IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Big Data Analytics in Data Mining – A Review

[1]Kinnari Mishra, [2] Hetal Bhaidasna

[1] Assistant Professor in Computer Engineering Department, PIET-DS, Parul University, Vadodara, Gujarat.

[2]Assistant Professor and Head of Computer Department, PIET-DS, Parul University, Vadodara, Gujarat.

Abstract

The Age of Big Data are already running because the traditional data analytics may not be able to handle such large quantities of data. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle the extract value and knowledge from these data sets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper presents brief introduction of data analytics, architecture of big data, knowledge discovery and big data algorithms.

Keywords: Big Data, Data Mining, Analytics, Analysis, Big data categories, Algorithm, Knowledge Discovery.

INTRODUCTION

As the information technology spreads fast, most of the data were born digital as well as exchanged on internet today.

On social media, billions of users connect daily, users share information, upload images, videos and many more. Big data refers to massive amounts of data produced by different sources like social media platforms, weblogs, sensors, IoT devices and many more. Structured data are predefined data which already define or we can say labeled data. Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model. Semi Structured is a combination of structured and unstructured data. It can be like Structured (eg. DBMS), Semi- structured (eg. XML file), Unstructured (eg. Images, Audios and Videos).

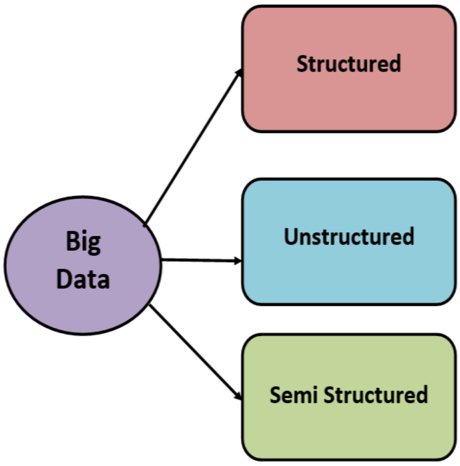

Figure 1: Big Data Categories

Each of these categories has its own characteristics and complexities as described in Figure 1. Data sources include internet data, sensing and all stores of transnational information, ranges from unstructured to highly structured are stored in various formats. [12] Most popular is the relational database that comes in a large number of varieties. As theresult of the wide variety of data sources, the captured data differ in size with respect to redundancy, consistency andnoise, etc.

Although the advances of computer systems and internet technologies have witnessed the development of computing hardware following the Moore’s law for several decades, the problems of handling the large-scale data still exist when we are entering the age of big data.

Big data initiatives were rated as “extremely important” to 93% of companies. Leveraging in big data Analytics solution helps organizations to unlock the strategic values and take full advantage of their assets.

A numerous researches are therefore focusing on developing effective technologies to analyze the big data. To discuss in deep the big data analytics, this paper gives not only a systematic description of traditional large-scale data analytics but also a detailed discussion about the differences between data and big data analytics framework for the data scientists orresearchers to focus on the big data analytics.

As a result, this paper is aimed at providing a brief review for the researchers on the data miningand distributed computing domains to have a basic idea to use or develop data analytics for big data.

DISTINCTNESS AND IMPORTANCE OF BIG DATA

Big data is a term utilized to refer to the increase in the volume of data that are difficult to store, process and analyze through traditional database technologies. The nature of big data is indistinct and involves considerable processes to identify and translate the data into new insight. The tern “big data” is relatively new in IT and business. However, several researchers and practitioners have utilized the term in previous literature.

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

Big data are important because it is cost savings, time savings, understand the market conditions, social media listening, boost customer acquisition and retention, and solve advertiser’s problem and offer marketing insights, the driver of innovations and product development.

CLASSIFICATION AND ARCHITECTURE OF BIG DATA

The classification is important because of large-scale data in the cloud. The classification is based on data type, content format, data source, data consumer, data usage, analysis type, processing purpose, processing method, data store, data frequency, data aggregation. [11]

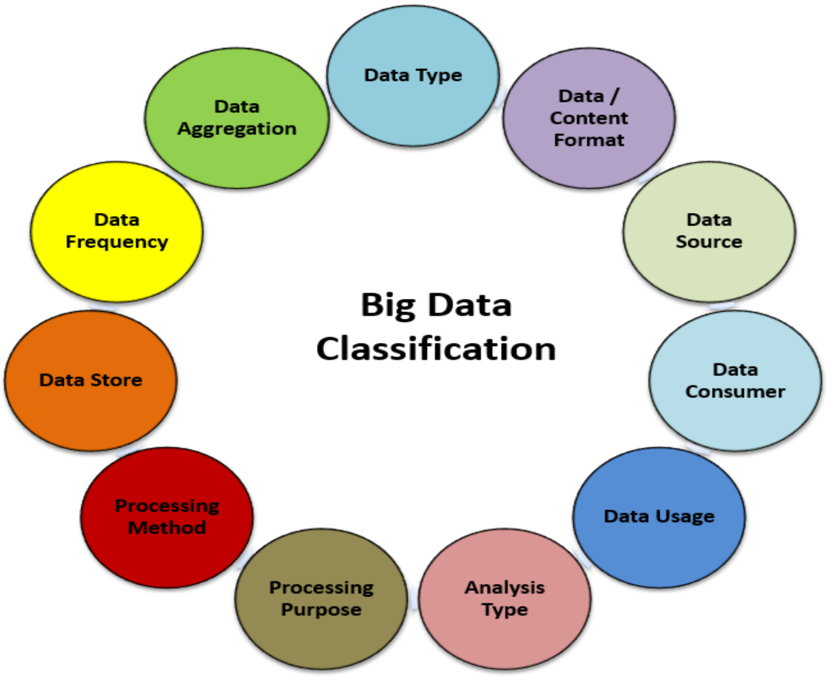

Figure: 2 Big Data Classification

BIG DATA ANALYTICS with DATA MINING

Nowadays, the data that need to be analyzed are not just large,but they are composed of various data types, and even including streaming data. [1]Since big data has the unique features of "massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous," which may change the statistical and data analysis approaches Although it seems that big data makes it possible for us to collect more data to find more useful information, the truth is that more data do not necessarily mean more useful.[5]

The big data may be created by handheld device, social network, and internet of things, multimedia, and many other new applications that all have the characteristics of volume, velocity, and variety. [3]

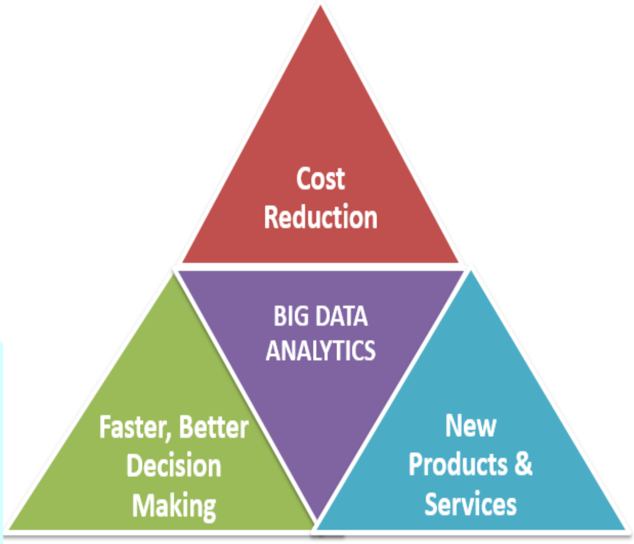

Figure: 3 Big Data Analytics

From the volume perspective, the deluge of input data is the very first thing that we need to face because it may paralyze the data analytics. [8] Different from traditional data analytics, for the wireless sensor network data analysis, pointed out that the bottleneck of big data analytics will be shifted from sensor to processing, communications, storage of sensing data. This is because sensors can gather much more data, but when uploading such large data to upper layer system, it may create bottlenecks everywhere.

In addition, from the velocity perspective, real-time or streaming data bring up the problem of large quantity of data coming into the data analytics withina short duration but the device and system may notbe able to handle these input data. This situation is similar to that of the network flow analysis for whichwe typically cannot mirror and analyze everything we can gather.

From the variety perspective, because the incoming data may use different types or have incomplete data, how to handle them also bring up another issue for the input operators of data analytics.

To make the whole process of Knowledge Discovery in Databases (KDD) more clear, Fayyad and his colleagues summarized the KDD process by a few operations in, which are selection, preprocessing, transformation, data mining, and interpretation/evaluation.

As shown in figure 3, with these operators at hand we will be able to build a complete data analytics system to gather data first and then find information from the data and display the knowledge to the user.[7] According to our observation, the number of research articles and technical reports that focus on data mining is typically more than the number focusing on other operators, but it does not mean that the other operators of KDD are unimportant.

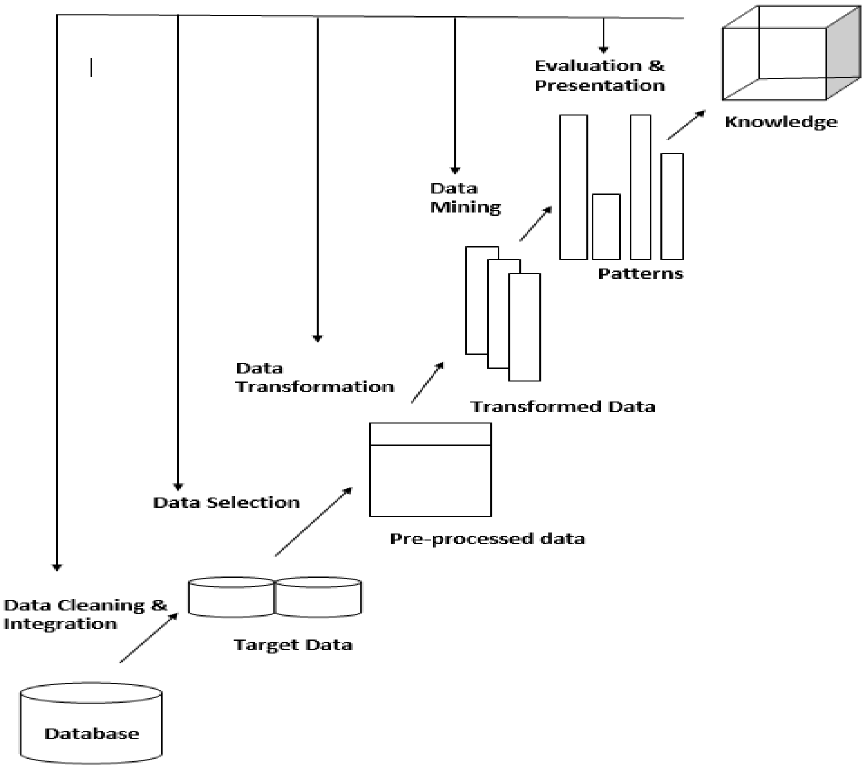

Figure 3: The process of knowledge Discovery in Database

Data mining is the process of finding complex data, patterns and similarity within large data sets to predict results. Using a wide range of methods, you can use this information to increase knowledge from the data. It is necessary to understand the data being processed, in order to ensure meaningful data mining results. Data mining approaches are generally affected by many factors, such as noisy data that include null values using preprocessing. For data mining different type of techniques are used like classification, regression, for Predictive data mining and Association Rules mining, clustering, Rough Sets Analysis for descriptive data mining.

DATA MINING TOOLS

The development of data mining algorithms requires the use of powerful software tools. As the number of available tools continues to grow, the choice of the most suitable tool becomes increasingly difficult. The tools for data mining categorization based on different user groups, data structures, data mining tasks and methods, visualization and interaction styles, import and export options for data and models, platforms, and license policies. Every tool has its own advantages and disadvantages.

Data mining tools predict behaviors, future trends, allowing business to make proactive, knowledge driven decisions. There are number of open source tools available for data mining like,

- Rapid Miner. Availability: Open source.
- Orange. Availability: Open source.
- Weka. Availability: Free software.
- KNIME. Availability: Open Source.
- Apache Mahout.
- Oracle Data Mining.

SECURITY ISSUES

Since much more environment data and human behavior will be gathered to the big data analytics, how to protect them will also be an open issue because without a security way to handle the collected data, the big data analytics cannot be a reliable system.
In spite of the security that we have to tighten for big data analytics before it can gather more data from everywhere, the fact is that until now, there are still not many studies focusing on the security issues of the big data analytics. According to our observation, the security issues of big data analytics can be divided into fourfold: input, data analysis, output, and communication with other systems.

The input, it can be regarded as the data gathering which is relevant to the sensor, the handheld devices, and even the devices of internet of things. One of the important security issues on the input part of big data analytics is to make sure that the sensors will not be compromised by the attacks. For the analysis and input, it can be regarded as the security problem of such a system. For communication with other system, the security problem is on the communications between big data analytics and other external systems. Because of these latent problems, security has become one of the open issues of big data analytics.

CONCLUSION

The availability of Big Data, low-cost commodity hardware, and new information management and analytic software have produced a unique moment in the history of data analysis. The convergence of these trends means that we have the capabilities required to analyze astonishing data sets quickly and cost-effectively for the first time in history. These capabilities are neither theoretical nor trivial. They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue, and profitability.

The Age of Big Data is here, and these are truly revolutionary times if both business and technology professionals continue to work together and deliver on the promise.

Big data analytics is trying to take advantage of the excess of information to use it productively. It must support and encourage fundamental research towardsaddressing these technical challenges if we are to achieve the promised benefits of big data.

REFERENCES

[1]. Lyman P, Varian H. How much information 2003? Tech. Rep, 2004. [Online]. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/ printable_ report. pdf.

[2]. Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009.

[3]. Ding C, He X. K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on machine learning; 2004. pp. 1–9.

[4]. Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Trans Knowl Data Eng. 2003;15 (5):1170–87.

[5]. Fisher D, DeLine R, Czerwinski M, Drucker S. Interactions with big data analytics. Interactions. 2012;19(3):50–9.

[6]. Apache Hadoop, February 2, 2015. [Online]. http://hadoop.apache.org.

[7]. Cuda, February 2, 2015. [Online]. http://www.nvidia.com/object /cuda_home_new . html.

[8]. Apache Storm, February 2, 2015. [Online]. http://storm.apache.org/.

[9]. Apache Mahout, February 2, 2015. [Online]. http://mahout.apache.org/.

[10]. L. Neumeyer, B. Robbins, A. Nair, A. Kesari, S4: Distributed Stream Computing Platform, Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, 2010, pp. 170–177.

[11] https://www.journals.elsevier.com

[12] https://journalofbigdata.springeropen.com