International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 9, ISSN No. 2455-2143, Pages 26-31
Published Online July – August 2016 in IJEAST (http://www.ijeast.com)

BIG DATA MINING AND TOOLS: A REVIEW

Sangeeta Bhagat
M.Tech CSE
(Big Data Analytics)
Pit Kapurthala
(Main campus PTU)
India

Mansi Gupta
M.Tech CSE
(Assistant Professor)
Pit Kapurthala
(Main campus PTU)
India

Kanchan Rni
M.Tech CSE
(Big Data Analytics)
Pit Kapurthala
(Main campus PTU)
India

Kanika Dhanjal
M.Tech CSE
(Big Data Analytics)
Pit Kapurthala
(Main campus PTU)
India

***Abstract*— Big data term can be used for large amount of data, generated from various different resources. With the difficulty of storing and processing.. Big data is now expanding in all area of science and engineering including, physical, biomedical and biological sciences. Big Data mining is the capability of extracting useful information from large datasets or streams of data, that due to its volume, veracity, and velocity, it was not possible before to do it. In this paper we will characterize the Big Data processing with the view of data mining. The Big Data challenge is becoming one of the most exciting deal for the next years. This paper presents an overview of Big data, challenges and tools for processing this big data's content. This paper also introduces the HACE theorem that characterizes the Big Data rising and represents the Big Data Processing Model. This data-driven model has lot of information resources, analysis and mining, user interest and privacy and security.**

Keywords— **Big Data Processing, Big Data Mining, HACE Theorem.**

I. INTRODUCTION

Today's digital world there petabyte data is generated from various social media like Facebook, Twitter and from e-commerce sites like Amazon, Flipkart which is beyond the capability of traditional tools to store and to process. And this data is growing day by day in massive amount. Big data and its processing is center issue of science and all business industries. Everyday 2.5 quintillion bytes of data are created and 90% of data in the world today were produced within the past two years.

Let us take an example Flicker which is a public picture sharing site.1.8 million photos receive by it per day. Indeed as an old saying "a picture is worth a thousand words." These pictures actually explore the human society interest, public affairs, events, disasters and so on, only if we have the ability to harness massive amount of data. This example shows the demonstrate rise of Big Data applications. There is need to record and process this vast amount of data. This process of recording and processing real time data is known as big data processing.

Information from various heterogeneous sources is growing at staggering rate. In 2012, the number of internet users reached 2.27 billion. Everyday Facebook generates more than 25 Terabyte of log data; Twitter generates more than 14 Terabyte of tweets. Add to this quantity the data generated by hundreds of millions of GPS devices sold every year and more than 30 million networked sensors currently in use. The volume of these data is expected to double every two years over the next decade.

The rest of the paper is organized as follows. Various challenges, issues and knowledge discovery from big data is presented in Section II. Section III presents Hace theorem and section IV presents the data mining challenges with big data and section V describes various open source tools. Concluding remarks are given in section IV.

II. CHALLENGES OF BIG DATA

To explore large volume of data and extracting the intelligent information from it for future actions is the biggest challenge for Big Data applications. Because capturing all observations in real time is almost infeasible. The various challenges can be described as follow:

A. **Privacy and Security --**

It is the most imperative issue with big data which is touchy and incorporates calculated, specialized and in addition legitimate hugeness.

- The individual data of a people when consolidated with outside extensive data sets prompts the deduction of new realities about that individual and it's conceivable that these sorts of actualities about the people are cryptic and the people may not need the any person to know about them.
- Information in regards to the client is gathered and utilized as a part of request to enhance the matter of the association. This is finished by making experiences in their lives about which they are unaware.
- Another vital outcome emerging would be Social stratification where a proficient person would take

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 9, ISSN No. 2455-2143, Pages 26-31
Published Online July – August 2016 in IJEAST (http://www.ijeast.com)

favorable circumstances (advantage) of the Big data predictive analysis and on the other hand underprivileged will be effortlessly recognized and treated more regrettable.

- Big Data utilized by law requirement will expand the odds of certain labeled person to experience the ill effects of antagonistic results without the capacity to battle back or even having information that they are being segregated.

B. Data Access and Sharing of Information --

In the event that data is to be utilized to set aside a few minutes it gets to be essential that it ought to be accessible in exact, complete and timely way. This makes the Data administration and governance process bit complex including the need to make data open and make it accessible to government organizations in standardized way with standardized APIs, metadata and configurations accordingly prompting better basic leadership, business knowledge and efficiency upgrades.
Expecting sharing of information between organizations is unbalanced on account of the need to get an edge in business. Sharing information about their customers and operations debilitates the way of life of mystery and intensity.

C. Storage and Processing Issues --

The storage accessible is insufficient for putting away the huge amount of information which is being delivered by practically everything: Social networking sites are themselves an extraordinary supporter along with the sensor gadgets and so on.
Due to the thorough requests of the Big information on systems, stockpiling and servers outsourcing the information to cloud may appear an alternative. Transferring this extensive measure of information in cloud doesn't solve the issue. Since Big data bits of knowledge require getting every one of the information gathered and afterward connecting it in an approach to remove critical data. Terabytes of data will take extensive measure of time to get transferred in cloud and in addition this information is changing so quickly which will make this data difficult to be transferred progressively. In the meantime, the cloud's distributed nature is additionally problem for Big data analysis. In this manner the cloud issues with Big Data can be sorted into Capacity and Performance issues.
The transportation of information from storage point to processing point can be stayed away from in two ways. One is to handle in the storage place put just and results can be exchanged or transport just that data to calculation which is essential. In any case, both these strategies would require trustworthiness and provenance of information to be kept up. Handling of such vast measure of information likewise takes huge measure of time. To discover appropriate components entire of data Set should be scanned which is to some degree unrealistic.

D. Technical challenges --

1) Fault Tolerance: With the approaching of new advancements like Cloud computing and Big data it is constantly expected that at whatever point the failure happens the harm done ought to be inside threshold as opposed to starting the entire task from the scratch. Fault tolerant figuring is to a great degree hard, including perplexing algorithm. It is just unrealistic to devise completely secure, 100% dependable fault tolerant machines or programming. In this way the principle undertaking is to diminish the failure of inability to a "worthy" level. Tragically, the more we Endeavour to decrease this probability, the higher the cost.
Two strategies which appear to build the adaptation to non-critical failure in Big data are as: First is to separate the whole computation being done into errands and allocate these tasks to various hubs for calculation. One hub is assign the work of watching that these hubs are working appropriately. In the case of something happens that specific assignment is restarted.
Be that as it may, now and then it's entirely conceivable that that the entirety calculation can't be separated into such free undertakings. There could be some errands which may be recursive in nature what's more, the contribution of the past assignment is the contribution to the following calculation. Consequently restarting the entire calculation gets to be awkward procedure. This can be evaded by applying Checkpoints which keeps the condition of the framework at certain interims of the time. If there should be an occurrence of any failure, the calculation can restart from last checkpoint kept up.

2) Scalability: The Scalability issue of Big Data has lead towards cloud environment, which now totals numerous unique workloads with differing execution objectives into vast groups. This requires an abnormal state of sharing of assets which is costly furthermore carries with it different difficulties like how to run and execute different jobs with the goal that we can meet the objective of every workload cost adequately. It likewise requires managing the system failures in a proficient way which happens all the more every now and again if working on huge groups(clusters). These elements consolidated put the worry on the best way to express the programs, even complex machine learning undertakings. There has been an immense movement in the innovations being utilized.

Hard Disk Drives (HDD) are being supplanted by the solid state Drives and Phase Change innovation which are not having the same execution amongst consecutive and arbitrary information exchange. In this way what sort of capacity

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 9, ISSN No. 2455-2143, Pages 26-31
Published Online July – August 2016 in IJEAST (http://www.ijeast.com)

gadgets are to be utilized is again a central issue for information stockpiling?

3) Quality of Data: Collection of large amount of data and its capacity includes some significant pitfalls. More information if utilized for basic leadership on the other hand for prescient investigation in business will prompt better results. Business Leaders will dependably need progressively and more information stockpiling (storage) though the IT Leaders will take all specialized perspectives as a top priority before putting away every one of the information. Big Data essentially concentrates on quality information stockpiling as opposed to having large irrelevant information with the goal that better results and conclusions can be drawn.
This further prompts different inquiries like how it can be guaranteed what information is significant, the amount of information would be enough for basic leadership and whether the put away information is precise or not to reach determinations from it and so forth.

4) Heterogeneous Data: Unstructured information speaks to just about each sort of Data being created like social networking connections, to recorded gatherings, to treatment of PDF archives, fax exchanges, to messages and that's only the tip of the iceberg. Structured Data is constantly sorted out into exceptionally motorized and sensible way. It indicates well coordination with database however unstructured information is totally crude and chaotic. Working with unstructured information is lumbering and obviously excessive as well. Changing over this unstructured information into organized one is moreover not achievable. Organized information is the one which is sorted out in a way so that it can be overseen effectively. Burrowing through unstructured information is unwieldy and exorbitant.

III. HACE THEOREM

Big Data means very large volume of data; independent sources of data having decentralized control and it also explore the relationship among the data. These are main attributes of the Big Data. Obtaining the useful information from this massive amount of data is the main challenge of Big Data Mining.
Assume there is an elephant and a number of blind men are sizing up the giant elephant. Each blind man will draw a picture of an elephant according to part of information that he collects during this process but each blind man's view is limited to his local region. Each blind man has different-different information of parts. Some will "feel" it is a rope or a wall or a hose. And then an expert will conclude all men's view in real time fashion. This is very difficult task same as in Big Data where we collect information from various heterogeneous sources and we have to extract the useful information from it [2]**.**

Characteristics of Big Data on The Basis Of HACE Theorem:

3.1 Massive Amount of Data with Heterogeneousness and multiple aspects:

One of the fundamental characteristics of the Big Data is the most of the data represented by heterogeneous and multiple aspects. Every data collector will prefer his local protocols for data collection and each different application of Big Data will result into different data representation for the same kind of data because different organization use their own schemata to represent the information, the data heterogeneousness and multiple aspects issues become major challenges if we are trying to enable data collecting by aggregating data from all different sources.

3.2 Distributed and Autonomous Sources: This is the main issue of Big Data application. Each source will obtain information without depending upon any unit which is centralized because of being decentralized. The large volume of data will make application vulnerable attack if it has to depend upon any centralizes control.
Face book, Twitter and Flicker these are Big Data applications which are used or deployed all over the world to ensure quick response for solution and nonstop services etc.

3.3 Complex and advance Relationship: Big Data application has to take into account complex relationship of data with the advance changes. So, that new way should be finding to accumulate the data.
For example, most of social sites of network are mainly characterized by social functions such as friend-circle and followers. The correlations between individuals naturally make the whole data representation complicated. In a progressive world, the features used to represent the individuals and the social links used to represent our connections may also evolve with respect to spatial, temporal and other factors. Such a complication is becoming part of the reality for Big Data applications, where the key is to take the nonlinear and many-to-many data relationships, along with the evolving changes taking into account to find useful patterns from Big Data collections [2].

IV. CHALLENGES OF DATA MINING WITH BIG DATA

Big Data is *too big,* or *too hard* for existing tools to process. Here too big means when we talk about Petabytes and Exabyte. "Too fast" means that this massive data must be processed. "Too hard" means analysis and tools are not easily provided for this.
For intelligently learning the database system to handle Big Data it is very important to scale up such large volume of data. Fig. 2 shows conceptual view of Big Data processing framework. It includes three Tiers from inside out with

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 9, ISSN No. 2455-2143, Pages 26-31
Published Online July – August 2016 in IJEAST (http://www.ijeast.com)

consideration on data accessing and computing(Tier I), data accessing ,data privacy and domain knowledge(Tier II) and Big Data Mining Algorithms (Tier III).

4.1 Tier I: In this Tier mainly mining is done by using computing units that analysis and compares the data. A computing platform has access to two type of resources.1.) Data 2.) Computing Processor. Data Mining means splitting the billions of records into small tasks which will run on one or more multiple computing nodes concurrently.

- For small scale of data mining single desktop is adequate. Many algorithms are also used for Data Mining.
- For medium scale it depends upon parallel computing programming.
- For Big Data mining, a typical data processing framework will rely on cluster computers and many parallel programming
- Tools are used like MapReduce, Enterprise Control Language (ECL) on multiple computing nodes.

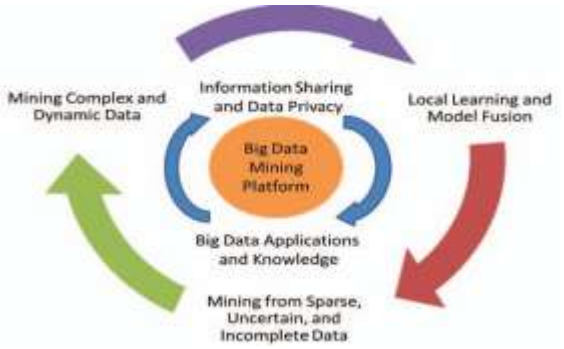

Fig. 1. A Big Data processing framework: The research challenges form a structure of three tier "Big Data mining platform"(Tier I), which focuses on low-level data accessing and computing, information sharing and privacy, and Big Data application, domains and knowledge form Tier II, The outer most circle shows Tier III challenges on actual mining algorithm[2].

4.2 Tier II: This tier addresses two issues
1) Data accessing and privacy
2) Domain and application Knowledge

Information Sharing And Data Privacy: Main goal of all system the sharing of information. To make the privacy of data two approaches are used:

i. Restrict access to data means adding certificate so that particular group has access to that data.
ii. Anonymize data field such that the data will not pinpoint to the individual person. So, its benefit is that the data can be independently shared without involving any restriction.

Domain and Application Knowledge: It provides essential knowledge through which we can design Big Data Mining algorithm. It helps to identify the right feature to model the data. .It also helps to design achievable business objectives by using Big Data analytical techniques.

4.3 Tier III:

4.3.1 Model Fusion for Global Decision: As Big Data applications are featured with self-governing sources and distributed controls. In this environment mining is consistently prohibitive due to the potential transmission cost and privacy issues. On the other side, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each site often leads to biased opinion, just like in the case of elephant and blind men. Under this circumstances a Big Data Mining system has to work on information sharing and joined mechanism to take Global decision.
Several data sources can be fused to meet the global mining goal. Global Mining can be featured at three levels:

a) At Data Level: Each site will calculate data statistics and do local mining and exchange the data among sites.
b) At Model Level: Local Mining takes place in this level.
c) At Knowledge Level: How to take decision by correlating model from autonomous sources.

4.3.2 Mining from improper data:
Sparse means a few data so it is very difficult to draw reliable conclusion from this incomplete data. This kind of problem generally arises in case of High Dimension space. For this various approaches are used like dimension reduction or feature selection to reduce the High Dimensions but the reduced data should contain same analysis result as that of original one.
Uncertain data are special type of data in which data fields are nondeterministic. For example an individual may not feel comfortable to reveal his/her actual salary, but will be comfortable to show a rough range like [100k, 150k]. For uncertain data, the major problem is that each data item is represented as sample distributions but not as an atomic value, so most existing data mining algorithms cannot be directly applied for processing. Common solutions are to take the data distributions into consideration to evaluate model parameters [1].
Incomplete Data means missing of some data field values for samples. Modern Data Mining Algorithms have in-built ability to handle such missing values. And Data Imputation research field seeks to impute missing values in given data to produce improved model that can give us realistic picture of information.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 9, ISSN No. 2455-2143, Pages 26-31
Published Online July – August 2016 in IJEAST (http://www.ijeast.com)

4.3.3 Mining Complex and real time Data:
Mining the complex data in Big Data is the biggest challenge in Big Data Mining.

Big Data Complexity is represented as:
Complex heterogeneous data types, Complex intrinsic semantic associations in data, Complex relationship network among data.

- **Complex Heterogeneous Data Types:**

In Big Data there are various data types are used like structure data, unstructured data, semi structured data and so on. Specifically data is in the form of tables, hyper text, audio and video and so on.

- **Complex Intrinsic Semantic Association in Data:** It is also a challenge for Big Data is to build the semantic association among various heterogeneous and among autonomous data sources.
- **Complex Relationship Network:** The web pages create hyperlinks to various other pages. For example on "www" where each web page has hyperlinks to other web pages also cause complex network.

V. OPEN SOURCE TOOLS FOR MINING BIG DATA

Big Data term is related to Open Source Software Revolution. Various companies like Yahoo, Facebook and Twitter etc. are using various open source software. These software are as follow:

- **Apache Hadoop**: This software is based on Map Reduce Programming Model and in this distributed file system is used which is known as Hadoop Distributed File System (HDFS). Hadoop Allows writing applications that rapidly process massive amount of data in parallel.
- The main task of **Map Reduce** is to split the large data set into small-small task and then process those tasks in parallel. It works like the Divide and conquer technique.
- **Apache Hadoop Related Project:**
- Apache Hive, Apache Pig, Apache Hbase, Apache Zookeeper, Apache Cassandra, Cascading, Scribe etc. [3].
- **Apache S4:** It is a platform to process continuous data stream. It is mainly designed for managing data streams.S4 apps mainly designed by combining stream and processing elements.
- **Storm:** Nathan Marz at Twitter has developed it. It is used for streaming data-intensive applications. It is highly scalable, robust and fault tolerable.
- **R:** It is an Open Source Programming Language. It is used for statistically analysis of large sets of data.
- **MOA:** It is stream Data Mining Open source software .The stream framework is able to use MOA software.
- **SAMOA**: (Scalable Massive Online Analysis) It is upcoming software that will combine S4 and storm with MOA.
- **Vowpal Wabbit:** This Open source Project started at Yahoo! Research and continuing in Microsoft Research to design a scalable, fast and learning algorithms. VW is able to learn from terafeatured data sets. By using this linear learning can be done via parallel learning and because of this throughput of any single machine network interface get increased [3].

For Big Graph Mining following Open Source Tools are used:

- **PEGASUS:** Big Graph Mining system is built on the top of Map Reduce. It allows finding patterns and various anomalies in large real-world graphs [5].
- **Graph Lab:** It is a high-level graph parallel system build without using Map Reduce. In Graph Labs algorithms are expressed as Vertex-Program which are executed on each vertex in parallel and also interact with neighboring vertices [5].

VI. CONCLUSION

Big Data is continuously growing day by day so it is becoming the biggest challenge to process such large volume of data. Big Data is going to be more large, diverse and faster. Big Data is becoming the upcoming final frontier for data research and for business applications The Big Data has geared up the Data Mining, knowledge discovery from data and it has introduced various complex questions to users and to researchers. It also has disclosed the limitations of existing data mining techniques. Big Data Mining is an assuring research field, which is still in the early stage. There is still much work is needed to overcome its challenges.

VII. REFERENCE

[1] B R Prakash1, Dr. M. Hanumanthappa: "Issues and Challenges in the Era of Big Data Mining", 4 July-August 2014.

[2] Xindong wu, Xinquan zhu, Senior member IEEE Gong-Qing wu and wei Ding : " Data Mining with Big Data" January 2014.

[3] Wei Fan, Albert Bifet : "Mining Big Data: Current Status, and Forecast to the Future", Volume 14, Issue 2.

[4] Bharti Thakur, Manish Mann, Computer Science Department LRIET, Solan (H.P), India, 5, May 2014.

[5] Yang Liu, Bin Wu_, Hongxu Wang, and Pengjiang Ma, " BPGM: A Big Graph Mining Tool" .

[6] International Journal of Computer Applications (0975 – 8887) National Level Technical Conference "X-PLORE 14: "Algorithm and Approaches to Handle Big Data".

[7] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer : "MOA: Massive Online Analysis" http://moa.cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.

[8] SAMOA, http://samoa-project.net , 2013

[9] Apache Hadoop, http://hadoop.apache.org

[10] Apache Mahout, http://mahout.apache.org