International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Volume 7 Issue 4, April-2018, Impact Factor: 3.578

Big Data Analytics: A Review Paper

Monika Anand[1], Meenu Kinra Nangia[2]

[1,2]Astt. Prof., Hindu Institute of Management &Technology, Rohtak

ABSTRACT

A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. The basic objective of this paper is to explore the potential impact of big data challenges, open research issues, and various tools associated with it. As a result, this article provides a platform to explore big data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues.

Keywords: Big data analytics; Hadoop; Massive data; Structured data; Unstructured Data

1. INTRODUCTION

Imagine a world without data storage; a place where every detail about a person or organization, every transaction performed, or every aspect which can be documented is lost directly after use. Organizations would thus lose the ability to extract valuable information and knowledge, perform detailed analyses, as well as provide new opportunities and advantages. Anything ranging from customer names and addresses, to products available, to purchases made, to employees hired, etc. has become essential for day-to-day continuity. Data is the building block upon which any organization thrives. Now think of the extent of details and the surge of data and information provided nowadays through the advancements in technologies and the internet. With the increase in storage capabilities and methods of data collection, huge amounts of data have become easily available. Every second, more and more data is being created and needs to be stored and analyzed in order to extract value. Furthermore, data has become cheaper to store, so organizations need to get as much value as possible from the huge amounts of stored data. The size, variety, and rapid change of such data require a new type of big data analytics, as well as different storage and analysis methods. Such sheer amounts of big data need to be properly analyzed, and pertaining information should be extracted.

2. LITERATURE SURVEY

Over the last many years, there are many researchers has completed their work successfully on big data. Hundreds of articles have appeared in the general business press (For example Forbes, Fortune, Bloomberg, Business week, The Wall street journal, The Economist). National Institute of Standards and Technology [NIST] said that Big Data in which data volume, velocity and data representation ability to perform effective analysis using traditional relational approaches. In March 2012, The Obama Administration announced that the US would invest 200 Million Dollars to launch a big data research plan. An IDC Reports predicts that from 2005 to 2020, the global data volume will grow by a factor of 300, from 130 Exabyte's to 40,000 Exabyte's, representing a double growth every two years. IBM estimates that everyday 2.5 quintillion bytes of data are created out of which 90% of the data in the world today has created in the last two years. It is observed that social networking sites like Facebook have 750 Million users, LinkedIn has 110 million users and Twitter has 250 million users. From industry, government and research community, Big Data has led to an emerging research field that has attracted tremendous interest. The broad interest is first exampled by coverage on both industrial reports and public media for example: The economist, New York Times. Mobile Phones becoming best way to get data on people from different aspect, the huge amount of data that mobile carrier can process to improve our daily life. In figure 1, From Year 2005, it would appear from this graph that the amount of data was practically increased. However, Consider exponential growth in data from 2005 year, when enterprise system and user level data was flooding into data warehouse.

When the capacity of Data Warehouse grew from 50 GB to 1 TB – 100TB. Data was in structured form when it creates from many organizations. Data goes from three properties like volume, Variety and velocity. Many companies were facing the problem on how to expand the capacity of data warehouse to accept the new requirement.

3. BIG DATA

"Big Data can be defined as volumes of data available in varying degrees of complexity, generated at different velocities and varying degrees of ambiguity, that cannot be processed using traditional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions."

Big data is the new term that contains large and complex datasets. It is difficult to manage these datasets without new technology. The Mckinsey Global Institute (MGI) published a report on big data that describes the various business opportunities that big data opens. Paulo Boldi, One of the authors says "Big Data does not need big machines, it needs big intelligence. There are two types of Big Data is as follows:

3.1 Structured Data These data can be easily analyzed. It is in numerical form, figures, and transaction data etc.

3.2 Unstructured Data These data contain complex information such as Email attachments, Images comments on social networking sites. These data cannot be easily analyzed. Doug Lancy was the first one talking about 3v's in big data management: Volume - It describes the amount of data. It refers to mass quantities of data.

Variety - It describes different types of data and sources including structured, semi-structured and unstructured data.

Velocity - It defines the motion of data. Data created rapidly, processed and analyzed.

4. BIG DATA ANALYTICS

The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time . Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before. Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage. In this section, we will start by discussing the characteristics of big data, as well as its importance. Naturally, business benefit can commonly be derived from analyzing larger and more complex data sets that require real time or near-real time capabilities; however, this leads to a need for new data architectures, analytical methods, and tools. Therefore the successive section will elaborate the big data analytics tools and methods, in particular, starting with the big data storage and management, then moving on to the big data analytic processing. It then concludes with some of the various big data analyses which have grown in usage with big data.

4.1 Structured Analytics In structured analytics, large quantity of data is generated from business and scientific research fields. These data is managed by RDBMS, Data warehousing, OLAP and BPM. Data grown by various research area like Privacy preserving data mining, E-commerce.

4.2 Text Analytics In Text analytics, Text is one of the most common forms of storing the information and it includes Email communication, documents, and Social media contents. Text analytics also known as Text mining, refers to the process of extracting useful information from large text. Text mining system is based on text representation and Natural Language Processing (NLP) with emphasis on the latter.

4.3 Web Analytics The aim of Web analytics is to retrieve, extract the information from Web Pages. Web Analytics also called Web mining. 4.4 Multimedia Analytics Recently multimedia data, including images, audio, and video has grown at a tremendous rate.

4.4 Multimedia analytics refers to extract interesting knowledge and semantics captured in multimedia data. Multimedia analytics covers many subjects like Audio Summarization, Multimedia annotation, Multimedia indexing and retrieval. 4.5

Mobile Analytics Mobile data traffic increased 885PBs Per Month at the end of 2012. Vast volume of application and data leads to mobile analytics. Mobile analytics involves RFID, mobile phones, sensors etc.

5. BIG DATA ANALYTICS TECHNIQUES

There are different techniques that are currently in use. In general, the techniques can be summed up to:

1) Association Rule Learning
2) Classification tree analysis
3) Genetic algorithms
4) Machine learning
5) Regression analysis
6) Sentimental Analysis
7) Social network analysis.

All classification techniques have come to inspire secondary definition types of forms of analytics like text analytics, and media analytics. The result for the categorization and classification of big data is that there is a variety of different manageable volumes of data can be transfer between persons at a high velocity.

5.1 Association Rule Learning This is the classification learning technique. It basically comprises of analysis of the data presented and finding relations between data presented. The result is categorization of data with similar characteristics together. It has been used in different spheres of life. For example, the use of association rule learning can be used in text analytics. Websites that depend on user frequency to determine their frequency of users on the site and hence the productivity of a particular site over another.

5.2 Classification Tree Analysis Classification tree analysis is the best way in which different text data can be analyzed. Text analytics can also manifest itself in the form of classification tree analysis. Large historical data can be classified chronologically in through classification tree analysis.

5.3 Genetic Algorithms Genetic algorithms are techniques that is used to identify the most possibly viewed videos, TV shows and other forms of media. There is an evolutionary pattern that can be identified by genetic algorithms. Video and media analytics can be done by the use of genetic algorithms.

5.4 Machine Learning Machine learning is another technique that can be used to categories and determine the probable outcome of a specific set of data. Machine learning defines software that can be able to determine the possible outcomes of a certain set of event. It is therefore used in predictive analytics. An example of predictive analytics is probability of winning legal cases or the success of certain productions (Watson, 2014).

5.5 Regression Analysis This is a technique that takes the use of independent variables and how they affect dependent variables. This can be a very useful technique in determining social media analytics like the probability of finding love over an internet platform (Ratner & Ratner, 2011).

5.6 Sentiment Analysis This is the ultimate technique that is used that is used in text analytics. It looks at the actual sentiments of different people and then cross references them with the experience that is described in the text or audio response. Sentiment analysis is a categorization technique that is text based but can have applications in audio analytics.

5.7 Social Network Analysis In today's world, social media has become one of the most important tools in today's communications. The principle majorly analyzes the different relationships that can be identified in between the different social interactions. The technique has been widely applied in determining the interpersonal relationships between human beings. Social network analytics is one of the forefront techniques that can be used to determine the influence of an individual amongst others.

The analysis of such kind of data can be very beneficial to the different parts of social interaction (Stimmel, 2015) Hadoop and Map/reduce frameworks are platforms upon which the documentation of the finding that big data analytics (Perera & Gunarathne, 2014). It is a software that used in the distribution and large processing of the different sizes and categories of big data.

6. APPLICATION OF BIG DATA ANALYTICS

There are several applications of big data analytics. The first and most evident applications are in business. Through business analytics, within big data, patterns in business can be identified so that the different niches in business are found can be maximized upon Big data analytics can also be used in the analysis of large text that is transferred over the internet. Security intelligence is one of the most important tools that any government looks into when it comes to data analytics. It can therefore be used in the different aspects of data analytics

CONCLUSION

Big data analytics has been one of the most important breakthroughs in the information technology industry. The growth of the data that is being transferred the Information Communication Industry is getting to a point where it is becoming unmanageable. The use of big data analytics and extended storage spaces, like the Cloud has made it easier to manage the amount of data is processed in the internet. However, big data analytics cannot be the solution to all of the different problems are present due to the lack of storage space. Compression should be incorporated in all analytic techniques so that the information that is realized at the end of analytic processing is reduced to a manageable size. Introduction of compression engines and techniques can improve the quality of information that is realized at the end of the analytics process.

REFERENCES

[1]. Sameera Siddiqui, Deepa Gupta," Big Data Process and Analytics : A Survey", International Journal Of Emerging Research in Management & Technology, ISSN: 2278-9359, Volume 3, Issue 7, July 2014.
[2]. Bharti Thakur, Manish Mann," Data mining for big data: A Review", International journal of advanced Research in Computer Science and Software Engineering, ISSN: 2277 128x, Volume 4, Issue 5, May 2014.
[3]. Anand V. Saurkar, Vaibhav Bhujade, Priti Bhagat and Amit Khaparde," A Review Paper on Various Data Mining Techniques", International Journal of Advanced Research in Computer Science and software
[4]. Puneet Singh Duggal, Sanchita Paul, "Big Data Analysis: Challenges and Solutions", International Conference On Cloud, Big Data and Trust 2013, Nov 2013
[5]. Albert Bifet, "Mining Big Data in Real Time", informatica, 2013.
[6]. Stephen Kaisler, Frank Armour, J. Alberto Espinosa and William Money," Big Data: Issues and Challenges Moving Forward", Hawaii International Conference on System Science, IEEE Computer Society, Page No. 995, 2013.
[7]. D.Fisher, R.Deline, M.Czerwinski and S. Drucker,"Interaction with big data analytics", Volume 19, No.3, May 2012.
[8]. J.Gantz, D. Reinset," The Digital Universe in 2020: Big Data, Bigger digital shadow, and biggest growth in the far east", in Proc: IDC iview, IDC Anal, Future, 2012.
[9]. Denis Guyadeen , Rob Peglar," Introduction to Analytics and Big data- Hadoop", SNIA Education Committee, 2012.
[10]. Neil Raden,"Big Data Analytics Architecture", Hired Brains Inc, 2012
[11]. James Manyika, Michael Chui, Brad Brown, Jacques Bhuhin, Richard Dobbs, Charles Roxburgh, Angela Hungh Byers, "Big Data: The next frontier for innovation, competition and productivity", June 2011.
[12]. Wei Fan, Albert Bifet, "Mining Big Data: Current Status and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2.