An outlook on Big Data

A. Nicchi

Version 2.1 - last update on 06/12/2015

1 Introduction

The consequences of information society are that we have a swelling flood of data that is being generated, collected and stored about almost every aspect of the real world, from sciences, government (Open Data) and healthcare, to banking, economics and the Internet. So it was almost impossible not to pay attention to this phenomenon that someone called "datafication" [12].

But what do we mean for the term data? "Data", mass noun plural of "datum", means pieces of information representing some fact. But what is important is that data refers to a description of something that can be recorded, tabulated, quantified, analysed and reorganized in some form suitable to be processed by a computer system.

Where do these mountains of data come from? The explosion of these enormous quantities of electronic data production happened in the last two or three years, since the latest advances of Information Technology made it easier to generate data and now every day there is a rate of 3 billion of kilobytes of data produced by both humans and machines themselves. Let us show some numbers just to give an idea. More than three billion of search queries every day are processed and saved by Google. More than 10 million new photos are uploaded every hour on Facebook. On 2012 more the 400 million messages a day were tweeted on Twitter, that is growing at around 200 percent a year. About 2.5 million photos per day are archived on Flicker. Cloud computing and the Internet of Things (IoT) push to improve technology for the collection of data produced by machines, that also communicate to each other to help the lives of citizens.
For example, we have sensors in the cars which direct drivers towards available parking places, sensors for street lights for consuming less electrical energy, sensors for measuring the level of air pollution and so on, to form a

grid that produces many types of data.

Smart phones, watches and other wearables interact with this infrastructure producing other data. Moreover, many types of sensors, together with the extensive use of CCTV, are the source of many streams of data that are stored and processed in the cloud. They turn out to be a valuable source of useful information when they are analysed together. This is possible now because of the low cost of storages which have given the possibility to record and keep huge amounts of data and to manage and access them in almost a real time way.[10]

What is the nature of this data? [11] In general, data can be of three types: structured data, semi-structured and unstructured.

- **structured data**: data managed by classical Data Base Management Systems. The DBMS itself guarantees consistency and respect of the semantic schema of the data. When you make queries on a DBMS, you get clear and accurate results according to the data. There are contexts in which this is necessary, like in banks to manage bank accounts.

- **semi-structured data**: digital items that can be categorized and analysed. Data generated by sensors, by smart phones, by global positioning system (GPS) devices and log data are examples of these kind of data.

- **unstructured data**: data of various formats that needs more attention and more work to be processed. Emails, photographs, e-books, music, videos, comments on social networking sites and review from websites and so on are examples of unstructured data.

Nowadays, why do we pay such high attention to the data? Because it gives concrete opportunities to make revolutionary challenges, that is:

- **fast processing**: to extract new insights from humongous volumes of data in a fast way and close to real time.

- **predictions**: to apply maths to huge quantities of data in order to infer probabilities or mine useful information or knowledge on which to base predictions with an high degree of likelihood.

2 Big Data: definition and features

There isn't a generally recognized definition of Big Data, according to the context we have several descriptions and characterizations.
According to McKinsey & Company [5], Big Data shall mean the datasets that could not be acquired, stored, processed and managed by classic database software within a tolerable time. This definition underlines two aspects: firstly, data volumes are changing and may grow overtime because of technological advances; secondly, data is different in relation to the applications.
Doug Laney [8] defined Big Data with a 3Vs model: Volume, Velocity and Variety. In this model *Volume* means that data's order of magnitude gets bigger and bigger; *Velocity* means that data collection and analysis must be reasonably fast in order to utilize the commercial value of Big Data; *Variety* means that the data is of several types.
For IDC "*Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.*" [7] With this definition, Big Data characteristics may be summarized by 4Vs: Volume, Variety, Velocity and Value. This description underlines the necessity of exploring the huge hidden values and indicates the most critical problem in Big Data: how to discover in tolerable time values from humongous and increasing datasets containing data of various types.
NIST, focusing on the technological aspects, elaborated this definition: "*Big Data consists of extensive datasets primarily in the characteristics of volume, variety, velocity, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis.*" [4]

Wu et al. defined and stated the characteristics of Big Data using the HACE Theorem [14] [1] [13].

HACE Theorem: *Big Data starts with large-volume, **H**eterogeneous, **A**utonomous sources with distributed and decentralized control, and seeks to explore **C**omplex and **E**volving relationship among data.*

The HACE Theorem highlights the characteristics of Big Data:

1. **Huge and Heterogeneous Data and Various data sources**: the adjective "big" could mislead. We are almost always in presence of huge quantities (order of magnitude of petabytes) of data related to a phenomenon. But to be more precise, for Big Data we mean gathering

as much data as possible about a phenomenon or if it is feasible to collect N=all data about it. As sources and applications change, the related data has diverse dimensionalities, different representations and semantics.

2. **Autonomous Sources with Distributed and Decentralized Control**: each data source is able to produce and collect information without involving any centralized control as in the World Wide Web in which each web server provides information and functions without relying on other servers. As a consequence, the sources of data don't have any critical points of failure.

3. **Complex and Evolving Correlations**: statistical methods are applied to huge quantities of data to reveal hidden associations. This analysis doesn't rely on casualties but on correlations. In a dynamic world data increases and changes overtime and so does the complexity and the relationships related to the same data. So, while looking for linear relationships could be simple, more sophisticated analyses need to identify non-linear relationships among data. This is an important aspect of Big Data because predictions based on correlations constitute the revolutionary aspects of the Big Data age.

From the aforementioned definitions of Big Data we can infer and understand several important aspects related to this revolutionary phenomenon.
The advances of technology in the area of communication, collection, storage and processing has created the condition to accumulate actively and passively really huge amounts of data of several types and natures.
What is important about this data is that we can have almost "all data" about a phenomenon or in general about any aspect of the real world. All means all=N. When we use statistics we are not using a sample but all the universe. We don't need to infer information and then knowledge from a sample with the risk that the chosen sample is not a good representation of the entire population.[12]
Then, if we investigate this data in a smart way, it can produce new values:

- More knowledge
- Valuable and hidden correlations
- Predictions with an high level of probability

3 Correlations and predictions

In the Big Data age a phenomenon could be represented by almost all data related to it. However, data often carries a significant amount of imperfections and errors. In addition, for privacy preserving purposes data is often made anonymous by man-made perturbations. So, generally, these enormous quantities of data are deteriorated by messiness, errors, man-made perturbations and imperfections. This data should be examined to study a phenomenon. Using flexibility and *schema on read* or *schema on need*, we have to figure out how to work with data. The process of analysis of the data entails data loading, extracting and transforming in order to discover patterns, connections and correlations.

But what is, in general, a **"correlation"**? Correlation is a measure of the strength of a linear or non linear relationship between two variables A and B. The values of this measure range between +1 and -1, in particular:

- 0 indicates no linear relationship;
- +1 indicates a positives linear relationship: as A increases in its values also B increases in its values;
- -1 indicates a negative linear relationship: as A increases in its values then B decreases in its values.

As an example of the calculation of a linear correlation, we can use the correlation coefficient of Pearson.

In the context of Big Data, if A and B are often together in the data, then, there is a correlation between A e B. We can use B as a proxy to indicate that probably also A is taking place. The connection between A and B can be expressed by a function in case of strict dependence or can be non functional in the case where the same value for A can correspond to several values of B or vice-versa.

Here are some practical examples that show how correlations can predict the future with a certain degree of likelihood.

An example of correlations' application is the "predictive model" system elaborated by the British insurer AVIVA PLC to identify risky clients [9] . Instead of costly blood and urine tests to assess peoples's health, AVIVA uses

data from on-line shopping details, catalogue purchases, magazine subscriptions, leisure activities and from social networking web sites. Assuming that many diseases are related to lifestyle factors such as exercise habits and food diets, the model conceived by AVIVA can estimate a person's risk for illness such as high blood pressure and depression. So, by just using smart correlations between people and all the data about their lifestyle, it is possible to predict if a person will have a health problem. Moreover the new evaluation system costs less to the company and, as a consequence, the customer can benefit from it having services at lower prices.
Another example is related to the quick spread of the new virus flu H1N1 in 2009. In the Unites States the Centres for Disease and Prevention (CDC) were always two or three weeks out of date even though the doctors immediately informed them of new flu cases. Engineers at Google wanted to use the more than three billion saved search queries per day. They tried to look for correlations between the frequencies of certain search queries and the spread of the flu. Their software discovered a combination of 45 search terms that when used together had a strong correlation between their prediction and the official data provided by CDC. Using these correlations the Google mathematical model could tell where the flu had spread in almost real time.[6]

4 Graphical Analysis of Big Data

In this section we talk about an approach to analyse data using non-planar graphs [2] [3]. In the Figure 1 we can see a schema that shows the architecture of the framework and the methodology used to pass from the raw data format to graphical representations of linked data.

The process of analysis is characterized by three main steps: Data Collection, Data Transformation and Interlinking and Graph Analysis.

In the **Data Collection phase** we have the storage of huge quantities of structured data, semi-structured data and unstructured data originating from many independent sources. This data is heterogeneous and raw.
In the **Data Transformation and Interlinking phase** the raw data is transformed in *smarted data*. The same data is expressed in an standard format using RDF (*Resource Description Framework*) that is the foundation upon which the web of semantic data is built. [3]. In this model any resource can be uniquely identified and can be enriched with descriptive meta-data.

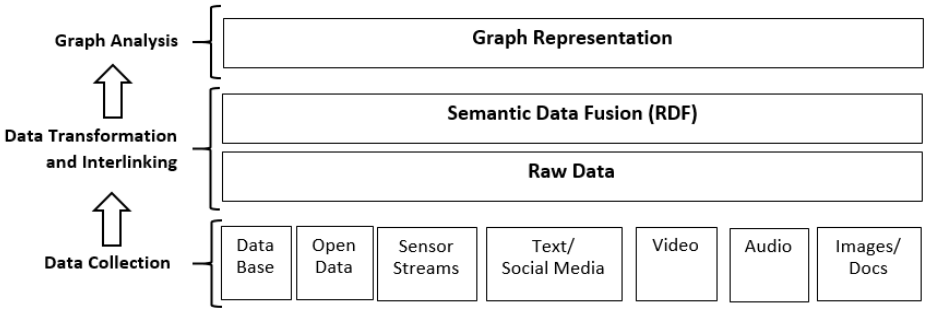

Figure 1: Graphical Big Data Analysis Framework

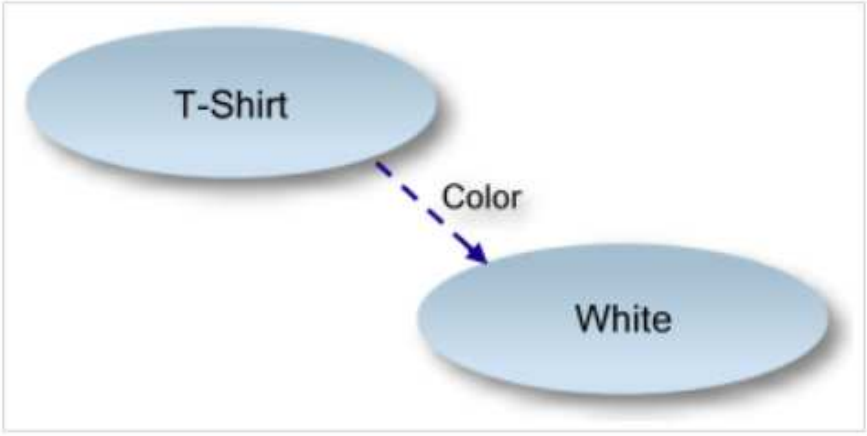

Figure 2: RDF Triple: Subject is the T-shirt, Predicate (property) is the color, Object is white (from http://www.linkeddatatools.com/introducing-rdf).

In RDF every statement (Triple) Figure 2 is expressed by three components: the *subject*, that is the identifier of the resource, the *predicate*, that indicates the property or attribute of the subject to be described and the *object*, the value of the predicate related to the subject. Using this semantic data model we can capture the semantic heterogeneity and the differences in the logical representation of data.

In the **Graph Analysis phase** we create a graph representation using semantic data fusion. Graphical representation of data enables graph analysis to discover more insights.
This framework is different from the Map Reduce framework and NoSQL that is based on data representation in the form of key-values pairs. This representation is very good for data parallelism processing but does not work when the data elements are related to each other because it does not capture explicit relationships between the data.

References

[1] Sultana N. Sayyad Deepak S. Tamhane. Big data analysis using hace theorem. *International Journal of Advanced Research in Computer Engineering and Technology*, 4:18–23, 2015.

[2] Narayanan Unny Geetha Manjunath Divanshu Gupta, Avinash Sharma. Graphical analysis and visualiation of big data in business domain. *Lecture Notes in Computer Science*, Volume 8883, 2014.

[3] Dario Cerizza Emanuele Della Valle, Irene Celino. *Semantic Web*. Pearson Addison Wesley, Milano, 2009.

[4] NIST Big Data Public Working Group. Nist big data interoperability framework: Definitions.

[5] Brad Brown Jacques Bughin Richard Dobbs Charles Roxburgh Angela Hung Byers James Manyika, Michael Chui. Big data: The next frontier for innovation, competition, and productivity.

[6] Rajan S.Patel Lynnette Brammer Mark S. Smolinski Larry Brilliant Jeremy Ginsberg, Matthew H. Mohebbi. Detecting influenza epidemics

using search engine query data. *Nature*, Volume 457, February 19, 2009, 2009.

[7] David Reinsel John Ganz. Extracting value from chaos.

[8] Doug Laney. 3d data management: Controlling data volume, velocity, and variety.

[9] Mark Maremont Leslie Scism. Insurers test data profiles to identify risky clients. *The Wall Street Journal*, November 19, 2010.

[10] Yunhao Liu Min Chen, Shiwen MAo. Big data: A survey. *Springer Science+Business Media New York*, pages 171–209, 2014.

[11] Microsoft Pat Helland. If you have too much data, then "good enough" is good enough. *Databases*, Volume 9, issue 5, May 23, 2011.

[12] Kenneth Cukier Viktor Mayer-Schnberger. *Big Data: A Revolution that will transform how we live, work and think*. John Murray, London, 2013.

[13] Kale Suvana Vilas. Big data mining. *International Journal of Computer Science and Management Research*, Ottobre:12–17, 2013.

[14] Gong-Qing Wu Wei Ding Xindong Wu, Xingquan Zhu. Data mining with big data. *IEEE Transactions on knowledge and data engineering*, Volume 26, No. 1, January 2014, 2014.