ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Survey on Big Data: Challenges, Tools and Technique

Buta Singh
Assistant Professor
Guru Nanak College
Budhlada, Punjab, India
er.bootasidhu138@gmail.com

Satish Kumar
Assistant Professor
Guru Nanak College
Budhlada, Punjab, India
satishahuja06@gmail.com

Gagandeep Kaur
Assistant Professor
Guru Nanak College
Budhlada, Punjab, India
gagan.manshahia@gmail.com

Mehakdeep Kaur
Lab Instructor
Guru Nanak College Budhlada, Punjab,India
mehakd3@gmail.com

Abstract: *The term "Big Data" refers to data sets that are not only big but also high in volume, variety and velocity which make them very difficult to handle by traditional data processing techniques. The three Vs (volume, variety and velocity) are responsible to create data sets that grow so large that it becomes awkward to work with data sets using traditional data management techniques. Here, the term volume of the data is its size and how large it is. Velocity refers to the rate with which data is changing and variety includes different formats and types of data as well as different types of uses and ways to analyzing the data. Big data is a data whose scale, distribution and timeliness require new technical architectures and tools in order to handle and extract useful knowledge from these datasets. This useful knowledge adds new insights into the business for decision makers. So, big data analysis is required to gain valuable insights from these large and changing datasets. This paper presents various tools and techniques used to analyze the big data along with various challenges that may be faced while analyzing the big data.*

Keywords: Big Data, Data Mining, decision making**.**

1. INTRODUCTION

Data is the building block upon which any organization survives. A world without data storage is a place where every detail about a person or organization, every transaction performed, or every aspect which can be lost directly after use. Organizations would thus lose the ability to extract valuable information and knowledge, as well as provide new opportunities and benefits. Nowadays through the advancements in technologies and the internet more and more data is being created. With the increase in storage capabilities and methods of data collection, huge amounts of data have become easily available. Furthermore, data has become cheaper to store, so organizations need to get as much value as possible from the huge amounts of stored data. Hence, the challenge is that these large data sets need to be stored and analyzed in order to extract value. 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 so that pertaining information can be extracted.

1.1 Big Data Analytics

The Term "Big Data" has been applied to data sets that are holding large amount of data whose size is beyond the ability of commonly used data processing tools. Today organizations are exploring large volume of data so as to discover those important facts which are never discovered before. Big data size is gradually increasing currently from few dozen terabytes to many petabytes of data in a single data set as discussed in [1][2]. Hence the key challenge is to capture, store, manage and process the big data within a tolerable time interval. Hence, big data analytics requires advanced analytical tools and techniques to manage and process these data sets. Advanced analytical techniques contribute to discover important facts and future decisions which leverage business change.

2. BIG DATA CHALLENGES

Opportunities are invented from challenges .No doubt, big data has created attractive and new opportunities; we are also facing a lot of challenges when handling big data. Major difficulties lie in data collection and storage, Data inconsistence and incompleteness, timeliness, analysis data visualization and data security. If we surpass these major challenges then big data will become a gold ore. But till today, the state of art tools and techniques can exploit these gold ores of big data. The brief discussion of each of these challenges is explained in the following sub sections.

2.1. Data Collection and Storage:

Data sets are growing in size since billion of data is created every day. There is a bulk storage requirement of databases, array storage and to store large output files. So the major challenge is the requirement of more storage mediums and comparatively higher input output speed .Data collection means to capture the valuable and related data in a structured manner from different data sets. The accessibility of big data is a knowledge discovery process. Big data should be collected and stored in such a way that it should be accessed easily and may be used for further analysis.

2.2 Data Inconsistence and Incompleteness:

Big data is a huge repository of various types of data sets consisting of structured, semi-structured and unstructured data. The most important challenge is that data should be complete and consistence from every view of accessibility. Large amount of data may be incomplete or unrelated and chances of inconsistency are more. So, powerful analysis is required to make it complete and consistence.

2.3 Timeliness:

Timeliness means data should be available at the right time i.e. when it is required. The capacity to store information has doubled after every 3 years since 1980s [1]. But hard disks are quite slower in case of input output speed performance. So, major challenge is to implement optimizing data access techniques, schema free databases to quickly modify the structure of data so that it do not need to rewrite tables. Data searching policy has to be improved.

2.4 Data Analysis:

Data analysis is a key challenge of big data. Various traditional techniques are invented but these cannot exploit the gold ores of big data. To capture the useful patterns of big data we need to develop extraordinary techniques. Besides it, we need extraordinary tools to make some sense of data. Most popular Apache Hadoop architecture has provided some tools to meet real time data requirements. Data analysis is a big challenge and it needs some state of art and extraordinary tools to meet this challenge.

2.5 Data Visualization:-

The main motive of visualization is to discover the hidden and complex knowledge in an understandable and interactive form. Current big data tools and techniques have poor response regarding data visualization. Data may be visualized in the form images, text, diagrams, graphics and many other interactive ways which is a great matter of concern as discussed in [1].

2.6 Data Security:-

Data security has a great attention in field of information technology. Big data security is also a key challenge. The size of big data is very large and it is stored in different data sets on remote computers in a distributed way. Hence, the network threat may create security issues. The personal data privacy protection is also a necessary challenge regarding big data security.

3. BIG DATA ARCHITECTURAL FRAMEWORK

The conceptual framework of big data analytics project has a pooled data. As shown in Figure 1, data may be collected from various sources in various formats. This form of data is called raw data. In the second step, data needs to be transformed via the steps of extract, transform and load (ETL) [2]. Another approach is to use a data warehouse where related data from various sources is aggregated. In the next part, Big data tools are used for analyzing the transformed data. Big data is analyzed by using many states of art big data techniques and technologies and useful knowledge is discovered as a final result.

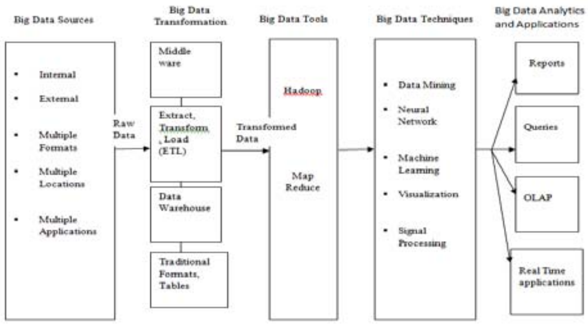

Figure1. Conceptual Architecture of Big Data Analytics [3]

4. BIG DATA TOOLS: HADOOP AND MAP REDUCE

The most significant and open source platform for big data analytics is called as Hadoop (Apache platform). Most of the big data systems are built on Hadoop. Hadoop is a powerful batch processing tool used to perform distributed data processing. Hadoop has the capability to process large amounts of data mainly by allocating all the partitioned data sets to different nodes and each node solves a sub part of the problem and then combines all sub parts to obtain final result. Hence Hadoop plays double role of data organizer and analytic tool. Hadoop along with the addition of map reduce framework works a powerful tool. Apache Hadoop platform is a combination of Hadoop kernel, map.reduce and Hadoop distributed file system (HDFS). HDFS enables the storage area of Hadoop cluster. It divides the larger problem into sub parts and distributes it across the various nodes. On the other hand, map reduce is a programming model based on divide and conquer approach. Map reduce provides an interface to distribute sub tasks and collects the output of each task to generate final output. Map reduce keeps track on processing of each node when tasks are executed. If any node fails it tries to find out the reason or it transfers the work of failed node to any other node. The Map reduce is implemented on Hadoop. There are two types of nodes in Hadoop architecture-the master nodes and worker nodes. The master node collects the input and distribute to all the worker nodes in map step. After this, the master node captures all the sub results and combine them to form the final result in reduce step. The master node in map reduce framework is called job tracker and all the slave nodes are called task trackers. The master node is responsible for job scheduling, fault tolerance and distribution of subtasks to all the slaves as shown in figure 2

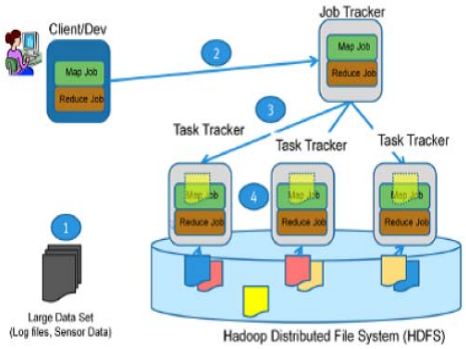

Figure2. Hadoop and Map Reduce platform [2]

BIG DATA TECHNIQUES

Big data analysis needs some extra-ordinary techniques to explore useful patterns from large volume of data. Each technique is driven by specified applications. For example a Chinese company like eBay [1] uses data mining techniques to browse large amount of user's data recorded on its website and its exploits a useful deal of valuable information for decision making. Big data techniques include data mining, machine learning, neural networks, signal processing and visualization approaches.

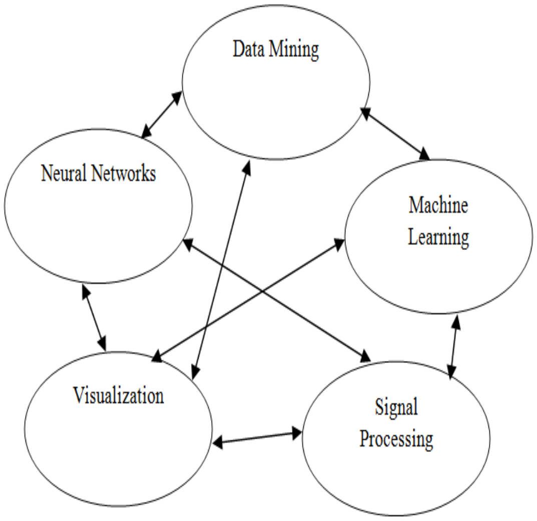

Figure3. Data Mining Techniques [1].

4.1 Data Mining Technique:

Data mining involves a set of techniques to extract useful patterns from large amount data. Data mining is a process which involves a number of methods such as clustering analysis, classification, regression and association rule learning. Big data mining is more challenging and useful technology as compared with traditional data mining algorithms. Classification is mostly applied data mining technique. Classification has two phases learning and classification. In the learning phase pre-classified data (training data) is analyzed by classification algorithm. In the classification phase test data is used to estimate the accuracy of classification rules. If the accuracy is acceptable then rules can be applied to new data. Fraud detection applications are well suited for this type of analysis. Clustering technique identifies a group of similar classes of objects. It is used to find correlations among data attributes. For example, to form a group of members based on their purchasing patterns based on some similarity. Regression technique is used to predict the relationship among one or more independent variables and dependent variables. In data mining, the variables already known are called independent variables and variables which we want to predict are called response variables. Association rule and correlation is used to find frequent decisions such as cross marketing and customer shopping behavior analysis. Association algorithms are used to generate rules with confidence value less than one. Association rule is used to find frequent item set among large data sets.

5.2 Neural Networks:

Neural networks have a remarkable ability to extract complex patterns and detect trends that are to complex to derive by either human beings or computer oriented techniques. For example, handwritten character recognition so that a computer can pronounce English text has already been implemented successfully using neural networks in many industries. Neural network is a network of continuous valued inputs and outputs i.e. each intermediate connection has some weight present on it. During the learning phase network learns by adjusting weights to predict the correct class labels of the input tuples. All complex patterns are extracted by neural networks.

5.3 Machine Learning:

Machine learning is an important application of artificial intelligence which allows computers to explore behaviors on the basis of empirical data. The algorithms designed for machine learning have an obvious characteristic to discover knowledge and make intelligent decisions automatically. There are many machine learning algorithms available for both supervised learning and unsupervised learning to cope with big data analytics. But all are facing the scalability problems. For example support vector machine (SVM) is a fundamental algorithm used for classification and regression analysis but SVM is also facing scalability problems like memory requirements and computational time [1]. Map -reduce big data tool has the capability to scale machine learning so that valuable knowledge can be learned.

5.4 Visualization Technique:

Visualization approaches are used to display the data in the form of tables, images, diagrams and other interactive display methods. Big data visualization is a very challenging task as data sets are growing rapidly. When we think about large scale data visualization, many researchers use different techniques such as feature extraction to reduce the size of data before displaying the actual data. Recently, social media analysis has become popular. Social network analysis (SNA) has grown as a key technology in modern era. But the main obstacle regarding SNA is the growing size of big data. The state of art techniques for visualization high dimensional data is still a prime demand.

5.5 Signal Processing Technique:-

Digital signal processing (DSP) is particularly motivated by the need to extend traditional signal processing techniques. DSP methods are applied to those data sets having complex and irregular structure. Patterns from different data sets are formulated and solved as standard signal processing problem. DSP involves data compression and decompression through Fourier and inverse Fourier transformations, recovery, denoising and classification of data by signal regualarization, anomaly detection by high pass filtering. A group signal is used to carry out this process.

6. CONCLUSION

This paper is a review study of some useful tools, techniques and challenges of big data. Each technique serves a specific application. Knowledge discovery from big data is a vast process and every individual technique plays a major role in knowledge discovery in big data analytics. The overall conclusion is that these techniques are helpful but not sufficient in bridging the gap between big data and knowledge discovery. There is still a need to develop some extraordinary techniques regarding big data analytics to meet real time data analysis problems.

REFERENCES

[1] C.L. Philip Chen, Chun-Yang Zhang "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data" (Elsevier 2014).

[2] Nada Elgendy and Ahmed Elragal " Big Data Analytics: A Literature Review Paper" (Springer International Publishing Switzerland 2014).

[3] Wullianallur Raghupathi and Viju Raghupathi " Big data analytics in healthcare: promise and potential" (Health Information Science and Systems 2014).

[4] Jafar Raza Alam1, Asma Sajid2, Ramzan Talib3, Muneeb Niaz4 " A Review on the Role of Big Data in Business" (International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 3, Issue. 4, April 2014, pg.446 – 453).

[5] W.-H. Weng and W.-T. Lin, "A Scenario Analysis Of Big Data Technology Portfolio Planning," in International Journal of Engineering Research and Technology, 2013.

[6] A. Bifet, "Mining Big Data in Real Time," Informatica (03505596), vol. 37, 2013

[7] U. G. Pulse, "Big Data for development: challenges & opportunities," Naciones Unidas, Nueva York, mayo, 2012.

[8] “Top ten big data security and privacy challenges,” Cloud Security Alliance White paper,2012.

[9] U. Rasheed, M. U. Sarwar, and R. Talib, "A Review on Data Warehouse Management."

[10] G. Halevi and H. Moed, "The evolution of big data as a research and scientific topic: overview of the literature," Research Trends, Special Issue on Big Data, vol. 30, pp. 3-6, 2012.

[11] “The Compelling Economics and Technology of Big Data Computing, For Big Data Analytics There’s No Such Thing as Too Big.” 4syth White Paper,2012.

[12] “Mining big data in the enterprise for better business inteligence.”Intel White Paper,2012.

978-93-85670-72-5 © 2016 (RTCSIT)

[13] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," Knowledge and Data Engineering, IEEE Transactions on, vol. 26, pp. 97-107, 2014.

[14] A. Bifet, "Mining Big Data in Real Time," Informatica (03505596), vol. 37, 2013.

978-93-85670-72-5 © 2016 (RTCSIT)