Challenges and opportunities in Big data analytics

[1]Anil Kumar Mahto, [2]Ranjit Biswas, [3]M. Afshar Aalam
[1]Assistant Professor, [2]Professor, [3]Professor
[1]Department of Computer Science & Engineering,
Jamia Hamdard, New Delhi, India

Abstract: **Big Data is very difficult to captured, stored, analyzed and managed by the traditional existing technology like distributed system, Haddoop ecosystem etc. Hence for cope up with the challenges arises due to big data we need a new storage media, storage technique, new mathematical model to access and manipulation of data new data mining techniques to analyzing the data. In this paper we have discussed the different challenges arises due to Big Data and the possible opportunities for the researcher in near future. In this paper we have describe the different challenges faced by the existing technology for dealing with the Big Data Analytics and possible solution given by the various researchers working in the area of Big data. We also discussed the different opportunities available for the researcher for working in the area of Big Data Analytics.**

Index Terms **–Big Data, Challenges in Big Data, Big Data Analytics, Hadoop, Map Reduce, Hadoop File System,**

1. Introduction:

The Big data [1] is defined as the dataset that is very difficult to captured, stored, managed and analyzed by the existing technology. The Big data is very difficult to store by the traditional storage media, difficult to manage by traditional DBMS tools, and difficult to find out the useful knowledge by the traditional data mining techniques. The Presently Big data expanding very rapidly mainly the dimension of 4Vs: Volume, Varity, Velocity and veracity, and also in many more dimensions. Here Volume represent the amount of data (In Petabytes, Zettabytes), Varity represent the types of data (Text, Image, Video), Velocity represent the growth of data day by day, and veracity represent the correctness of data.

Demonetization of 500 and 1000 currency in India was done on 8th November 2016. This was one of the most challenging and controversial decision taken by the Indian government. Searching on Google with "Demonetization in India" resulted in about 75,00,000 web links on internet as on 6 June 2018. Many people praise the demonetization while many people criticize the decision across the social media, print media, news channels etc. Can we summarize the different opinions came from different sources like twitter, Facebook, News channels, blogs of critics in real time fashion. This type of summarization program is an excellent example for Big Data processing, because the large amount of data (volume), came from different sources in different variety like text, images, video (Variety), and the amount of data keeps growing (Velocity), and the data may also come from the unauthenticated sources(Veracity).

1.1 Characteristics of Big Data:

The Big Data is the large data set that is very difficult to manage by the traditional existing technology due to its characteristics given by the researchers mainly in 3 V's, 4 V's, and 5 V's. These V's are Volume, Velocity, Variety, Veracity and Value. Some of the characteristics are as follows in terms of V's:

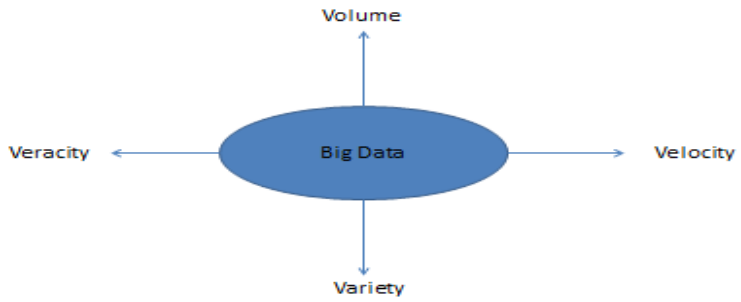


Figure 1: Attributes of Big Data in terms of V's

Volume:

Big Data implies huge amount of data. Now these data is generated by the our day to day life activities (like shopping of daily needs items, social medias), health care industries, Social media and print media, ERP etc. According to [5] "It's obvious that data volume is the primary attribute of big data. With that in mind, most people define big data in terabytes—sometimes petabytes. For example, a number of users interviewed by TDWI are managing 3 to 10 terabytes (TB) of data for analytics. Yet, big data can also be quantifed by counting records, transactions, tables, or files." 90% of all data ever created, was created in the past 2 years. From now on, the amount of data in the world will double every two years. By 2020, we will have 50 times the amount of data as that we had in 2011. The sheer volume of the data is enormous and a very large contributor to the ever expanding digital universe is the Internet of Things with sensors all over the world in all devices creating data every second.

Velocity:

Velocity means the speed of new data generation and transfer of data from one source to other sources. We can take an example of social media messages, which is going to be viral with in very less span of time. According to [1] the speed at which data is

created currently is almost unimaginable: Every minute we upload 100 hours of video on YouTube. In addition, every minute over 200 million emails are sent, around 20 million photos are viewed and 30,000 uploaded on Flickr, almost 300,000 tweets are sent and almost 2.5 million queries on Google are performed.

Variety:

Variety means data comes from different sources (Databases, spreadsheets, Social Media, News Channels etc.) in different formats like structured, semistructered and unstructered in the form of Text, images, video etc. Hence due to unstructured nature of data it is very difficult to store, analyzing, mining etc. Today for industry variety of data is one of biggest challenges in the field of big data analytics.

Veracity:

The meaning of veracity is conformity to fact. Its synonyms are: correctness, accuracy, realism, faithfulness etc. Because today the source of data can be untrusted, unauthorized like social media where anyone can give their own opinion without taking any serious responsibility. The main issue is the "The data is going to store and analyze is meaningful or not?" Hence we should have a mechanism that fairly deals with the correctness of data that is going to be analyzed for knowledge discovery.

In [2] Big Data characteristics is explain with the help of HACE Theorem as" Big Data starts with large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data." Due to these characteristics it is extremely challenging for discovering useful knowledge from the Big data.

2. Challenges with Big Data Analytics:

Challenges with Big Data Analytics are mainly divided into three stages. Challenges in Stage-1 are also known as characteristics of Big Data, already discussed in the above section. The challenges in stage-2 are mainly related to the processing of big data like data capturing, data cleaning, data integration, analysis and modeling. In this stage we mainly concerned about how to apply the data mining techniques for analyzing the big data. Challenges in stage-3 are mainly related to data security, privacy, data ownership and data governance.

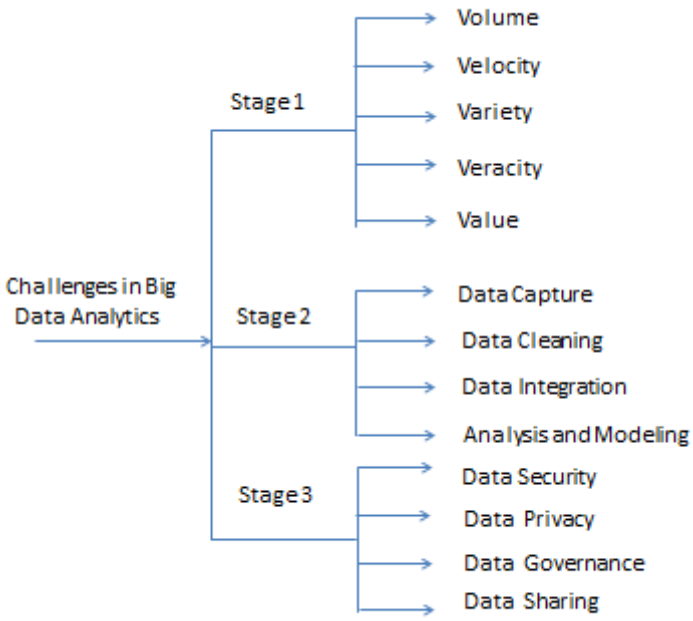


Figure 2: Challenges in Big Data Analytics

In [4] various issues with big data like Storage and Transport Issues, Management Issues, Processing issue was discussed.

Memory Units:

1 Bit = 1 Binary Digit(0 or 1)	1024 TB = 1 Petabyte (PB)
8 bits = 1 Byte	1024 PB = 1 Exabyte (EB)
1024 Bytes = 1 Kilobytes (KB)	1024 EB = 1 Zetabyte (ZB)
1024 KB = 1 Megabytes (MB)	1024 ZB = 1 Yottabyte (YB)
1024 MB = 1 Gigabytes (GB)	1024 YB = 1 Broontobyte (BB)
1024 GB = 1 Terabyte (TB)	1024 BB = 1 Geopbyte

Assume that an Exabyte of data needs to be processed in its entirety. For simplicity, assume the data is chunked into blocks of 8 words, so 1 Exabyte = 1024 petabytes. Assuming a processor expends 100 instructions on one block at 5 gigahertz, the time required for end-to-end processing would be 20 nanoseconds. According to work presented in [4] "to process 1K petabytes would require a total end-to-end processing time of roughly 635 years." Thus, effective processing of Exabytes of data will require

extensive parallel processing and new analytics algorithms in order to provide timely and actionable information. According to work presented in [11] the following two sets is defined as

4V = { Volume, Varity, Velocity, Veracity}

4N = { New Theories, New Hardware, New Software, New Models}

The first set defined the problems and characteristics of Big Data while the 4N set contains the solution to handle the existing challenges concerned with the Big Data. But the main problem is the development in the set 4N, New theories, New Hardware, New Software and New Models is much lesser than the expansion of 4V set. Hence for proper solution to handle the big data is to develop a proper 'New Data Structure', 'New type of distributed system' that should be compatible with the new data structure, 'New Network Topology' to support the new distributed system and new mathematical theories and models.

2.1 Challenges with classical Data Processing Approach:

As per the work presented in [20] the Big data mining system were very rare and expensive. The main reason for that is system should be upgraded to process the huge dataset, because traditionally a normal computer system having limited storage and processing power.

Hence the two main concepts: Scale-up and scale-out, is used to upgrade the processing power as well as the storage of the existing computer system. Scale-up is mainly define as the growing a system onto bigger and bigger hardware. Suppose if the amount of data doubles, then simply doubled the hardware of a single system. Scale-out approach spreads the processing onto more and more machines. If the data set doubles, simply use two machines instead of a single machine of size double. If it doubles again, move to four systems.

But the main problem with the scale-up approach is there is a limit to the size of an individual server that can be purchased from the hardware suppliers. Also at some point, the amount of data will be beyond the capacity of single scale-up system. Hence we have the scale-out approach as a second option. Here the main concept is of having two servers instead of single large system. If the requirement will increase then we can increase the number of servers. Due to end of road tendency in scale-up approach, this technique very rarely used in Big data analytics and the best choice is to choose scale-out approach. But the problem with the scale-out approach is reliability.

The terms "five nines" referring to 99.999 percent uptime or availability. Though this is absolute best-in-class availability, it is important to realize that the overall reliability of a system comprised of many such devices can vary greatly depending on whether the system can tolerate individual component failures. Assume a server with 99 percent reliability and a system that requires five such hosts to function. The system availability is 0.99*0.99*0.99*0.99*0.99 which equates to 95 percent availability. But if the individual servers are only rated at 95 percent, the system reliability drops to a mere 76 percent.

Hence the traditional DBMS and distributed system is not good for the Big Data Analytics. One of the popular alternatives used popularly today for this type of problem is Hadoop. It is open source software that is used for distributed computation for high volume of data. It is designed using the scale out.

2.2 Hadoop:

Hadoop [3] is an Apache open source project which consists of HDFS, Map Reduce, HBase, Hive and ZooKeeper and other projects.Hadoop has two primarily parts:Hadoop distributed file system (HDFS) and MapReduce programming model .HDFS is an open source version of the Google GFS implementation, as a highly fault-tolerant distributed file system, which provides high throughput data access, suitable for mass storage (PB-class) of large files.

2.2.1 Elementary parts of Haddop:

Nodes: It is the elementary part of the Hadoop. A node is a simple a computer having traditional processing power, memory etc..

Rack: Rack is generally defined as the collection of nodes (Generally 30- 40 Nodes). Diagrammatically a node may be described as shown in the below figure.

Hadoop Cluster: A hadoop cluster is defined as the collection of racks.

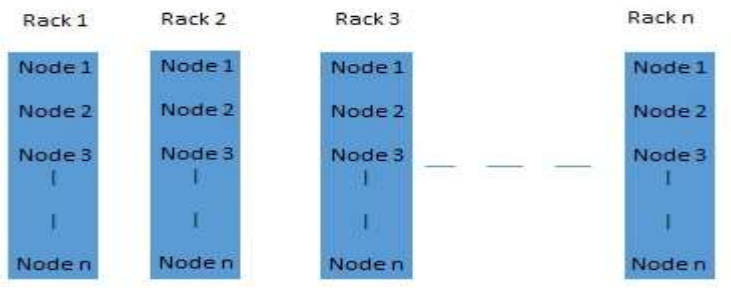


Figure 3: A Hadoop Cluster: Collection of Racks

2.2.2 Components of Hadoop:

Hadoop mainly having two components: Distributed File System and Map Reduce.

Distributed File System:

Hadoop File system that runs on top of existing file system, is known as Hadoop Distributed File system (HDFS). It is designed to handle very large files with streaming data access pattern. It is using sequential data access rather than random data access. It is using file blocks to store file or parts of a file. A file block in Hadoop is of default size 64 MB and recommended size is of 128

MB whether the block size in UNIX system is 4 KB. So behind the scene, 1 HDFS block is supported by multiple Operating System Blocks. The main advantage of having the blocks is, the blocks can be replicated to multiple nodes. Also due to the nature of having fixed size of the block, it is easy to calculate that how to fit on a disk.

Map Reduce:

A MapReduce program consists of map and reduce functions. A MapReduce job is broken into tasks that run in parallel.

HDFS nodes are divided in two categories: Name Nodes and Data Nodes. Map Reduce nodes are also divided in two categories: Job Tracker and Task Tracker.

2.2.3 Hadoop Framework:

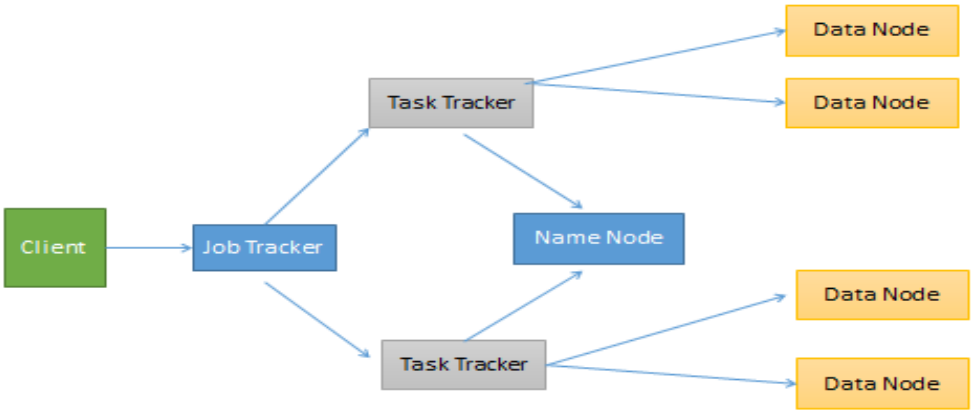


Figure 4: Hadoop Framework

HDFS Nodes: HDFS nodes are divided in two categories: Name Nodes and Data Nodes.

Map Reduce Nodes: Map Reduce nodes are also divided in two categories: Job Tracker and Task Tracker.

Name Node: Only one Name node exist per Hadoop cluster. It manages the file system namespace and metadata. But one of the main problem is single point of failure will effect the whole cluster.

Data Node: We can have more than one Data Node per Hadoop cluster. It manages the Blocks with data and serves them to clients. It periodically reports to the Name Node about the list of locks it have stored. It uses cheap hardware commodity.

Job Tracker Node: We can have one Job tracker Node per Hadoop cluster. It revives job request submitted by the client. It mainly schedules and monitors MapReduce job on Task Tracker.

3. The Opportunities in big data analytics:

We consider the most common problem on Big Data Analytics here in four sequential phases and each phases having lots of opportunity for the researchers. The phases are as follows:

Phase-1: Hardware Issue: Where to store big data for real time processing?

Phase-II: How to Store: Which Data Structure and Distributed system is suitable for Big Data?

Phase-III: How to improve traditional Data Mining Algorithms on Big Data?

Phase-IV: If possible then how to develop a new language for Big Data by developing New Data Structure for Big Data, New Network Topology and Distributed System for Big Data.

3.1 Phase-I: Hardware Issue: How to store big data for real time processing?

We know that one of the most important characteristic of big data is volume. The global data explosion can be imagine by the following diagram:

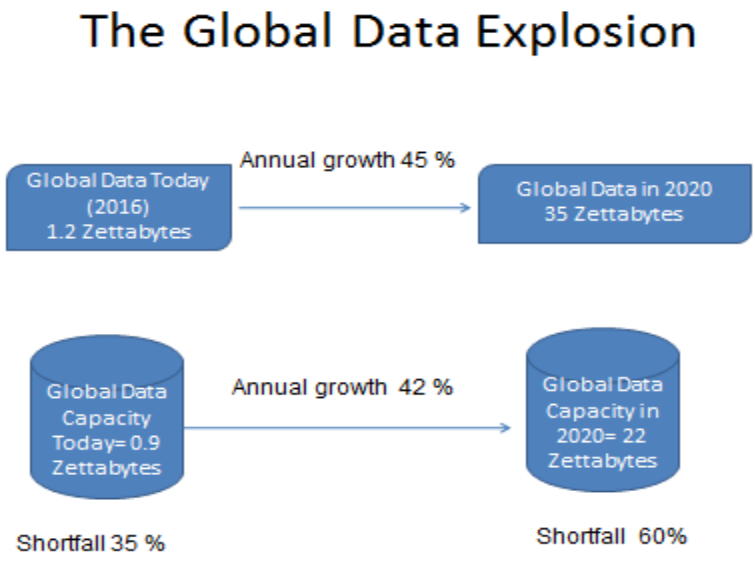


Figure 5: Global Data Explosion

Example of a Sample Problem:

Here we consider a normal computer having 1TB hard disk. For any purpose suppose we need to download 1 petabyte (PB) from internet. Hence due to shortage of memory we cannot store the desired data and if we cannot store then we cannot process the data.

Here we can have two solutions: Either we can have a single system (scale-in) with the desired storage (1 PB) or we can have at least 1024 normal computer systems(scale-out).

Not Stored ===> No Processing Possible!!

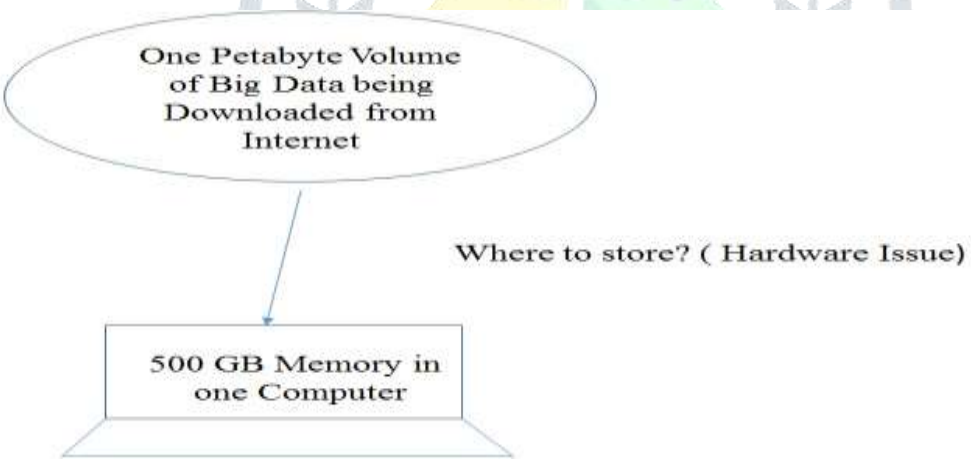
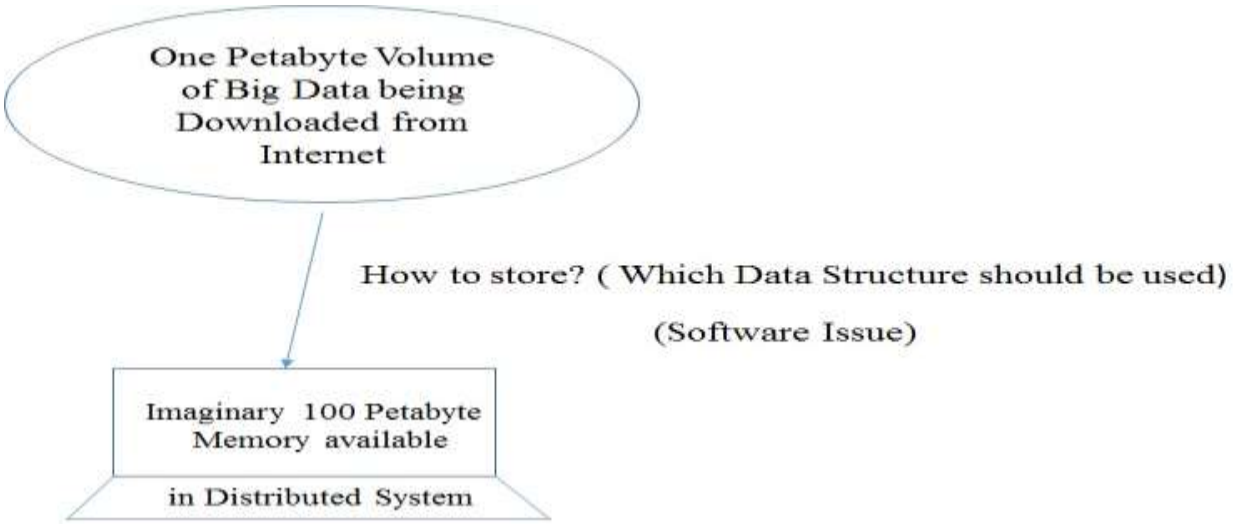


Figure 6: Storage (Hardware) Issue in a single Simple Computer

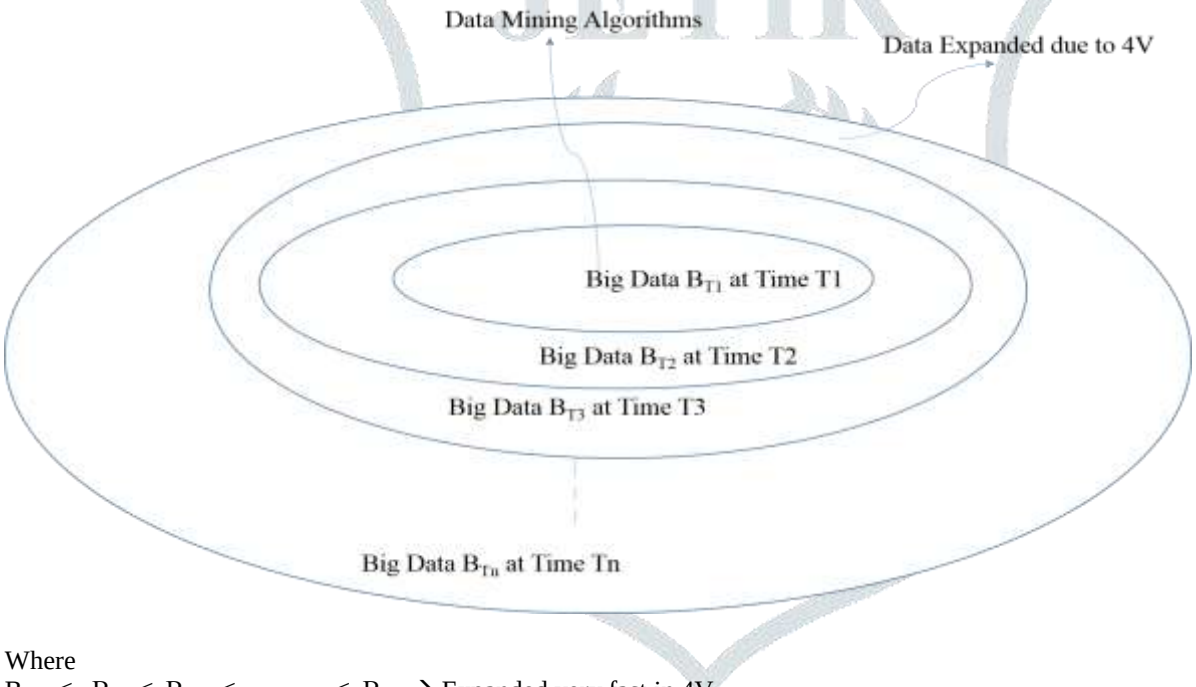
3.1 Phase-II: How to Store:

Here we assumed that we have enough hardware resources to store the data. The main problem in this phase is what kind of distributed system will be suitable for storing the data. As we know the distributed system having the challenges like fault tolerance, availability etc. A new distributed system of high fault tolerance and compatible with the new storage model should be develop. To support the new distributed system a very new network topology model should be develop. Here we can have an additional challenge of suitability of the data structure to be used for accessing the big data. We know that big data is heterogeneous and unstructured in nature. Hence homogeneous data structure will not be suitable to access big data.



3.3 Phase-III: Application Issue: How to improve traditional Data Mining Algorithms on Big Data?

Assumption: Let us assume that by rigorous research work we solved the previous problems then next problem will be to improve the existing Data Mining Algorithms so that the improved algorithms will be applicable for the Big Data Processing.



Where

$B_{T1} \leq B_{T2} \leq B_{T3} \leq \ldots\ldots\ldots \leq B_{Tn}$ → Expanded very fast in 4V

We know that big data expanding very fast in 4V's direction. Hence the big data at time T_1 denoted by B_{T1} will be different from the big data at time T_2 denoted by B_{T2}. Hence it might be a case that the data mining algorithm is suitable for B_{T1} but not suitable for B_{T2}.

3.1 Phase-IV: Programming Language Issue: how to develop a new language for Big Data?

Is it possible to develop a new language especially for big data? For example we have structured query language especially for databases, R language for data analytics, C language for procedural problem etc. Hence researcher can work on to develop a separate language for big data.

Conclusion:

Big Data is similar to 'small data' but the amount is much bigger and it is expanding very fast in 4V (Volume, Velocity, Variety and Veracity) and it is very difficult to handle by the tradition existing technology. Hence for cope up with the challenges arises due to big data we need a new storage media/technique, new mathematical model to access and manipulation of data. In this paper we have discussed the different challenges arises due to Big Data and the possible solving techniques in near future. In this paper we are going to describe the different challenges faced by the existing technology for dealing with the Big Data Analytics and

possible solution given by the various researchers working in the area of Big data. We also discussed the different opportunities available for the researcher for working in the area of Big Data Analytics.

References:

1. Vassakis, Konstantinos, Emmanuel Petrakis, and Ioannis Kopanakis. "Big Data Analytics: Applications, Prospects and Challenges." *Mobile Big Data*. Springer, Cham, 2018. 3-20.
2. Xindong, et al. "Data mining with big data." *IEEE transactions on knowledge and data engineering* 26.1 (2014): 97-107.
3. Zhang, Da-Wei, et al. "Research on hadoop-based enterprise file cloud storage system." *Awareness Science and Technology (iCAST), 2011 3rd International Conference on*. IEEE, 2011.
4. Kaisler, Stephen, et al. "Big data: Issues and challenges moving forward." *System Sciences (HICSS), 2013 46th Hawaii International Conference on*. IEEE, 2013.
5. Russom, Philip. "Big data analytics." *TDWI best practices report, fourth quarter* 19 (2011): 40.
6. Xindong Wu, Gong-Qing, Data Mining with Big Data, IEEE Transaction on Knowlrdge And Data Engineering, Vol. 26, No. 1, January 2014.
7. Ahad, Mohd Abdul, and Ranjit Biswas. "Comparing and Analyzing the Characteristics of Hadoop, Cassandra and Quantcast File Systems for Handling Big Data." *Indian Journal of Science and Technology* 10.8 (2017).
8. Stephen Kaisler, Frank Armour, J. Alberto Espinosa and William Money, Big Data: Issues and Challenges Moving Forward, 46th Hawaii International Conference on System Sciences, 2013
9. Puneet Singh Duggal, Sanchita Paul, Big Data Analysis: Challenges and Solutions, International Conference on Cloud, Big Data and Trust 2013, Nov 13 -15, RGPV
10. Garry Turkington, Hadoop Beginner's Guide, PACKT publishing 2013
11. Ranjit Biswas, Heterogeneous Data Structure "R-Atrain" ,'Information' An international Journal (Japan), Vol.15(2) February 2012, pp 879-602.
12. Ranjit Biswas, Processing of Heterogeneous Big Data in an Atrain Distributed System (ADS) using the Heterogeneous Data Structure r-Atrain, International Journal on Computing and Optimization, Vol.1(1) 2014, pp 17-45.
13. Divyakant Agrawal, Amr El Abbadi, Shyam Antony, Sudipto Das,Data Management Challenges in Cloud Computing Infrastructures, University of California, Santa Barbara
14. Sangeeta Bansal, Dr. Ajay Rana, Transitioning from Relational Databases to Big Data, International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 1, January 2014 ISSN: 2277 128X.
15. Michael A. Bender, Bradley C. Kuszmaul, Data Structures and Algorithms for Big Databases.
16. Du Zhang, Inconsistencies in big data, 12th IEEE International Conference on Cognitive Informatics & Cognitive Computing, 2013
17. D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, and U. Dayal, Challenges and opportunities with big data, Cyber Center Technical Report 2011-1, Purdue University, January 1, 2011.
18. Ana Lucia Varbanescu, Alexandru Iosup, On Many-Task Big Data Processing:from GPUs to Clouds, The International Conference for High Performance Computing, Networking, Storage and Analysis 2013.
19. R. Ramakrishnan, "Big data in 10 years," in IPDPS, 2013, p. 887.
20. http://www-01.ibm.com/software/data/infosphere/hadoop/
21. http://www.dataintensity.com/characteristics-of-big-data-part-one/