International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

A survey on Big Data : Techniques and Technologies

Vinay Chaorasiya
Computer Science
S.R.C.E.M., Banmore, (M.P.)
vinviki3@gmail.com

Aparajit Shrivastava
Dept of Computer Science
S.R.C.E.M., Banmore, (M.P.)
aparajit2k5@gmail.com

Abstract--**Nowadays companies are starting to realize the importance of using more data in order to support decision for their strategies. It was said and proved through study cases that "More data usually beats better algorithms". With this statement companies started to realize that they can chose to invest more in processing larger sets of data rather than investing in expensive algorithms. The large quantity of data is better used as a whole because of the possible correlations on a larger amount, correlations that can never be found if the data is analyzed on separate sets or on a smaller set. A larger amount of data gives a better output but also working with it can become a challenge due to processing limitations. This paper intends to define the concept of Big Data and stress the importance of Big Data Analytics.**

Keywords: Big Data, Big Data Analytics, Sentiment analysis , Cloud computing

1. Introduction

Nowadays the Internet represents a big space where great amounts of information are added every day.. Also according to this study there are 100 Terabytes updated daily through Facebook, and a lot of activity on social networks this leading to an estimate of 35 Zettabytes of data generated annually by 2020. Just to have an idea of the amount of data being generated, one zettabyte (ZB) equals 10^{21} bytes, meaning 10^{12} GB. [1] We can associate the importance of Big Data and Big Data Analysis with the society that we live in. Today we are living in an Informational Society and we are moving towards a Knowledge Based Society. In order to extract better knowledge we need a bigger amount of data. The Society of Information is a society where information plays a major role in the economical, cultural and political stage. In the Knowledge society the competitive advantage is gained through understanding the information and predicting the evolution of facts based on data. The same happens with Big Data. Every organization needs to collect a large set of data in order to support its decision and extract correlations through data analysis as a basis for decisions. In this paper we will define the concept of Big Data, its importance and different perspectives on its use. In addition we will stress the importance of Big Data Analysis and show how the analysis of Big Data will improve decisions in the future.

2. Big Data Concept

The term "Big Data" was first introduced to the computing world by Roger Magoulas from O'Reilly media in 2005 in order to define a great amount of data that traditional data management techniques cannot manage and process due to the complexity and size of this data. A study on the Evolution of Big Data as a Research and Scientific Topic shows that the term "Big Data" was present in research starting with 1970s but has been comprised in publications in 2008. [2] Nowadays the Big Data concept is treated from different points of view covering its implications in many fields. According to MiKE 2.0, the open source standard for Information Management, Big Data is defined by its size, comprising a large, complex and independent collection of data sets, each with the potential to interact. In addition, an important aspect of Big Data is the fact that it cannot be handled with standard data management techniques due to the inconsistency and unpredictability of the possible combinations. [3]

In IBM's view Big Data has four aspects: Volume: refers to the quantity of data gathered by a company. This data must be used further to obtain important knowledge; Velocity: refers to the time in which Big Data can be processed. Some activities are very important and need immediate responses, that is why fast processing maximizes efficiency; Variety: Refers to the type of data that Big Data can comprise. This data can be structured as well as unstructured; Veracity: refers to the degree in which a leader trusts the used information in order to take decision. So getting the right correlations in Big Data is very important for the business future. [4]. In addition, in Gartner's IT Glosarry Big Data is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. [5] According to Ed Dumbill chair at the O'Reilly Strata Conference, Big Data can be described as, "data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it." [6]

In a simpler definition we consider Big Data to be an expression that comprises different data sets of very large, highly complex, unstructured, organized, stored and processed using specific methods and techniques used for business processes. There are a lot of definitions on Big Data circulating around the world, but we consider that the most important one is the one that each leader gives to its one company's data. The way that Big Data is defined has implication in the strategy of a business. Each leader has to

define the concept in order to bring competitive advantage for the company.

3. The importance of Big Data

The main importance of Big Data consists in the potential to improve efficiency in the context of use a large volume of data, of different type. If Big Data is defined properly and used accordingly, organizations can get a better view on their business therefore leading to efficiency in different areas like sales, improving the manufactured product and so forth.

Big Data can be used effectively in the following areas:

- In information technology in order to improve security and troubleshooting by analyzing the patterns in the existing logs;
- In customer service by using information from call centers in order to get the customer pattern and thus enhance customer satisfaction by customizing services;
- In improving services and products through the use of social media content. By knowing the potential customers preferences the company can modify its product in order to address a larger area of people;
- In the detection of fraud in the online transactions for any industry;
- In risk assessment by analyzing information from the transactions on the financial market. In the future we propose to analyze the potential of Big Data and the power that can be enabled through Big Data Analysis.

4. Big Data challenges

The understanding of Big Data is mainly very important. In order to determine the best strategy for a company it is essential that the data that you are counting on must be properly analyzed. Also the time span of this analysis is important because some of them need to be performed very frequent in order to determine fast any change in the business environment. Another aspect is represented by the new technologies that are developed every day. Considering the fact that Big Data is new to the organizations nowadays, it is necessary for these organizations to learn how to use the new developed technologies as soon as they are on the market. This is an important aspect that is going to bring competitive advantage to a business. The need for IT specialists it is also a challenge for Big Data. According to McKinsey's study on Big Data called Big Data: The next frontier for innovation, there is a need for up to 190,000 more workers with analytical expertise and 1.5 million more data-literate managers only in the United States. This statistics are a proof that in order for a company to take the Big Data initiative has to either hire experts or train existing employees on the new field.

5. Big Data Analytics

The world today is built on the foundations of data. Lives today are impacted by the ability of the companies to dispose, interrogate and manage data. The development of technology infrastructure is adapted to help generate data, so that all the offered services can be improved as they are used. As an example, internet today became a huge information-gathering platform due to social media and online services. At any minute they are added data. measured in gigabytes, since data is bigger there are used etabytes, exabytes, zettabytes and yottabytes. In order to manage the giant volume of unstructured data stored, it has been emerged the "Big Data" phenomena. It stands to reason that in the commercial sector Big-Data has been adopted more rapidly in data driven industries, such as financial services and telecommunications, which it can be argued, have been experiencing a more rapid growth in data volumes compared to other market sectors, in addition to tighter regulatory requirements and falling profitability. At first, Big Data was seen as a mean to manage to reduce the costs of data management. Now, the companies focus on the value creation potential. In order to benefit from additional insight gained there is the need to assess the analytical and execution capabilities of "Big Data". To turn big data into a business advantage, businesses have to review the way they manage data within data centre. The data is taken from a multitude of sources, both from within and without the organization. It can include content from videos, social data, documents and machine-generated data, from a variety of applications and platforms. Businesses need a system that is optimised for acquiring, organising and loading this unstructured data into their databases so that it can be effectively rendered and analysed. Data analysis needs to be deep and it needs to be rapid and conducted with business goals in mind.

6. Sentiment analysis

The social communication is the great tool for exploring the thoughts and interests based on single topic. The sentiment analysis is combination of opinion, sentiment, emotions from the given texts. The datasets are considers as comments created from the user profiles, likes, share, mutual interest in communication. According to estimates, the volume of business data worldwide, across almost companies, doubles every year. Taking advantage of sophisticated machine learning techniques to exploit the knowledge hidden in this huge volume of data, they successfully improve efficiency of their pricing strategies and advertising campaigns. The management of their inventory and supply chains also significantly benefits from the large-scale warehouse. In the era of information, almost every big company encounters Big Data problems, especially for multinational corporations. On the one hand, those companies mostly have a large number of customers around the world. On the other hand, there are very large volume and velocity of their transaction data. The first impression of Big Data is its volume, so the biggest and most important challenge is scalability when we deal with the Big Data analysis tasks. In the aspect of Big Data analytical techniques, increment algorithms have good scalability property, not for all machine learning algorithms. This shift in processors leads to the development of parallel computing. For

those real-time Big Data applications, like navigation, social networks, finance, biomedicine, astronomy, intelligent transport systems, and internet of thing, timeliness is at the top priority. How can we grantee the timeliness of response when the volume of data will be processed is very large? It is still a big challenge for stream processing involved by Big Data. It is right to say that Big Data not only have produced many challenge and changed the directions of the development of the hardware, but also in software architectures.

7. Opinion Mining

Opinion Mining also called sentiment analysis is a Process of finding user‟s opinion towards a topic. Opinion mining concludes whether user‟s view is positive, negative, or neutral about product, topic, event etc. Opinion mining involves analyzing user‟s opinion, attitude, and emotion towards particular topic. This consists of first categories text into subjective and objective information, and then finding polarity in subjective text. Opinion Mining can be performed word, sentence or document level. Fig 1 Opinion retrieval is a process of collecting reviews text from review websites. Information retrieval techniques such as web crawler can be applied to collect review text data from many sources and store them in database. This step involves retrieval of reviews, micro-blogs, comments etc of user. We should only consider the data which contain subjective data but not the objective data. Reviews are retrieved by query based information retrieval techniques. [14]

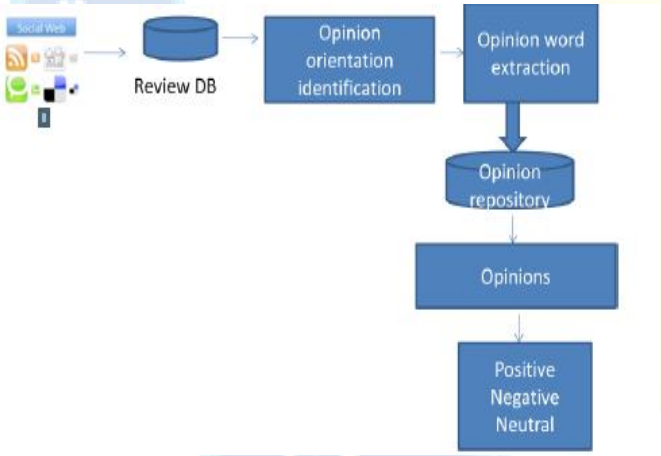

Figure. 1. opinion mining process

8 Related Work

When the datasets are large, some information fusion algorithms might not scale up well. For example, if an algorithm needs to load data into memory constantly, the program may run out of memory for large datasets. A simple and complete system for sentiment mining on large datasets using a Na¨ıve Bayes Classifier with the Hadoop framework [18].Facebook post identification (ID) is needed to allow the extraction of all the comments from the selected Facebook post (N. azmina m. zamani, siti z. z. abidin) [19]. To collect data, they created and registered a Facebook Connect application, called *iFeel*. Allowing the app to require specific Facebook privileges allowed them to host the app from Stanford.edu accounts, making it available to anyone and allowing to gather status updates quickly and efficiently [19].The sentiment classification of user posts in Twitter during the Hurricane Sandy and visualize these sentiments on a geographical map centred around the hurricane. Then it show how users' sentiments change according not only to users‟ locations, but also based on the distance from the disaster [13]. Generally, IFrames offer a variety of manual configurations in regards with the FBML canvas pages in which most of the contents are automatically configured from Facebook [13]. Therefore, this raises the need of considering the different user characteristics in order to adapt the system according to the user needs and other relevant aspects. [17] Finally, in recent years, due to the increasing amount of information delivered through social networks, many researches are focusing on applying sentiment analysis to these data [12].

9. Conclusion

In this paper we will define the concept of Big Data, its importance and different perspectives on its use. In addition we also discussed the importance of Big Data Analysis and show how the analysis of Big Data will improve decisions in the future. When conducting serious research or making every-day decisions, we often look for other people's opinions. We consult political discussion forums when casting a political vote, read consumer reports when buying appliances, ask friends to recommend a restaurant for the evening. The Internet is increasingly both the forum for discussion and source of information for a growing number of people. As a response to the growing availability of informal, opinionated texts like blog posts and product review websites, a field of Sentiment Analysis has sprung up in the past decade to address the question What do people feel about a certain topic? Bringing together researchers in computer science, computational linguistics, data mining, psychology, and even sociology, sentiment analysis expands the traditional fact-based text analysis to enable opinion-oriented

REFERENCES

[1] G. Noseworthy, Infographic: Managing the Big Flood of Big Data in Digital Marketing, 2012 http://analyzingmedia.com/2012/infographic-big-flood-of-big-data-in-digital-marketing/
[2] H. Moed, The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature, 2012, ResearchTrends, http://www.researchtrends.com
[3] MIKE 2.0, Big Data Definition, http://mike2.openmethodology.org/wiki/Big_Data_Definition
[4] P. Zikipoulos, T. Deutsch, D. Deroos, Harness the Power of Big Data, 2012, http://www.ibmbigdatahub.com/blog/harness-power-big-data-book-excerpt
[5] Gartner, Big Data Definition, http://www.gartner.com/it-glossary/big-data/
[6]E. Dumhill, “What is big data?”, 2012 ,http://strata.oreilly.com/2012/01/what-is-big-data.html
[7] A Navint Partners White Paper, “Why is BIG Data Important?” May 2012, http://www.navint.com/images/Big.Data.pdf
[8] Greenplum. A unified engine for RDBMS and Map Reduce, 2009. http://www.greenplum.com/resources/mapreduce/.

[9] For Big Data Analytics There's No Such Thing as Too Big The Compelling Economics and Technology of Big Data Computing, White Paper, March 2012, By: 4syth.com, Emerging big data thought leaders
[10] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers." Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute. May 2011
[11] Oracle Information Architecture: An Architect's Guide to Big Data, An Oracle White Paper in Enterprise Architecture August 2012
[12] Alvaro Ortigosa, José M. Martín, Rosa M. Carro "Sentiment analysis in Facebook and its application to e-learning ",Department of Computer Science, Universidad Autónoma de Madrid, Francisco Tomás y Valiente 11, 28049 Madrid, Spain Computers in Human Behaviour 31 (2014) 527–541.
[13]Georgios Nomikos "Stumbl: Using Facebook to collect rich datasets for opportunistic networking research" May 2010 – October 201 M.Sc. Thesis vol-2 pp-38.
[14 Vijay B. Raut et al, "] Survey on Opinion Mining and Summarization of User Reviews on Web", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1026-1030.
[15] Sentiment Analysis on Big Data Machine Learning Approach span white paper.
[16] Sunil B. Mane et al, "Real Time Sentiment Analysis of Twitter Data Using Hadoop" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 – 3100 3098
[17] Michael Jakl "REST Representational State Transfer", mj@int-x.org 0226072 – 033-534 University of Technology Vienna
[18]Bingwei Liu, "Scalable Sentiment Classification for Big Data Analysis Using Na¨ıve Bayes Classifier" gcheng@intfusion tech.com, erik.blasch.1 @u.af.mil vol.3, 2013 (pp. 519–528).
[19]Ana Mihanović, Hrvoje Gabelica, "Big Data and Sentiment Analysis using KNIME: Online Reviews vs. Social Media" inteligencija.com vol. 2, No. 1–2, 2008, pp 1-135.
[20] Mo Karimkhan, Jitendra B. Bhatia "Sentiment Analysis and Big Data Processing" IJCSC Volume 5 • Number 1 March-Sep 2014 pp. 136-142.
[21] A. Srivastava et. al., "An Approach for Personalization on Web Based Systems" International Journal of Research and Development in Applied Science and Engineering, Volume 5, Issue 1, March 2014.