International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Big Data : Features, Architecture, Research and Applications

Marwane Smaiti[1], Mostafa Hanoune[2]

[1, 2]Doctoral Studies, Mathematics, Computer Science, and Information Technologies, Hassan II Mohammedia-Casablanca University, Ben M'sik School of Sciences, BP 7955 Casablanca, Morocco

Abstract : ***In this article, we present a description of the dimensions of Big Data, for the study of the application of this notion in several fields of application. We focus on the volume, variety and frequency of generations of mass data. We are flying over areas where effective processing of large amounts of data has become a necessity for the survival and development of organizations. Indeed, users of information today must ensure a unique competitive advantage by taking advantage of the data coming from the various players in their environment.***

Keywords: Big Data, dimensions

1. Introduction

Big Data is a term that refers to the processing of "too large" and "too complex" quantities of data. Conventional methods of data processing are useless when an average size is exceeded. Challenges include: data analysis, data collection, data retrieval, data storage, data sharing, and data protection. In common language, Big Data refers to prediction techniques, to draw useful conclusions from typical data behaviors, regardless of size [1].

The study of data sets has applications in several areas, such as preventing epidemics, predicting trends in a physical or virtual market, and combating crime. Several actors, including scientists, governments and researchers face real difficulties in dealing with important data flows in several areas, such as the Internet, metrology, e-learning, and particle simulation. In the scientific field, non-linearity is a source of phenomenal quantity of data varying with position and time. This limits many advances in the field of scientific research.

The available quantity of information exponentially increases by what the data, nowadays, are collected by tools more and more accessible and effective. Large companies are wondering about the impact of having access to large amounts of data across the organization.

The use of conventional data processing means, such as database management, and conventional statistics, is not possible. Indeed, this may require thousands of servers working in parallel [2].

The term "Big Data" is relative. In fact, it depends on the users and their means. In addition, improving data processing tools makes Big Data a changing concept over time. Big Data is a collection of data whose processing is not feasible using conventional tools. Studies also looked at the impact of Big Data on risk minimization and corporate decision-making. The concept of Big Data is one of the most complex challenges of the second decade of the century [3 The term "Big Data" was introduced in the 1990s [4], to designate quantities of data whose processing exceeds the capabilities of the technologies available at the time. The size of a typical Big Data system varies over time. In 2012, it was a dozen terabytes. Today, we are talking more of several petabytes. Big Data requires a set of techniques that go beyond the classical framework to focus on their overall behavior rather than the behavior of individuals. In 200, Doug Laney demonstrated that the challenges and opportunities created by the appearance of Big Data have three-dimensional (3D) behavior [5], distinguishing between volume (quantity of data), speed Input and output), and the variety (sources and families of data). Gartner in 2012 [6] has updated its definition of Big Data as follows: "Big Data is high volume, high velocity, and / or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and Process optimization ". So the notion of Big Data depends on the available data processing tools, as well as the conclusions drawn from the data, using a given method of processing.

Today, the 3Vs convention is used to define the Big Data: Volume, Speed, Variety. A new V (Veracity) has recently been introduced by some organizations [7].

2. Big Data Features

We summarize the 3Vs defining the characteristics of the Big Data in Figure 1, and in a second step we explain them in detail.

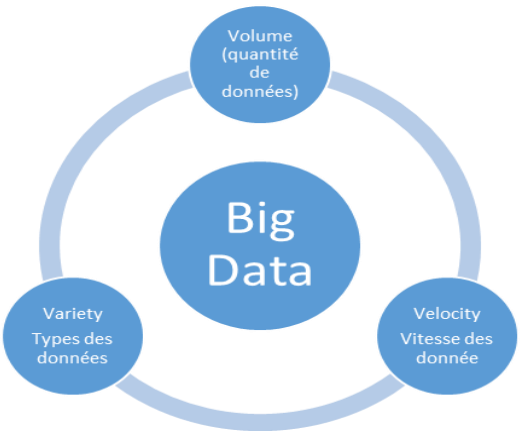

Figure 1: Big Data Features

3. Volume : definition et examples

The volume of massive data refers to the large amounts of data generated every second. These are not terabytes, but zettabytes or brontobytes. This is an infinitely large amount of data: if we take the data generated in the world between the oldest date and 2008, the same amount will be generated every minute. This makes conventional processing methods unnecessary for handling Big Data. 90% of the data created had been created in the last two years. From now on, the data created in the world will double every year. By 2020, we will have 50 times more data than we had in 2011. The volume of data is enormous and is a huge contributor to the digital universe, which is still expanding: the Internet has Objects with sensors all over the world. All devices creating data every second. The era of a trillion of sensors is ours.

For aircraft, they generate about 2.5 billion terabytes of data each year from the sensors installed in the engines. The Auto-driving cars will generate 2 Petabyte data each year. In addition, the agricultural industry generates massive amounts of data with sensors installed in tractors. Shell, for example, uses super sensitive sensors to find extra oil in wells and if they install these sensors at every 10,000 wells they collect about 10 exabytes of data per year. Again, this is absolutely nothing compared to the square kilometer telescope that generates 1 exabyte of data per day.

In the past, creating so much data would have caused serious problems. Nowadays, with lower storage costs, better storage solutions like Hadoop, as well as powerful algorithms to draw conclusions from all these data do not pose a problem.

4. Velocity : definition et examples

The speed indicates the speed at which new data is generated and the speed at which the data moves. Technology now allows us to analyze data as it is generated (sometimes called in-memory analysis), without ever putting it into databases: Speed is the rate at which data is created, stored, analyzed, and Displayed.

In the past, when batch processing was a common practice, it was normal to receive an update of the database every evening or even weekly. Computers and servers required time to process data and update databases. In the era of large data, data is created in real time or almost in real time. With the availability of devices connected to the Internet, wireless or via cables, machines and peripherals can transmit their data as soon as they are created.

The speed at which data is created is infinitely large: every minute, we watch 100 hours of video on YouTube. In addition, every minute, more than 200 million emails are sent, about 20 million photos are viewed and 30,000 downloaded on Flickr, nearly 300,000 tweets are sent and almost 2.5 million queries on Google are made.

The challenge for information science researchers is to cope with the enormous speed with which data is created and used in real time.

5. Variety : definition et examples

Variety refers to the different types of data we can use. Over time, we focused on structured data that is perfectly integrated into tables or relational databases, such as financial data. With large data technologies, we can now analyze and aggregate data of different types, such as messages, conversations on social networks, photos, sensor data, and video or voice recordings.

In the past, all the data that was created was structured data, well adjusted in columns and rows. On the other hand, today, 90% of the data generated by a typical organization is unstructured data. Data are presented in a variety of formats: structured data, semi-structured data, unstructured data, and even complex structured data (amalgamating structured and unstructured data). The wide variety of data requires a different approach and different techniques to store all the raw data in their original format.

There are many types of data, and each of these types of data requires different types of analysis and different tools to use. Social media like Facebook messages or Tweets can give different ideas, such as feeling sentiment on your brand, while sensory data will give you information about how a product is used and what are the possible failures In the marketed product.

In short, with better storage solutions, the challenge of rapid data creation, and the development of new data storage and analysis tools, there are new, practical approaches to performing data analyzes to Informed decisions: these are the tools of Big Data.

6. Big Data Architecture

In 2000, Seisint developed a distributed file-sharing platform in C ++ for data storage and query. The system stores and distributes structured, semi-structured and unstructured data across multiple servers. Users could create queries in a C ++ language called ECL. ECL uses an "apply schema on read" method to infer the structure of the stored data when it is queried, instead of when it is stored. In 2004, LexisNexis acquired Seisint Inc. [8] And in 2008 acquired ChoicePoint, Inc. [9] and their parallel high-speed processing platform. Both platforms were merged into High Performance Computing Cluster (HPCC) systems and in 2011, HPCC was an open source under the Apache v2.0 license. Quantcast File System was available at about the same time. [10]

In 2004, Google published an article on a process called MapReduce that uses a similar architecture. The MapReduce concept provides a parallel processing model, and an associated implementation has been released to handle huge amounts of data. With MapReduce, queries are distributed between the parallel nodes and processed in parallel (Map step). The results are then collected and distributed (reduction step). This technology was very successful [11], while others wanted to reproduce the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an open source Apache project named Hadoop [12].

A new methodology [13] deals with the processing of large data in terms of useful permutations of data sources, the complexity of interrelations and the difficulty of removing (or modifying) individual records [14].

Studies in 2012 have shown that a multi-layered architecture is an option for dealing with mass data problems. A distributed parallel architecture distributes data across multiple servers; These parallel execution environments can dramatically improve the speed of data processing. This type of architecture inserts data into a parallel DBMS, which implements the use of the MapReduce and Hadoop platforms. This type of framework seeks to make the processing power transparent to the end user by using a front-end application server. [15]

Large data analysis for manufacturing applications is marketed as a 5C architecture (connection, conversion, cyber, cognition and configuration). [16]

The concept of "data lake" allows an organization to move from a centralized control to a shared model to respond to the changing dynamics of information management. This allows rapid segregation of data, thus reducing the overhead time. The architecture of a typical modern mass data processing system is described in FIG 2.

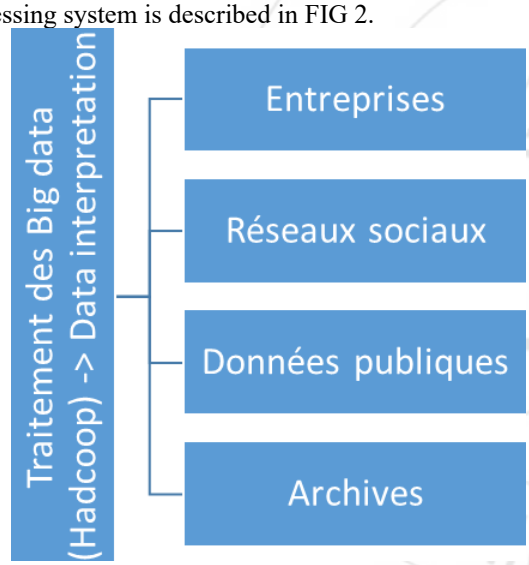

Figure 2: Architecture d'un système de traitement de Big Data

7. Big Data Applications

The emergence of Big Data has increased the demand for information management specialists to such an extent that Software AG, Oracle, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $ 15 billion on business Specialized in data management and analysis. In 2010, this industry was worth more than $ 100 billion and increased to nearly 10 percent per year: about twice as fast as the entire software industry. [17]

Developed economies are increasingly using data-intensive technologies. There are 4.6 billion mobile phone subscriptions worldwide and between 1 billion and 2 billion people accessing the Internet [18]. Between 1990 and 2005, more than one billion people around the world entered the middle class, which means that more people are becoming more literate, resulting in information growth. The global capacity to exchange information through telecommunications networks was 281 petabytes in 1986, 471 petabytes in 1993, 2.2 exabytes in 2000, 65 exabytes in 2007 [19], and 667 exabytes According to one estimate, one-third of the overall stored information is in the form of alphanumeric text and still image data, [20] which is the most useful format for most important data applications. This also shows the potential for unused data (ie, in the form of video and audio content). A tremendous search effort is needed to take advantage of the data potential available in the globe.

While many vendors offer solutions in the market for Big Data, experts recommend the development of customized internal solutions to solve the problem of the company if the company has sufficient technical capacities. [21] Let us see through our proposed ontology To model the parameters related to the definition of the decisional problems during the process of the economic intelligence, and to show the relations connecting these parameters with a company.

8. Research in Big Data

Encrypted research and training in the field of clusters in the Big Data were demonstrated in March 2014 at the American Society of Engineering Education. Gautam Siwach is committed to the challenges of Big Data by the MIT Computer Science and Artificial Intelligence Laboratory and Dr. Amir Esmailpour to the Research Group UNH has studied the main features of mass data such as cluster construction and Their interconnections. They focused on the security of Big Data and the actual orientation of the term to the presence of different types of data in encrypted form at the cloud interface by providing raw definitions and real-time examples within the technology. In addition, they proposed an approach to identify the encoding technique to move forward to an accelerated search on encrypted text leading to security improvements in the Big Data domain. [22]

In March 2012, the White House announced the creation of a "Big Data Initiative": a national institution consisting of six federal departments and agencies that spent more than $ 200 million on major research projects in the treatment of Data [11].

The initiative included a $ 10 million grant over five years to AMPLab [23] at the University of California, Berkeley, as part of the National Science Foundation [24]. AMPLab has also received funding from DARPA and more than a dozen industrial stakeholders and uses Big Data techniques to address a wide range of issues: prediction of traffic congestion [25] To the fight against cancer [26].

The White House Big Data Initiative also included a commitment from the Department of Energy to provide $ 25 million in five-year funding to create the Evolving Data Management, Analysis and Visualization Institute [27], Led by Lawrence Berkeley National Laboratory. The SDAV Institute aims to bring together the expertise of six national laboratories and seven universities to develop new tools to

help scientists manage and visualize ministry supercomputer data.

The State of Massachusetts announced the Massachusetts Big Data Initiative in May 2012, which provides state and private sector funding to a variety of research institutions. [28] The Massachusetts Institute of Technology hosts the Intel Science and Technology Center for large data in the MIT Computer Science and Artificial Intelligence Laboratory, combining government, business, and institutional funding and research efforts.

The European Finance Commission, through its seventh framework program, the two-year Big Public Forum, to engage companies, academics and other stakeholders to discuss major data issues. The project aims to define a research and innovation strategy to guide the European Commission's support actions in the successful implementation of the large data economy. The results of this project will be used as a contribution to Horizon 2020, their next framework program. [29].

The British government announced in March 2014 the founding of the Alan Turing Institute, named after the computer pioneer and code-breaker, which will focus on new ways to collect and analyze large datasets. [thirty]

On the inspirational day of the Canadian University of Waterloo at Stratford Campus, participants demonstrated how the use of data visualization can increase the understanding and appeal of large datasets and communicate their history to the world [31].

To make manufacturing more competitive in the United States (and the world), it is necessary to integrate more American ingenuity and innovation in manufacturing; As a result, the National Science Foundation has awarded the University of Industry's Intelligent Maintenance Systems (IMS) research center at the University of Cincinnati the opportunity to focus on the development of advanced predictive tools and Techniques to be applied in a Big Data environment. In May 2013, IMS Center held a large-data industry-oriented industry advisory meeting where presenters from various industrial companies discussed their concerns, challenges and future goals in the Big Data environment.

Anyone can use the Application Programming Interfaces (APIs) provided by Big Data holders, such as Google and Twitter, to conduct research in the social and behavioral sciences. Often, these APIs are provided free of charge. [32] Tobias Preis et al. Used Google Trends data to demonstrate that Internet users in countries with higher per capita gross domestic product (GDP) are more likely to look for information about the future than information about the past. The results suggest that there may be a link between online behavior and real economic indicators [16] [17] [18]. The authors of the study examined Google query records based on the research volume ratio for the upcoming year (2011) versus the previous year's volume of research (2009), which they call [33]. They compared the index of future direction with the GDP per capita of each country and found a strong trend for the countries where Google users learn more about the future to have a higher GDP. The results suggest that there may be a relationship between a country's economic success and the information retrieval behavior of its citizens captured in large data.

Tobias Preis and colleagues Helen Susannah Moat and H. Eugene Stanley have introduced a method for identifying online precursors for stock movements using trading strategies based on search volume data provided by Google Trends [34]. Their analysis of Google search volume for 98 terms of variable financial relevance, published in Scientific Reports [35], suggests that search volume increases for financially relevant search terms tend to precede large losses in financial markets [36] [37] [38] [39] [40].

Large sets of data come with algorithmic challenges that did not exist before. It is therefore necessary to modify fundamentally the treatment methods [41].

Workshops on algorithms for modern mass data sets (MMDS) bring together computer scientists, statisticians, mathematicians and data analysts to discuss the algorithmic challenges of Big Data.

9. Conclusion

Large data has been called a "fashion" in scientific research and its use has even been amused as an absurd practice in a satirical example on "data on pigs". [42] Researcher Danah Bboyd raised concerns about the use of important science data neglecting principles such as selecting a representative sample by being too preoccupied with manipulating the huge amounts of data [43]]. This approach can lead to bias results in one way or another. The integration of heterogeneous data - some of which could be considered as "large data" and some not - presents hefty logistical and analytical problems, but many researchers claim that these integrations are likely to represent the new frontiers The most promising.

References

[1] "The World's Technological Capacity to Store, Communicate, and Compute Information". MartinHilbert.net. Retrieved 13 April 2016.
[2] New Horizons for a Data-Driven Economy – Springer. doi:10.1007/978-3-319-21569-3.
[3] boyd, dana; Crawford, Kate (September 21, 2011). "Six Provocations for Big Data". Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. doi:10.2139/ssrn.1926431.
[4] "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.
[5] "Community cleverness required". Nature. 455 (7209): 1. 4 September 2008. doi:10.1038/455001a.
[6] Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011). "Challenges and Opportunities of Open Data in Ecology". Science. 331 (6018): 703–5. doi:10.1126/science.1197962. PMID 21311007.
[7] Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". Gigaom Blog.

[8] Segaran, Toby; Hammerbacher, Jeff (2009). Beautiful Data: The Stories Behind Elegant Data Solutions. O'Reilly Media. p. 257. ISBN 978-0-596-15711-1.
[9] Hilbert, Martin; López, Priscila (2011). "The World's Technological Capacity to Store, Communicate, and Compute Information". Science. 332 (6025): 60–65. doi:10.1126/science.1200970. PMID 21310967.
[10] "IBM What is Big Data? – Bringing Big Data to the enterprise". www.ibm.com. Retrieved 2013-08-26.
[11] Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity", December 2012
[12] Jacobs, A. (6 July 2009). "The Pathologies of Big Data". ACMQueue.
[13] Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data". Release 2.0. Sebastopol CA: O'Reilly Media (11).
[14] John R. Mashey (25 April 1998). "Big Data ... and the Next Wave of InfraStress" (PDF). Slides from invited talk. Usenix. Retrieved 28 September 2016.
[15] Steve Lohr (1 February 2013). "The Origins of 'Big Data': An Etymological Detective Story". New York Times. Retrieved 28 September 2016.
[16] Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science. 7: 1–5.
[17] Ibrahim; Targio Hashem, Abaker; Yaqoob, Ibrar; Badrul Anuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "Big Data" on cloud computing: Review and open research issues". Information Systems. 47: 98–115. doi:10.1016/j.is.2014.07.006.
[18] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety" (PDF). Gartner. Retrieved 6 February 2001.
[19] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. Archived from the original on 10 July 2011. Retrieved 13 July 2011.
[20] De Mauro, Andrea; Greco, Marco; Grimaldi, Michele (2016). "A Formal definition of Big Data based on its essential Features". Library Review. 65: 122–135. doi:10.1108/LR-06-2015-0061.
[21] "What is Big Data?". Villanova University.
[22] Grimes, Seth. "Big Data: Avoid 'Wanna V' Confusion". InformationWeek. Retrieved 5 January 2016.
[23] Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review.". martinhilbert.net. Retrieved 2015-10-07.
[24] What is Big Data?. 12 August 2015 – via YouTube.
[25] Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: a revolution that will transform how we live, work and think. London: John Murray.
[26] "Digital Technology & Social Change".
[27] http://www.bigdataparis.com/presentation/mercredi/PDelort.pdf?PHPSESSID=tv7k70pcr3egpi2r6fi3qbjtj6#page=4
[28] Billings S.A. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains". Wiley, 2013
[29] "le Blog ANDSI » DSI Big Data".
[30] Les Echos (3 April 2013). "Les Echos – Big Data car Low-Density Data ? La faible densité en information comme facteur discriminant – Archives". lesechos.fr.
[31] Wu, D., Liu. X., Hebert, S., Gentzsch, W., Terpenny, J. (2015). Performance Evaluation of Cloud-Based High Performance Computing for Finite Element Analysis. Proceedings of the ASME 2015 International Design Engineering Technical Conference & Computers and Information in Engineering Conference (IDETC/CIE2015), Boston, Massachusetts, U.S.
[32] Tobias Preis (24 May 2012). "Supplementary Information: The Future Orientation Index is available for download" (PDF). Retrieved 2012-05-24.
[33] Philip Ball (26 April 2013). "Counting Google searches predicts market movements". Nature. Retrieved 9 August 2013.
[34] Tobias Preis, Helen Susannah Moat and H. Eugene Stanley (2013). "Quantifying Trading Behavior in Financial Markets Using Google Trends". Scientific Reports. 3: 1684. doi:10.1038/srep01684. PMC 3635219Freely accessible. PMID 23619126.
[35] Nick Bilton (26 April 2013). "Google Search Terms Can Predict Stock Market, Study Finds". New York Times. Retrieved 9 August 2013.
[36] Christopher Matthews (26 April 2013). "Trouble With Your Investment Portfolio? Google It!". TIME Magazine. Retrieved 9 August 2013.
[37] Philip Ball (26 April 2013). "Counting Google searches predicts market movements". Nature. Retrieved 9 August 2013.
[38] Bernhard Warner (25 April 2013). "'Big Data' Researchers Turn to Google to Beat the Markets". Bloomberg Businessweek. Retrieved 9 August 2013.
[39] Hamish McRae (28 April 2013). "Hamish McRae: Need a valuable handle on investor sentiment? Google it". The Independent. London. Retrieved 9 August 2013.
[40] Richard Waters (25 April 2013). "Google search proves to be new word in stock market prediction". Financial Times. Retrieved 9 August 2013.
[41] David Leinweber (26 April 2013). "Big Data Gets Bigger: Now Google Trends Can Predict The Market". Forbes. Retrieved 9 August 2013.
[42] Jason Palmer (25 April 2013). "Google searches predict market moves". BBC. Retrieved 9 August 2013.
[43] E. Sejdić, "Adapt current tools for use with big data," Nature, vol. vol. 507, no. 7492, pp. 306, Mar. 2014.