Big Data technology

Nicolae Sfetcu

08.08.2019

Sfetcu, Nicolae, "Big Data technology", SetThings (August 8, 2019), URL = https://www.setthings.com/en/big-data-technology/

Email: nicolae@sfetcu.com

A partial translation of:
Sfetcu, Nicolae, "Etica Big Data în cercetare", SetThings (6 iulie 2019), DOI: 10.13140/RG.2.2.27629.33761, MultiMedia Publishing (ed.), ISBN: 978-606-033-228-2, URL = https://www.setthings.com/ro/e-books/etica-big-data-in-cercetare/

The term Big Data refers to the extraction, manipulation and analysis of data sets that are too large to be routinely processed. Because of this, special software is used and, in many cases, also dedicated computers and hardware. Generally, for these data the analysis is done statistically. Based on the analysis of the respective data, predictions of certain groups of people or other entities are usually made, based on their behavior in various situations and using advanced analytical techniques. Thus, tendencies, needs and behavioral evolutions of these entities can be identified. Scientists use this data for research in meteorology, genomics, (Nature 2008) connectomics, complex physical simulations, biology, environmental protection, etc. (Reichman, Jones, and Schildhauer 2011)

With the increasing volume of data on the Internet, in social media, cloud computing, mobile devices and government data, Big Data is both a threat and an opportunity for researchers to manage and use this data while maintaining the rights of the involved people.

1. Definitions

Big Data usually includes sets of data that exceed the capacity of ordinary software and hardware, using unstructured, semi-structured and structured data, with an emphasis on unstructured data. (Dedić and Stanier 2017) Big Data has grown in size since 2012, from dozens of terabytes to many data exabytes. (Everts 2016) Making data efficient with Big Data involves machine learning to detect patterns, (Mayer-Schönberger and Cukier 2014) but often this data is a by-product of other digital activities.

A 2018 definition states that "Big data is where parallel computing tools are needed to handle data," which represents a turning point in computing, using parallel programming theories and the lack of assurances assumed by previous models. Big Data uses inductive statistics and concepts of identifying nonlinear systems to deduce laws (regressions, nonlinear relationships and causal effects) from large data sets with low information density to obtain relationships and dependencies or to make predictions of results and behaviors. (Billings 2013)

At European Union level there is no mandatory definition but, according to the Opinion 3/2013 of the European Working Group on data protection,

> "Big Data is a term that refers to the enormous increase in access to and automated use of information: It refers to the gigantic amounts of digital data controlled by companies, authorities and other large organizations which are subjected to extensive analysis based on the use of algorithms. Big Data may be used to identify general trends and correlations, but it can also be used such that it affects individuals directly." (European Economic and Social Committee 2017)

The problem with this definition is that it does not consider reusing personal data.

Regulation no. 2016/679 defines personal data (Article 4, paragraph 1) as

"any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person." (European Economic and Social Committee 2017)

The definition applies at EU level also to unidentified persons, but which can be identified by correlating anonymous data with other additional information. Personal data, once anonymized (or pseudo-anonymized), can be processed without the need for authorization, however, taking into account the risk of re-identifying the data subject.

2. Big Data dimensions

The data is shared and stored on servers, through the interaction between the entity involved and the storage system. In this context, Big Data can be classified into active systems (synchronous interaction, entity data is sent directly to the storage system), and passive systems (asynchronous interaction, data is collected through an intermediary and then entered into the system).

Also, the data can be transmitted directly, consciously or non-consciously (if the person whose data is transmitted is not notified on time and clearly). The data is then processed to generate statistics.

Depending on the target of the respective statistics analyzes, the data dimensions may be a) individual (only one entity is analyzed); social (there are analyzed discrete groups of entities within a population); and hybrids (when an entity is analyzed from the perspective of its belonging to an already defined group).

The current huge output of user-generated data is expected to grow by 2000% worldwide by 2020 and are often unstructured. (European Economic and Social Committee 2017) In general, Big Data is characterized by:

- Volume (amount of data);
- Variety (products from different sources in different formats);

- Speed (speed of online data analysis);
- Accuracy (data is uncertain and must be verified);
- Value (evaluated by analysis).

The volume of data produced and stored is currently evolving exponentially, over 90% of them being generated in the last four years. (European Economic and Social Committee 2017) Large volumes require high speed of analysis, with a strong impact on veracity. Incorrect data has the potential to cause problems when used in the decision process.

One of the major problems with Big Data is whether the complete data is needed to draw certain conclusions about their properties, or a sample is enough. Big Data contains in its name a term related to size, which is an important feature of Big Data. But (statistical) sampling allows the selection of correct data collection points from a larger set to estimate the characteristics of the entire population. Big Data can be sampled across different categories of data in the process of sample selection with the help of sampling algorithms for Big Data.

3. Technology

Data must be processed with advanced collection and analysis tools, based on predetermined algorithms, in order to obtain relevant information. Algorithms must also take into account invisible aspects for direct perceptions.

In 2004 Google published a paper about a process called MapReduce that offers a parallel processing model. (Dean and Ghemawat 2004) MIKE2.0 is also an open source application for information management. (MIKE2.0 2019) Several studies from 2012 have shown that the optimal architecture for addressing Big Data issues is multi-layered. A distributed parallel architecture distributes data on multiple servers (parallel execution environments) thus dramatically improving data processing speeds.

According to a report from the McKinsey Global Institute in 2011, the main components and ecosystems of Big Data are: (Manyika et al. 2011) data analysis techniques (machine learning, natural language processing, etc.), big data technologies (business intelligence, cloud computing, databases), and visualization (charts, graphs, other data views).

Big Data provides real-time or near real-time information, thus avoiding latency whenever possible.

4. Applications

Big data in government processes increases cost efficiency, productivity and innovation. Civil records are a source for Big Data. The processed data helps in critical areas of development, such as health care, employment, economic productivity, crime, security and management of natural disasters and resources. (Kvochko 2012)

Also, Big Data provides an infrastructure that allows for highlighting uncertainties, performance, and availability of components. Trends and predictions in the industry require a large amount of data and advanced prediction tools.

Big Data contributes to the improvement of healthcare by providing personalized medicines and prescriptive analyzes, clinical interventions with risk assessment and predictive analysis, etc. The level of data generated in health systems is very high. But there is a pressing problem with generating "dirty data", which increase with increasing volume of data, especially since most are unstructured and difficult to use. The use of Big Data in healthcare has generated significant ethical challenges, with implications on individual rights, privacy and autonomy, transparency and trust.

In the field of health insurance, data is collected on the "determinants of health", which helps to develop forecasts on health costs and to identify clients' health problems. This use is controversial, due to the discrimination of clients with health problems. (Allen 2018)

In the media and advertising, for Big Data, numerous information points are used about millions of people, to serve or transmit personalized messages or content.

In sports, Big Data can help improve competitors' training and understanding using specific sensors and predict future performance of athletes. Sensors attached to Formula 1 cars collect, inter alia, tire pressure data to make fuel burning more efficient.

Big data and information technology complement each other, helping together to develop the Internet of Things (IoT) for interconnecting smart devices and collecting sensory data used in different fields.

5. In research

In science, Big Data systems are used extensively in particle accelerators at CERN (150 million sensors transmit data 40 million times per second, for about 600 million collisions per second, of which they are used after filtering only 0.001% of the total data obtained), (Brumfiel 2011) in astrophysical radio telescopes built from thousands of antennas, decoding the human genome (initially it took a few years, with Big Data can be done in less than a day), climate studies, etc. .

Big IT companies use data warehouses of the order of tens of petabytes for search, recommendations and merchandising. Most data is collected by Facebook, with over 2 billion monthly active users (Constine 2017) and Google with over 100 billion searches per month. (Sullivan 2015)

The research uses a lot of encrypted search and cluster formation in Big Data. Developed countries are currently investing heavily in Big Data research. Within the European Union, these researches are included in the Horizon 2020 program. (European Commission 2019)

Often, research programs use API resources from Google and Twitter to gain access to their Big Data systems, for free or at no cost.

Large data sets come with algorithmic challenges that previously did not exist, and it is imperative to fundamentally change the processing methods. To this end, special workshops have

been created that bring together scientists, statisticians, mathematicians and practitioners to discuss the algorithmic challenges of Big Data.

Bibliography

Allen, Marshall. 2018. "Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates." Text/html. ProPublica. July 17, 2018. https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates.

Billings, Stephen A. 2013. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains." Wiley.Com. 2013. https://www.wiley.com/en-ro/Nonlinear+System+Identification%3A+NARMAX+Methods+in+the+Time%2C+Frequency%2C+and+Spatio+Temporal+Domains-p-9781119943594.

Brumfiel, Geoff. 2011. "High-Energy Physics: Down the Petabyte Highway." *Nature* 469 (7330): 282–83. https://doi.org/10.1038/469282a.

Constine, Josh. 2017. "Facebook Now Has 2 Billion Monthly Users… and Responsibility." *TechCrunch* (blog). 2017. http://social.techcrunch.com/2017/06/27/facebook-2-billion-users/.

Dean, Jeffrey, and Sanjay Ghemawat. 2004. "MapReduce: Simplified Data Processing on Large Clusters." http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf.

Dedić, Nedim, and Clare Stanier. 2017. "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery." In *Innovations in Enterprise Information Systems Management and Engineering*, edited by Felix Piazolo, Verena Geist, Lars Brehm, and Rainer Schmidt, 114–22. Lecture Notes in Business Information Processing. Springer International Publishing.

European Commission. 2019. "Horizon 2020." Text. Horizon 2020 - European Commission. 2019. https://ec.europa.eu/programmes/horizon2020/en.

European Economic and Social Committee. 2017. "The Ethics of Big Data: Balancing Economic Benefits and Ethical Questions of Big Data in the EU Policy Context." European Economic and Social Committee. February 22, 2017. https://www.eesc.europa.eu/en/our-work/publications-other-work/publications/ethics-big-data.

Everts, Sarah. 2016. "Information Overload." Science History Institute. July 18, 2016. https://www.sciencehistory.org/distillations/magazine/information-overload.

Kvochko, Elena. 2012. "Four Ways to Talk About Big Data." Text. Information and Communications for Development. December 4, 2012. http://blogs.worldbank.org/ic4d/four-ways-to-talk-about-big-data.

Manyika, James, Michael Chui, Jaques Bughin, and Brad Brown. 2011. "Big Data: The next Frontier for Innovation, Competition, and Productivity." 2011. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Reprint edition. Boston: Eamon Dolan/Mariner Books.

MIKE2.0. 2019. "Big Data Solution Offering - MIKE2.0, the Open Source Methodology for Information Development." 2019. http://mike2.openmethodology.org/wiki/Big_Data_Solution_Offering.

Nature. 2008. "Community Cleverness Required." *Nature* 455 (7209): 1. https://doi.org/10.1038/455001a.

Reichman, O. J., Matthew B. Jones, and Mark P. Schildhauer. 2011. "Challenges and Opportunities of Open Data in Ecology." *Science* 331 (February): 703. https://doi.org/10.1126/science.1197962.

Sullivan, Danny. 2015. "Google Still Doing At Least 1 Trillion Searches Per Year." Search Engine Land. January 16, 2015. https://searchengineland.com/google-1-trillion-searches-per-year-212940.