International Journal on Emerging Technologies (Special Issue NCETST-2017) **8**(1): 72-77(2017)
(Published by Research Trend, Website: www.researchtrend.net*)*

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Big Data Issues and Challenges in 21st Century

Krishna Kumar[1] and Akhilesh Dwivedi[2]
[1]Assistant Professor (Computer Science), Roorkee Institute of Technology, Roorkee, (Uttarakhand), India
[2]Assistant Professor (Computer Science), School of Computing GEHU Bhimtal Campus, (Uttarakhand), India

ABSTRACT: In recent years, the rapid development of Internet, Internet of Things, and Cloud Computing have led to the explosive growth of data in almost every industry and business area. Big data has rapidly developed into a hot topic that attracts extensive attention from academia, industry, and governments around the world. In this position paper, we first briefly introduce the concept of big data, including its definition, features, and value. We describe the core concept behind the big data and framework required to handle big data with various software components. We then identify from different perspectives the significance and challenges of big data. We describe the grand security challenges, as well as possible solutions to address these challenges. Finally, we conclude the paper by presenting several suggestions on carrying out big data projects.

Keywords: Big data, Data complexity, Computational complexity, System complexity, Hadoop, map-reduce.

I. INTRODUCTION

This is the era of digital media. The huge records and information is collecting day to day. We all using digital devices connected to the web servers and producing new real time data every time. This data becomes large and increasing along the time. When we collect the data from all the different resources it becomes the unstructured and unmanageable with the standard DBMS tools and normal system hardware [1]. To handle data in terabytes we use map reduce, and Hadoop distributed file system (HDFS). The big data tool handles all the challenges that a traditional relational database can't handle. In section II we define the kind of data is called big data. In section III we discuss the Framework for Big Data and in section IV we discuss the significances of Big Data and section V shows grand security challenges and issues in big data and in section VI we conclude the paper with the suggestions to carry out big data projects and further research areas.

II. BIG DATA

Mainly the big data is defined by 3Vs (Volume, Variety and Velocity) but in many papers it defined by 5Vs. Big Data is characterized by the IBM [1] as the following:
Volume: The volume of data in petabytes to zetta bytes is considered as big data. Because such amount of data is not processed by traditional data management system.
Variety: The data may be relational and non-relational and includes Text, Audio, Video, Sensor Data, and Transaction Logs etc.

The 20 percent of the data in the world is structured and the 80 percent is semi structured or unstructured.
Velocity: "How fast the new data is generated?" In the big data environment the many terabytes of data generated every day. Big Data generation rate is increasing every year in the form of multiply.
The other characteristics of big data which are currently added are following:
Veracity: In the traditional database system we have certain data in structured form but in big data there are imprecise and uncertain data which is only handled by the big data tools. Uncertain data is the data in which we are not sure about the correctness of data and ambiguity and incomplete data.
Variability and Complexity: In big data there is not a fix rate of data generation, but high speed of data generation. The data is so complex because it is from the various sources and in structured, semi-structured format.
Value: The big data is data generated by various sources and may not contain high value in comparison with the volume. We obtain small valuable data after analyzing high volume data [6].The big data can also be classified by supervised, semi-supervised and unsupervised learning. In which classification and clustering is used for data mining purpose [9].

III. FRAMEWORK FOR BIGDATA

The Big data requires the separate framework to process and analyze the data. The most popular framework for big data is Hadoop framework.

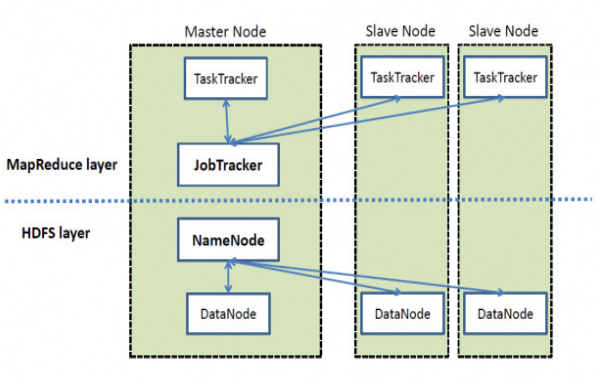

Fig. 1. High Level Architecture of Hadoop.

We will discuss about the Hadoop framework here in higher level as shown in fig.1. Hadoop framework use Google's map-reduce method to process the data on the number of nodes in parallel. The parallel processing of the big data the map-reduce is the most important method in big data to do work fast. The throughput of big data cluster depends on the number of nodes in the Hadoop Cluster. By increasing the number of nodes we can increase the throughput. In Hadoop framework the nodes are characterized in two types [8].

Master Node: Master node has the component NameNode and JobTracker in addition to component in slave node that's why it is called Master Node as shown in fig. 1.

NameNode: The NameNode is the node in Hadoop cluster or in the set of nodes which controls the all other nodes in the cluster. It works under the master node to manage the work and distribution of data processing across the cluster. It is responsible to perform map reduce task across the cluster. It stores the metadata about the DataNode such as name, file permissions, file modification, etc. so it keep track of all the files and replicated files and status of each slave node in the cluster. We can say that the Master Node is the controller of the all slave nodes.

JobTracker: JobTracker is a software daemon runs on the master node. In Hadoop cluster when a job is created it is submitted to the JobTracker. The work of job tracker is to find the required data across the cluster by communicating with NameNode.

Then it breaks the job into two different tasks: map and reduce task these task are assigned to those node on the cluster where the required data is available to do this task. Here the ***concept of data locality*** is applied to work efficiently. Because if there is not required data for the task then a node would have to send the huge amount of data across the cluster and this is extra work for the in case of big data where the data is in many terabytes. So the work is assigned to only that node that contains the sufficient data to do a particular task.

Slave Node: The Slave node is the node in which the data is stored in the blocks for processing on same node. Slave node contains two components:

DataNode: The DataNode is the node which works under the command from the name node and process data which he has and stores the result in its own storage.

TaskTracker: TaskTracker is a software daemon on the slave node. It monitors the status of each task. If any task is not completed due to any fault the TaskTracker informs to the JobTracker. Then JobTracker assign this task to other node in the cluster.

When we start Hadoop in single node cluster then it works in two parts DFS (data file system) and YARN. In DFS it contains DataNode Secondary NameNode, NameNode, jps and YARN contains node manager and resource manager. We can check that name node, DataNode are running by jps command which means Java virtual machine Process Status tool.

There are three modes [7] in which Hadoop run:

Standalone (local) mode: In this mode Hadoop runs on a single machine with its own file system and uses JVM installed on that machine. It is the smallest Hadoop environment. This environment is used to development of map reduce programs.

Pseudo-distributed mode: In this mode the Hadoop runs with its all demons to development, testing and quality assurance purpose on a single machine.

Fully distributed mode: When we have n number of nodes makes a Hadoop cluster is called fully distributed nodes. In which the one single node or host running as a name node and all other are running as a data node. But in the case of above two mode the name node and data node runs on the same node or host.

At the lower level Hadoop Framework is like the stack of software used to improve the processing power, management, data access, data storage of Hadoop and usability. There are so many product developed by Apache, Facebook, and other big data companies to Same as every product in Hadoop stack has to perform unique task. In Hadoop Framework there are so many products in which we classify all these in common class or type according to their functionality. There are the some common components in the stack which are commonly used are shown in fig. 2.

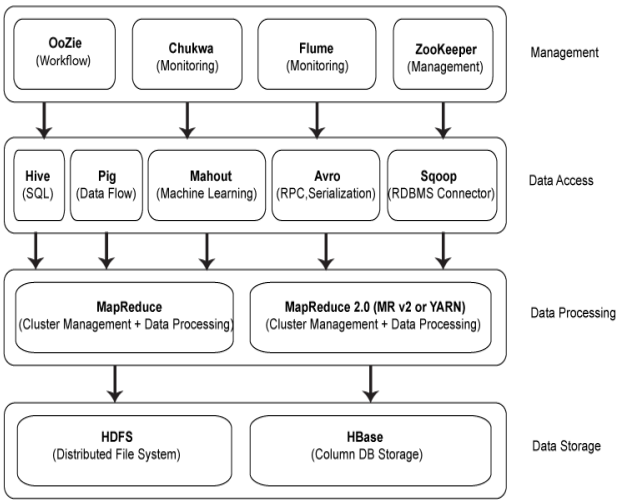

Fig. 2. Hadoop Ecosystem.

Map Reduce. The map reduce is the powerful tools in Hadoop which is used to work in distributed environment introduced by Dean and Ghemawat [15]. The map function extracts the values form the big data in key-value pair. The reduce stores the set of values for a particular key.

Hadoop Distributed File System (HDFS). The HDFS is the storage system in the Hadoop cluster. It is open source, reliable, highly scalable data storage and accessible by all nodes across the Hadoop cluster. It does not need any backup and fault tolerance because it automatically replicates same data to three nodes [5][16]. It also provides the high bandwidth to data transfer in the Hadoop cluster [20].

YARN. The YARN stands for Yet Another Resource Negotiator. It used to job scheduling and resource management across the Hadoop cluster. It is the one of important element of Hadoop Framework.

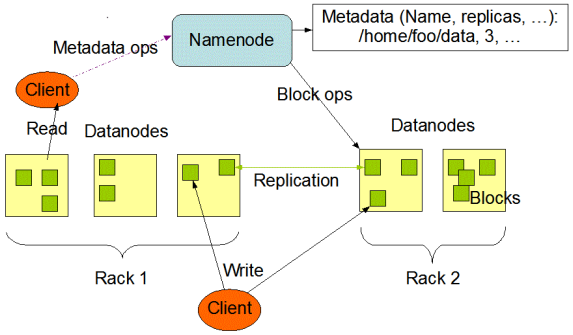

Fig. 3. HDFS Architecture.

Pig. It is the programming language used to write the map-reduce programs. It is initially developed by Yahoo. It can handle any kind of data. In Pig the Pig Latin language is used and it is a procedural scripting language.

Hive. It is an opened source data warehouse system initially developed by Facebook. It does three tasks on the large amount of data stored in Hadoop (HDFS or HBase) are summarization of data, query and analysis by using metadata.

It uses query language called HiveQL.

HBase. It is column-oriented, non-relational, distributed database management which runs on the top of HDFS. The application for HBase is written in Java Language same as map reduce. It is more flexible than relational database system.

ZooKeeper. The ZooKeeper is a centralized infrastructure and services in the Hadoop distributed Environment to provide cross node Synchronization, configuring information, name services, configuration management and store the state of entire system in a local log file. There may be multiple ZooKeeper servers to support large Hadoop Cluster.

OoZie. The OoZie is the Hadoop component which is used to create, execute, modify and coordinate the work flow of a map reduce job. The work flow in OoZie is defined by a Directed Acyclical Graph (DAG), Acyclical means there is not any loop in the graph.

Sqoop. It is an Extract, Load and Transform (ETL) tool to exchange the data between structured data sources and the Hadoop HDFS. It uses the comma separated values (CSV) file format to exchange data.

Avro. The Avro [18] is a data serialization system that provides the rich data structure with compact, fast and binary data format. It has a container file to store persistence data. It is also used in Remote Procedure call (RPC) in which the schema are shared between client and server.

Mahout. Mahout is used for creating machine learning application with the help of machine learning algorithms. The Mahout may be useful tool to enhance the big data usability for the machine learning.

Flume. The Flume [17] acts like a channel for transferring the data from sources to the Hadoop ecosystem. It has three entities called Source, Decorators and Sink. The Source is the source from where the data is coming to the Hadoop Ecosystem. The Sink is the target of data for specific operation and the decorators which decorate the data as by compressing or decompressing, modify, add or remove information during the data stream flow.

Chukwa. Chukwa [19] is the open source data collection system built on top of map-reduce and HDFS system and used for displaying, monitoring, analysing results to make better use of collected data. So it inherits the scalability and robustness of Hadoop.

IV. APPLICATION OF BIG DATA

A. To Generate Business Intelligence

The big data is used to extract the business intelligence .by analysing the big data we make the better decision for good business growth. It helps to understand the customer needs and most demanded products and better business choice. There are some business intelligence work are given below [5].

(i) To know customer requirements.
(ii) Analyse customer experience
(iii) To know maintenance requirements
(iv) Risk Prediction

B. Fraud Detection

All the data generated by the machine is more valuable to detect the suspicious task. The big data has two types of data: human generated and machine generated. The most of the data is machine generated like sensor data, logs, clicks etc. [2]

C. Dealing with Unstructured Data

Today's most of the data is unstructured. The unstructured data includes the text file, audio, video etc. This type of data is unmanageable and unprocessed by DBMS tools. By applying big data tools we can extract useful data and can manage it.

D. Reduction of Data Complexity:

There are many sources of data like websites, social media such as Facebook, Twitter, and LinkedIn. To maintain a useful record of all the data in multidimensional way is so complex. And if we use Hadoop and map-reduce technology then it's easy to store a worthy data into any secure storage and easily accessible when needed.

E. To Solve NP hard Problems

By using map reduce we can solve the NP hard problems like Max-k Cover [4] and Maximum Clique [3].The max -k Cover problem is to find the maximum number of connected nodes to broadcast an advertisement and to compute the single one result for a matching query in the web server or search engine instead finding same answer for each time for each user. The max clique problem is to find a sub graph from a given graph such that every vertex of that graph is adjacent to each vertex. The example of max clique is Facebook where each group is close and each friend knows everyone in that group. So both the problems are solved by using the map reduce method.

F. Research and Development

The research work needs to analyse the huge recorded sensory data and historical data in the field of space research and defence area. The satellite agencies like NASA, ESA and ISRO needs high computing environment, where big data framework Hadoop useful for handling many terabytes of data. In Indian Space Science Data Centre (ISSDC) established by ISRO have more than 150 servers and more than 250 terabytes of data across the servers [11]. The satellite sensor sends large volume of data to the ISSDC which is very difficult to store in traditional database system and needs big data framework to store, manipulate and extract useful information from that big data.

G. Big data in E-Governance

The Big Data tools are also used by many countries to make governance more accountable, transparent, and fast decision making with all citizen involvement resulted in corruption free governance.

H. Big Data in Healthcare

The use of big data in healthcare ecosystem provides the five pathways [12].

Right Living*:* By using big data in health care ecosystem the people choose the right life style, right exercise for fitness and right food. By choosing we can prevent from any particular disease

Right Care: In this case the patient will get optimal treatment which is already proved by ensuring safety of the patient.

Right Provider: In this we will get best professionals that needs patient for best outcomes.

Right Value: By using big data tools in healthcare's Electronic Media Records (EMR), we will get right treatment with less cost. That means reducing extra or unwanted treatment and quality is improved.

Right Innovation: By using the data stored in healthcare departments the researchers can discover new therapies for better and less costly treatment for any particulars disease and answer various clinical questions, using data acquired from the molecular, tissue, and patient levels of Health Informatics by using big data Technology [22].

V. SECURITY ISSUES IN BIG DATA

The organizations which have the big data to analysis are worried about its security. In the Hadoop cluster the sensitive data need to be secure in HDFS (stored state) and also in the data transmission from slave node to master node or data transfer among the Hadoop cluster [10]. There are some common security challenges in big data [12].

To Design Global Authentication Model

In Hadoop there requires to design the Authentication system for all types of users and the application and each type of user interface. In Big data platform there are many tools like HBase, Pig, YARN and Hive etc., and each tools has different authentication mechanism makes it more complex.

Securing Process on DataNode

The sensitive data is residing on the data node and any unauthorized user can access data blocks by creating a Task-tracker job and tasks on data node.

Access Control

There are the challenges in implementation of Attribute-Based Access Control (ABAC) and Role-Based Access Control.

Integration with existing Security Services

If any enterprise have its own security services then it is difficult to integrate adopt with its services.

Encryption of data in Transit and Rest

There is need to apply a better encryption and decryption technique to secure data in transit and at rest.

Network Security in Hadoop Cluster

In Hadoop Cluster there is a distributed environment for parallel computing and sensitive data across the network which needs to apply a tight security in the network.

Security issues in Cloud Computing Environment

The many challenges and issues solved if we apply modern cloud based infrastructure that provides a basis for building interoperable data or project centric SDI using modern technologies and best practices [21]. But as per the modern requirement big data technology needs the cloud computing environment where the security issues in cloud computing also a big issue for the Big Data analysis by using cloud services.

VI. CONCLUSION

In this paper we have discussed the big data and need to use big data in the real world environment for enhancing the quality of business, academia, healthcare and research areas. And also discussed the application of big data with the security challenges for managing the big data and protect the sensitive data across the Hadoop cluster. There are so many tools and techniques to support the big data and there are different tools are used by companies but some tools are common. And there are not any standard model for big data framework, this makes it more complex and needs more attention on the configuration and management. There is need to reduce the task on the Hadoop cluster for make resource available for high scalability. There is many research areas related to the using secure cloud services which overcomes many challenges of big data. The use of cloud computing and distributed system is requires to handle big data. And it is need of current century to implement the common framework for all areas which are using the big data tools for innovation and research.

REFERENCES

[1]. Zikopoulos, Chris, Thomas, Lapis, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", McGraw-Hill, ISBN 978-0-07-179053-6.

[2]. Danyel, Rob, Mary, "Interactions with Big Data Analytics",7th ACM interactions, may -June 2012.

[3]. Jongsawat, N., Premchaiswadi, W., "Solving the NP-hard computational problem in Bayesian networks using apache hadoop MapReduce", *11th International Conference on ICT and Knowledge Engineering (ICT&KE),* November, 2013, Pages 1-5.

[4]. Chierichetti, Ravi, Andrew, Tomkins "Max-Cover in Map-Reduce", *Proceedings of the 19th international conference on World wide web (WWW 2010),Raleigh*, North Carolina, USA, April 26–30, 2010, Pages 231-240.

[5]. Oracle, "An Enterprise Architect's Guide to Big Data", April 2015.

[6]. Gandomi, Haider, “Beyond the hype: Big data concepts, methods, and analytics”, *International Journal of Information Management* **35** (2015) p. 137–144.
[7]. http://www.edureka.co/blog/hadoop-interview-questions-hadoop-cluster/
[8]. Murthy, Padmakar, Reddy, “Hadoop Architecture and its Functionality”, *IJCSIS* Vol. **13**, No. 4, April 2015.
[9]. N. Khan, M.S Husain, M. R. Beg, “Big Data Classification using Evolutionary Techniques: A Survey”, *IEEE International Conference on Engineering and Technology (ICETECH),* 20 March 2015.
[10]. M. RezaeiJam, L. M. Khanli, M. K. Akbari, “A Survey on Security of Hadoop”, *4th IEEE International Conference on Computer and Knowledge Engineering(ICCKE),* 2014, p.716-721
[11]. J. D. Rao, “ISRO Telemetry, Tracking and Command Network”, Banglore, Link: http://nkn.in/nkn-workshop2013/images/presentation/NKN_ISSDC-V1-17-10-13-v4.pdf
[12]. K. T. Smith “Big Data Security: The Evolution of Hadoop's Security Model” August 14, 2013, Link: http://www.infoq.com/articles/HadoopSecurityModel.
[13]. J. Wiener, N. Bronson "Facebook’s Top Open Data Problems", 16 September 2014 Link:https://research.facebook.com/blog/1522692927972019/facebook-s-top-open-data-problems/
[14]. IBM, "Performance and Capacity Implications for Big Data", January 2014.
[15]. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Proc. Operating Systems Design and Implementation (OSDI), pp. 137-150, 2004.
[16]. https://hadoop.apache.org
[17]. http://www-01.ibm.com/software/data/infosphere/hadoop/
[18]. http://avro.apache.org/docs/current/
[19]. http://chukwa.apache.org/
[20]. V. Chavan, N. Phursule et al, “Survey Paper on Big Data”, *International Journal of Computer Science and Information Technologies,* Vol. **5** (6), 2014, 7932-7939.
[21]. Y. Demchenko Z. Zhao P. Grosso A. Wibisono and C. De Laat "Addressing big data challenges for scientific data infrastructure" in *IEEE 4th International Conference on Cloud Computing Technology and Science* (CloudCom 2012) 2012 pp. 614-617.
[22]. M. Herland T. M. Khoshgoftaar R. Wald "Survey of clinical data mining applications on big data in health informatics", *Proceedings of the 2013 12th International Conference on Machine Learning and Applications*, vol. **02** pp. 465-472 2013.