IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727

Survey on Big Data Tools and Techniques

Veenakshi Devi[1], Meenakshi Sharma[2]
[1,2,](Department of Computer Science & Engineering, Sri Sai College of Engineering and Technology)

ABSTRACT : *Collection of huge quantity of data is generally called as Big data. Its playing a crucial role in todays world and is one of the necessary technology for the upcoming era..Big data technologies plays a vital role in providing accurate analysis that leads to making of decisions in a tangible manner that too results in reduction of cost, risk and better efficiencies in terms of various operations. Moreover, Big data does not mean concern of data only, it implies various tools and techniques and framework. The Apache open framework that is used to handle Big data is Hadoop, Hadoop is basically designed to handle one server to number machines, each machine offering processing and storage .Hadoop framework works on the concept of map reduce. In this paper we summarized the Big data concept, its tool (Hadoop) ,various challenges and different platforms that Big data is using.*

Keywords – *Hadoop , HDFS, Spark , Map Reduce, Big data platforms.*

I. INTRODUCTION

Big data is nothing but the data. This data should be in any form either structured or unstructured. It contains a huge amount of data and all the time it is generated by us. Everything creates the data i.e. websites, blogs, social sites, database etc. and this data transfer via sensors or hardware systems. These sensors also generate data. So the main issue is how to handle this huge amount of data. If hardware is suitable for an amount of data, it does not mean that it will also be suitable for other type of data. Continuous increment of data always outdates the particular hardware and software. Every time researchers update the software/hardware to handle the increased data. Generally big data works of v3 concept. V3 is nothing but the velocity, variety and variability. Velocity means that the data is coming in which speed? What type of format it has? Also the decision making power relate with the big data because if the data will be accurate the decision will be confidently.

The platforms that are provided by big data have different characteristics and from different perspective different platform should be used. So it is necessary to know the depth knowledge of that platform [1]. It is the responsibility of the platform to provide the analytics solutions and it is possible only when the particular platform accepts our whole data.

Due to advancement in new technologies, devices and communication means like social networking sites, the amount of data produced by human being is growing at a very fast manner. The amount of data produced by mankind from beginning of time till 2000, if we count roughly was just 5 billion Gigabytes. But now a days the same amount was created in every two days in 2011 and every minute in 2013,the rate of data is growing more enormously in 2014.The large amount of data sets is called as Big data ,it involves bulk of data that is unstructured and such kind of data need much more real time analysis. Data stored in traditional ware houses differ to a great extent from Big data ,the data that is stored in warehouses need to be cleaned properly ,documented completely ,also it should be compatible with the basic structure of the particular warehouse ,Big data involves not only handling of data that is needed to be stored in warehouse but it also deals with the data not suitable for storing in the ware houses .Thus better business strategies and proper analysis of data is needed for handling of Big data. New opportunities in order to discover new values are also brought by Big data, It also help us to discover new challenges that is how to manage and organize data set in an effective manner [1].

Big data include data generated by different sources. Here are some of the sources from where most of Big data is generated :

- Black box data : It captures voices of the flight crews, recording of microphones and ear phones and the performance information of aircraft.

- Social media data: Social networking sites generates huge amount of data in the form of comments and posts by billions of people all over world.
- Stock exchange data: This includes data about the assessments of the various buyers and sellers made regarding share of different companies.
- power grid data : Power grid data is collection of various information regarding power uses by specific node with respect to another node taken as base station
- Transport data: Transport data includes information regarding transport vehicle .Such information contain data like model of vehicle, capacity and vehicle availability.
- Search engine data: Search engine contain information regarding the data that is retrieved from various databases.

The data produced by all these fields is generally of three types structured, semi structured and unstructured. Moreover, Big data can be defined with the following properties associated with it [2]

- Variety: Data is of various categories such as various pages of website, files, social networking sites, electronic- mail, documents and also data related to sensor devices.
- Volume: Big data means the large quantity of data .Till now data is available in petabytes and is supposed to increase at faster rate in upcoming years.
- Velocity: Velocity that deals with what speed that data is getting collected from different sources.
- Variability: It deals with the inconsistencies related to the flow of data; Big data become a challenge to be maintained when the data is increasing in a frequent manner.

.

II. BENEFITS OF BIG DATA

Big data is becoming important in our day to day life .Few are the benefits of Big data:

- By applying the information from social networking sites such as Facebook , the various agencies can learn better ideas about their campaign, Promotions and advertising medium.
- Social media usage information for preferences and product perception of their consumers, retail organization and companies which are planning for the production
- Hospitals are providing efficient and effective services by collecting historical patient data and previous medical history of patients.

III. TOOL FOR BIG DATA: HADOOP

Dough cutting ,Mike Cafarella and team took the solution provided by Google and started an open source Project called Hadoop in 2005.Now Apache Hadoop is a registered trade- mark of the Apache software foundation .Hadoop basically runs applications using Map Reduce algorithms ,Map reduce is a programming model [3]where data is processed in parallel on different CPU nodes. Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for a huge amount of data.

Hadoop is open source software, it involves many sub projects, and these belong to the type of infrastructure for developing distributed computing [4].Hadoop frame work consists of the following modules:

- Hadoop common
- Hadoop Yarn
- Hadoop file system
- Map reduce

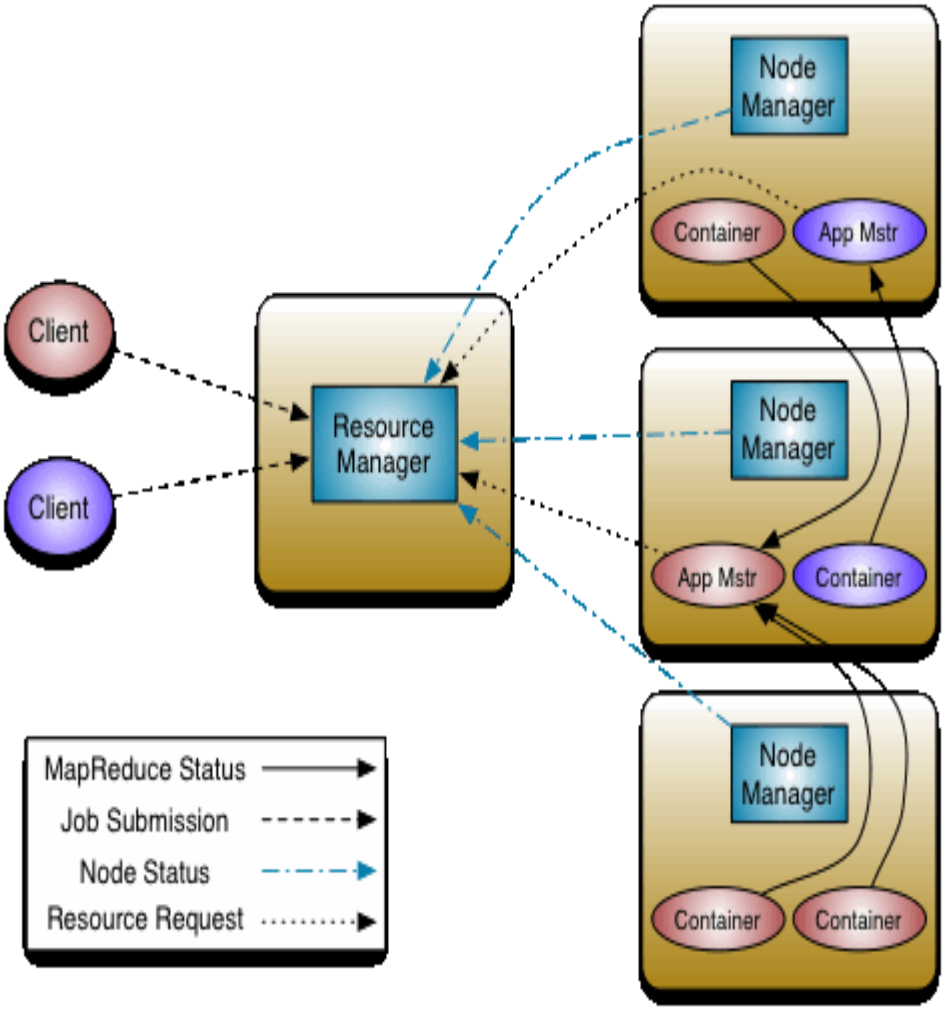

Fig. 1: Hadoop architecture

Hadoop common are java libraries and utilities required by Hadoop module .These libraries provides file systems and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop. Hadoop yarn is a framework for job scheduling and cluster resource management. The core of Apache Hadoop consist of storage part-known as Hadoop distributed file system (HDFS) [5], [6] and another processing part Map reduce [7]. Hadoop distributed file systems designed for storage of files with large data.

Moreover HDFS block size is much larger than that of normal file system i.e. 64 MB by default .A HDFS cluster has two types of nodes that is name node and data node [8].Name node is called master node and data nodes are called workers. The function of the name node is to manage the system name spaces; it is also responsible for maintaining the file system tree, maintaining the metadata of all the files and directories in the tree [4]. The data nodes works according the instruction of the name node .It is not possible to access the file without name node, thus name node should not be prone to any kind of failure. Hadoop Map reduce is a

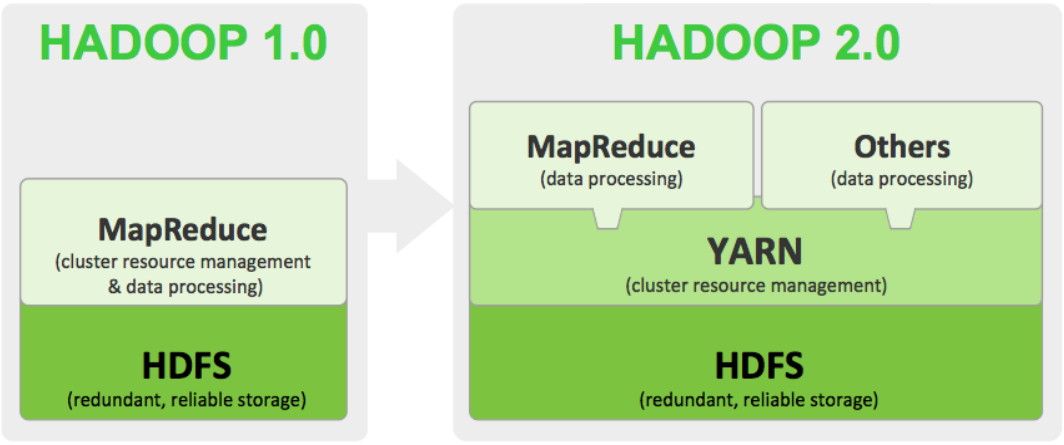

Fig. 2: HDFS architecture

software framework for writing applications in an easy manner .It process vast amount of data in parallel on large clusters of a hardware in a reliable manner. Map reduce usually splits the input data sets into independent chunks, these chunks are then processed by map task in a complete parallel manner .The sorting of the outputs of the map is performed by the framework, further these outputs are input to the reduced tasks .Thus finally both the input and output of the job are stored in the file system .Also Map reduce framework consist of a single master for job tracker and one slave task tracker per cluster node. The master is responsible for scheduling the jobs component task on the slave, it monitors them and also executes the failed task, and also the slave executes the task as directed by the master.

IV. BIGDATA PLATFORMS

Platforms are available for performing Big data analytics, these platforms are based upon the type of scaling. Thus on basis of this we have horizontal scaling and vertical scaling. We will discuss these as follows [9] Scaling it generally a measure how a system is able to adapt itself to increasing demand in terms of processing of data .Scaling is of two types:

Horizontal scaling— In this technique, the load of the work distributes with the multiple servers into pool of resources and that is also called scale out scaling. The servers or machines that have been added into your pool are not dependent on each other. Every machine has different instances of work. By doing this the processing power is increased of the server.

In the horizontal scaling, it is not difficult to increase the scaling in this scaling dynamically. Examples of horizontal scaling are Cassandra etc. It is not a big deal to increase the scaling. Like Cassandra, we have to add only hardware so the fault tolerance is zero for it, as it runs on thousands of servers. When we need space, we can add the servers. When we increase the capacity, it provides the administration powers. It is not a simple task to maintain multiple servers because they are in distributed manner.

Vertical scaling—Under this scaling technique, not the complete server but the processors are added to increase the power of single machine. The reason behind it is that more memory is directly related to the fast processing. This scaling is based on partitioning the data, that is, each machine has some amount of data of a complete process. In this scaling the complete process belongs to a single server, the distribution is according to the resources of that machine. In vertical it is difficult to scale up because there is only one machine. Mysql is an example of vertical scaling for managing the vertical scaling is not a big task because there is only one machine that we have to handle. But sometimes cost matters because machine has multiple number of CPU and RAM. There are some platforms on which these scaling techniques work. So they are as follows.

4.1 Horizontal Scaling Platforms

4.1.1 Peer to Peer Network: It involves about millions of machines connect to a network .There is availability

of decentralized and distributed network architecture. Each node is capable of storing and processing data [10], [11].

4.1.2 Apache Hadoop: It is an open source framework for storing and processing large data sets. There is high fault tolerance and it is designed to be used with commodity hardware.

- Hadoop hdfs : It is used to store cluster of commodity machine thus providing high availability and fault tolerance.
- Hadoop yarn: It is the resource management layer.
- Hadoop map reduce: It is basic data processing scheme used in hadoop .It include breaking the entire scheme into mappers and reducers.

4.1.3 Spark: This platform behaves like map reduce but with the overcome of the limitations of map reduce. It performs analytics and distributed computing tasks. As we know that the execution of map reduce hadoop is not fast so it is the solution of this problem. It generates memory computations for fast execution process. It also works with the live streaming data from Hadoop Distributed File System.

It is a framework in which huge data process parallel and responses of the query within a second. It also eliminates the disks as it use the cache memory to store the data. We can say that it is the best option for big data.

4.2 Vertical Scaling Platforms

4.2.1 High performance computing (hpc) clusters: It is also known as blades or supercomputers with thousands of processing cores. It contain well-built powerful hardware optimized for speed and throughput cost of scaling is high.

4.2.2 Multicore cpu: It involves one machine having dozens of processing cores .Parallelism is achieved through multithreading .In this number of cores per chip and number of operations a core can perform has increased to a great extent.

4.2.3 Graphics processing unit: It includes hardware with massively parallel architecture .It has large number of processing core typically more than 2500 currently. Recent development in GPU hardware and programming framework has given rise to General purpose computing on graphics processing unit.

4.2.4 Field programmable gate arrays: It has highly specialized hardware unit and it can be easily optimized for speed, in this the cost of development is much higher.

.

Scaling type	Platforms (Communication Scheme)	System/Platform			Application/Algorithm		
		Scalability	Data I/O performance	Fault tolerance	Real-time processing	Data size supported	Iterative task support
Horizontal scaling	Peer-to-Peer (TCP/IP)	★★★★★	★	★	★	★★★★★	★★
	Virtual clusters (MapRedce/MPI)	★★★★★	★★	★★★★★	★★	★★★★	★★
	Virtual clusters (Spark)	★★★★★	★★★	★★★★★	★★	★★★★	★★★
Vertical scaling	HPC clusters (MPI/Mapreduce)	★★★	★★★★	★★★★	★★★	★★★★	★★★★
	Multicore (Multithreading)	★★	★★★★	★★★★	★★★	★★	★★★★
	GPU (CUDA)	★★	★★★★★	★★★★	★★★★★	★★	★★★★
	FPGA (HDL)	★	★★★★★	★★★★	★★★★★	★★	★★★★

Fig. 3: Comparison of Various Big Data Platforms

V. BIG DATA CHALLENGES

The sharp increase in the rate of data in big data brings huge challenges on acquisition of data; its proper storage management and analysis .Following are the few challenges for big data:

- Processing and storage issues: Since Big data involves large amount of data ,this processing of such large data takes more time ,more over it also require proper storage for storing huge data [4],uploading this large data in cloud [13] [14] [15] does not alone solve the problem ,thus right indexes should be build up in the beginning itself while the data is being collected and stored, this will make a good practice and the time for processing is reduced to a great extent.
- Representation of data: Data in Big data is of heterogeneous in nature .The representation of data should be in such a manner that it reflects the appropriate data structure, class type, this will ensure better operations on various data sets .Thus it imposes a big challenge on Big data.
- Security and privacy: The data stored in big data should be in such a form so that it has proper security parameters imposed on it to avoid unauthorized access to data, this will help in maintaining privacy also [12]. Big data tools should be analyzed to the massive amount of threats received from data daily, thus security vendor needs to continuously update their global threat intelligence.
- Management and sharing of data: Data in Big data need to be properly managed in terms of accuracy and completeness, this will allow in better decision making. Also the making of data sets available to the public and between various agencies should be done to the extent made possible by private laws [13] .This is also a difficult task.
- Fault tolerance: Whenever there is any kind of failure in Big data, the damage should be done within certain threshold so that there is no need to begin task from the start .To avoid any kind of fault, we can use the concept of applying check points, so that if there is any failure the task can be restarted from the check points maintained .Thus this sometimes becomes cumbersome process.
- Scalability: The scalability in Big data require to run and execute large jobs so that goal of each job can be maintain in an effective manner .Thus scalability is big challenge for Big data.
- Quality of data: The main focus of Big data is to have good quality rather than having large data that is irrelevant This is necessary for better results and conclusion that are needed to be drawn from big data.
- Requirement of better skills: Big data should have better skill sets and these skills need to be developed in individuals, for this there is need to organize various training skills within the organization. Various universities need to introduce curriculum on Big data so that intelligent and knowledgeable employees can produced.

VI. CONCLUSION

This paper described the Big data in an effective way, to better adapt to this era of Big data technology there are many challenges are issues that are encountered, these should be brought up and tackled in an appropriate way .Thus our future research will focus on understanding the Big data issues in a complete manner and also to focus on those various factors that lead to contribution on Big data analysis and methodology.

REFERENCES

[1] M. Chen, S. Mao, and Y. Liu, Big data: A survey, Mobile *Networks and Applications, vol. 19(2)*, 2014, pp. 171–209.
[2] Y. Demchenko, C. De Laat, and P. Membrey, Defining architecture components of the big data ecosystem, *in IEEE International Conference on Collaboration Technologies and Systems (CTS)*, 2014, pp. 104–112.
[3] J. Dean and S. Ghemawat, Mapreduce: simplified data processing on large clusters, *Communications of the ACM, vol. 51 (1)*, 2008, pp. 107–113.
[4] A. Katal, M. Wazid, and R. Goudar, Big data: Issues, challenges, tools and good practices, *in Sixth IEEE International Conference on Contemporary Computing (IC3)*, 2013, pp. 404–409.

[5] D. Dev, Hadoop distributed file system-experiment and analysis for optimum performance, *International Journal of Cloud Computing and Services Science (IJ-CLOSER), vol. 3 (5),* 2014.

[6] T. White, Hadoop: The definitive guide, O'Reilly Media, Inc., 2012. [7] J. Dean and S. Ghemawat, Mapreduce : simplified data processing on large clusters, *Communications of the ACM, vol. 51 (1),* 2008, pp. 107–113.

[7] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, The hadoop distributed file system, *in IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST),* 2010, pp. 1–10.

[8] D. Agrawal, S. Das, and A. El Abbadi, Big data and cloud computing: current state and future opportunities, *in Proceedings of the 14th ACM International Conference on Extending Database Technology,* 2011, pp. 530–533.

[9] A. Rowstron and P. Druschel, Pastry: Scalable, decentralized object lo- cation, and routing for large-scale peer-to-peer systems, *in Springer Middleware,* 2001, pp. 329–350.

[10] S. L. Garfinkel, A. Juels, and R. Pappu, Rfid privacy: An overview of problems and proposed solutions, *IEEE Security & Privacy, (3),* 2005, pp. 34–43.

[11] H. Chen, R. H. Chiang, and V. C. Storey, Business intelligence and analytics: From big data to big impact, *MIS quarterly, vol. 36 (4),* 2012, pp. 1165–1188.

[12] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, Big data: Issues and challenges moving forward, *in 46th IEEE Hawaii International Conference on System Sciences (HICSS),* 2013, pp. 995–1004.

[13] Anjali, Jitender Grover, Manpreet Singh, Charanjeet Singh, Hemant Sethi, "A New Approach for Dynamic Load Balancing in Cloud Computing", IOSR-JCE, Special Issue AETM'15, Vol. 3, pp. 30-36, April 4, 2015.

[14] Jitender Grover and Shivangi Katiyar, "Agent Based Dynamic Load Balancing in Cloud Computing", IEEE International Conference on Human Computer Interactions (ICHCI'13), Saveetha University, Chennai, DOI: 10.1109/ICHCI-IEEE.2013.6887799, pp. 1-6, 23-24, August 2013.

[15] Jitender Grover, Shikha and Mohit Sharma, "Cloud Computing and Its Security Issues - A Review", IEEE Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT 2014), Hefei, Anhui, China, DOI: 10.1109/ICCCNT.2014.6962991, pp. 1-5, July 11-13, 2014.