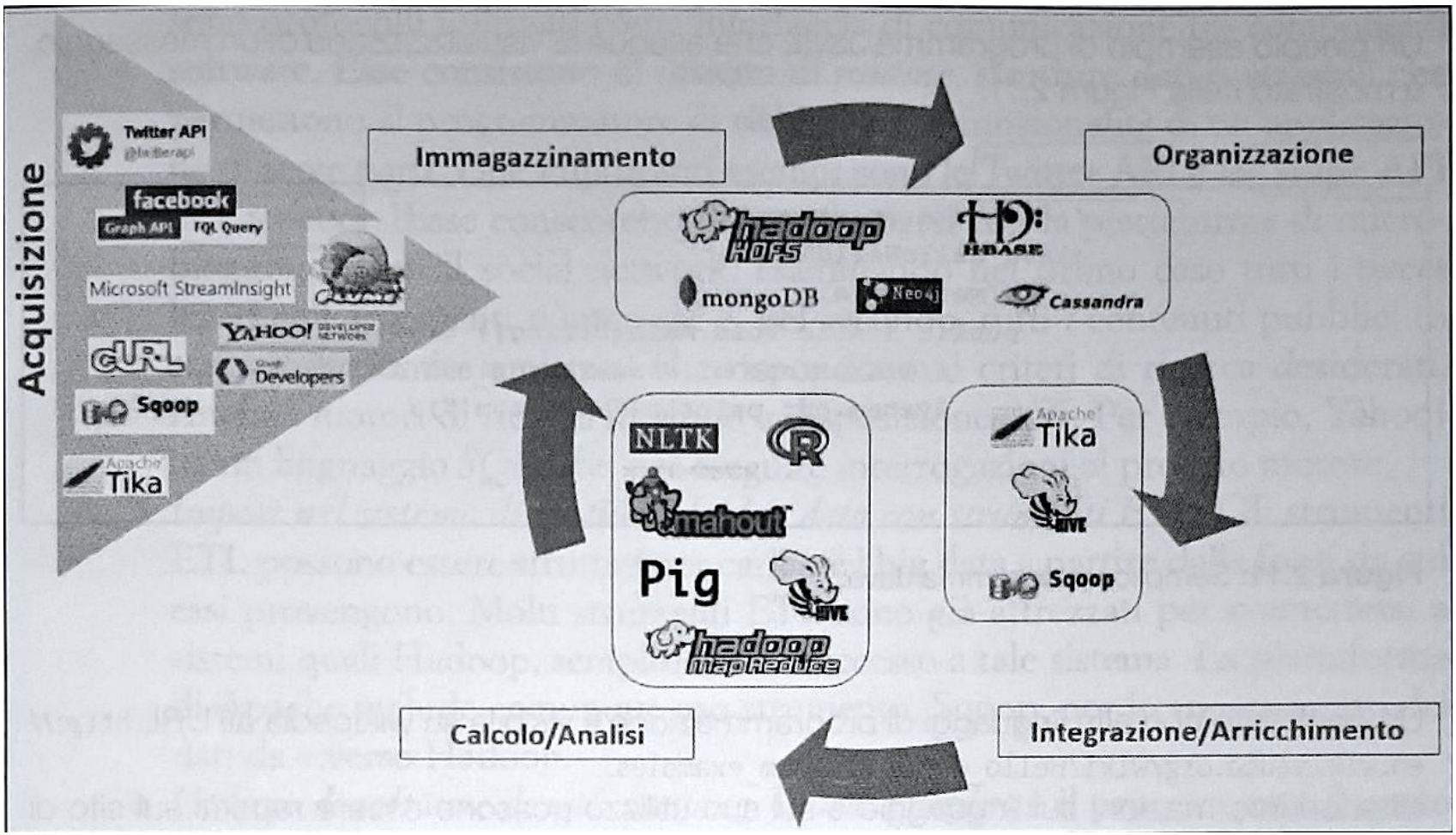
- The Web, with tons of *user-generated content*, is a mine of personal behaviours, preferences and even thoughts. From social networks political, sexual or religious opinions can be extracted
- Confidential data, such as health issues, raise concerns about security: are they safe enough from possible 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

Big data life cycle



Big Data Tools



Data acquisition may involve different means:

- *APIs available directly from data source (e.g., Twitter API)*
- *Import using ETL tools, possibly custom tools* (e.g., Hadoop Sqoop to transfer data to/from Hadoop)
- *Web scraping software* (e.g., Apache Tika to extract text from documents)
- *Data stream reading* (e.g., Apache Flume)

Storing a huge amount of unstructured, or semi-structured data is a big challenge. There are several technologies that are of common use:

- Distributed file system solutions like HDFS (*Hadoop Distributed File System*, together with an application ecosystem such as *HBase* and *MapReduce*)
- No SQL databases like *Cassandra*, *Berkeley DB*, *Mongo DB*, and *Neo4J* (graph database)
- New SQL approaches like *VoltDB*, *H-Store* and *MemSQL*

Once data has been acquired, stored and organized inside a suitable big data database, an additional integration or transformation step is likely required to prepare them for analysis:

- Integration with external data sources to and from Hadoop can be performed using *Squoop*
- In enterprise document repositories there is the need to deal with many formats like PDF, Word, Excel, HTML, XML; to this end, a tool like *Apache Tika* can manage different formats with a uniform approach
- A tool that can organize, manipulate, and analyze data is *Hive*: it can be defined as a data warehousing system based on Hadoop

The data analysis can be performed with many tools. Some of them rely upon distributed job functions made available by MapReduce and HDFS:

- *MapReduce* is a framework for processing parallelizable problems across large datasets using a large number of computers (nodes)
- *Pig* with its language Pig Latin simplifies the underlying use of MapReduce
- Hive supports *HiveQL*, a language similar to SQL that exploits MapReduce masking its complexity
- Statistical suite *R* has a package named RHadoop that exploits Hadoop and MapReduce

- *Apache Mahout* is a machine learning platform devoted to recommendation engines, clustering and classification
- *Apache SAMOA* is a platform for mining on big data streams. It is a distributed streaming machine learning (ML) framework that contains a programming abstraction for distributed streaming ML algorithms
- *Elasticsearch* provides a distributed, full-text search engine with an HTTP web interface and schema-free JSON documents
- *Apache Spark* is a general-purpose cluster-computing framework; built on MapReduce, with the objective of improving it under many aspects