ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

A Study on Big Data: Concepts, Architecture, and Technologies

Priyanka Rana

Assistant Professor, Department of Computer Science, KC Group of Institutions, Una, India

ABSTRACT: The term 'Big data' refers to the large volumes of structured and unstructured data or the complex data sets that cannot be handled by using a traditional data processing approach. While working with Big Data, it's not the amount of data that matters but the quality of information that can be extracted from the database. In an organization, 'Big data' is evaluated for insights that direct to better strategic decisions. Advanced data analytics techniques like predictive analytics, location intelligence, and data mining are used to process hundreds of terabytes of data for financial decision making or business informatics. To manage these large data sets (called 'Big data') Hadoop and MapReduce can be used. Hadoop is an open source framework that follows the distributed computing and parallel processing approach for the efficient and cost- effective processing of data sets. Another feature of Hadoop Model that is beneficial for 'big data' is its scalability. It can scale up from a single server to large clusters of commodity servers, with a very high degree of fault tolerance.

KEYWORDS: Big data, Lambda architecture, Hadoop, HDFS

I. INTRODUCTION

A. *Big Data:*

Before the term 'Big data' came into existence, a large amount of data was handled by using traditional approaches like Relational database management systems, data warehouses, and visualization-packages. Advancement in technology leads to the tremendous increase in the data sets and their storage media. Like, nowadays, we use mobile devices, radio-frequency identification (RFID) readers, cameras, wireless sensor networks, software logs, and aerial (remote sensing) [1]. Though large data sets provide more potential to study key insights, service, and products [2] but it's only possible if data curation of business information is efficiently implemented, which is almost impossible with traditional data processing techniques. In 2001, Gartner.Inc analyst Doug Laney cited the fundamental definition of 'Big data' as three V's:

Volume: It refers to the size of the data. From the last decade, there has been an exponential growth in the data because of the rapid increase in the data generation and storage devices. Business organizations collect terabytes or petabytes of data from various sources like social media, business transactions, sensor networks or machine-to-machine data transfer [2] and the size of this data determines whether it can actually be considered as 'Big Data' or not [1].

Velocity: Velocity refers to the speed at which data is processed. The growth of data has been unprecedented in the last few years and therefore it is imperative for the business organizations to handle the data with the same frequency to meet the demands and expectations of their clients. The data transfer is almost real-time nowadays [3] and the validity of data have also reduced to shorter intervals [2]. For example, on social media sites (twitter, facebook) sometimes a few minutes old updates are discarded and users start paying attention to the recent updates post within few seconds [3].

Variety: Variety implies multiple formats in which data can be stored. Different types of data like structured, semi-structured and unstructured can be stored in various file formats like audio, video, image, text and much more. New formats are coming to life with the development of the new applications [4]. The companies have to organize the data available in numerous formats to extract the meaningful information from it. This is the challenge the organizations struggle to over with 'Big data' tools [3].

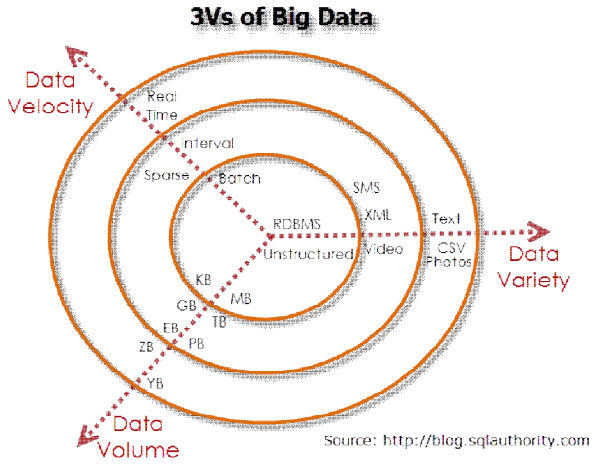
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017



With the passage of time, the enormous increase in the data leads to the addition of 2 V's in the definition of 'Big data'. They are:

Variability: Variability implies that the meaning of the same data varies according to the context in which it is used. Variability and Variety are two different concepts. For example, in the social media application, say twitter, a thread of tweets related to a particular topic is variety and usage of a particular word in different contexts in the same thread is variability. The concept of variability is prominently used in Sentiment Analysis [5]. Algorithms are written that are able to understand the context in which a particular word is used and are able to decode the exact meaning of the word in that context. This is a very difficult task and a lot of work is required to be done in this field of natural language processing.

Veracity: The quality of the stored data varies largely as it comes from different sources. Therefore, it becomes difficult for the organizations to clean, correlate and to link multiple data sets to do accurate analysis and gain profitable results. The veracity of data can exist due to various reasons like noise, biases, and abnormality in data [6]. Gathering large volumes of data is worthless if it is not correct, affecting both the organization and consumers. Hence, it's the need of the organizations to ensure the integrity of the data to obtain the accurate analysis.

B. ***Lambda Architecture for Big Data:***

Lambda architecture consists of three key layers through which raw or unstructured data is processed to obtain significant information. Lambda architecture provides a generic, fault-tolerant, scalable and robust data processing system to carry on a big data project successfully [7]. It uses both batch and real-time stream processing to provide ample and accurate views of batch and online data respectively [8].

Apache Hadoop can be used to implement lambda architecture. This framework is used to store and process large data sets and optimization techniques are used to handle queries. The data entering the system is distributed among batch layer and speed layer.

Batch layer: Batch layer uses distributed processing system to handle big data. It manages the master data set, pre computes the batch views and fixes the errors in the existing data. In the case of Hadoop framework, only append operation can be performed on master data set stored in Hadoop File System (HDFS) and arbitrary batch views are computed using MapReduce program [9]. Finally, the output is stored in a read-only database.

Speed layer: Unlike batch layer, the speed layer processes the recently available data and increments the real-time views respectively. It sacrifices throughput to compensate the high latency of data by the batch layer [7]. The real-time views provided by the speed layer are not as accurate as the batch views but they provide time constraint or immediate information as and when required. It provides modular designing and is useful in the applications where data transfer rate is more critical as compared to the reliability of the data, though, we can always replace these real-time views with the batch views as needed. Lambda architecture applies "complexity isolation" [10] as the complex data is shifted to

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

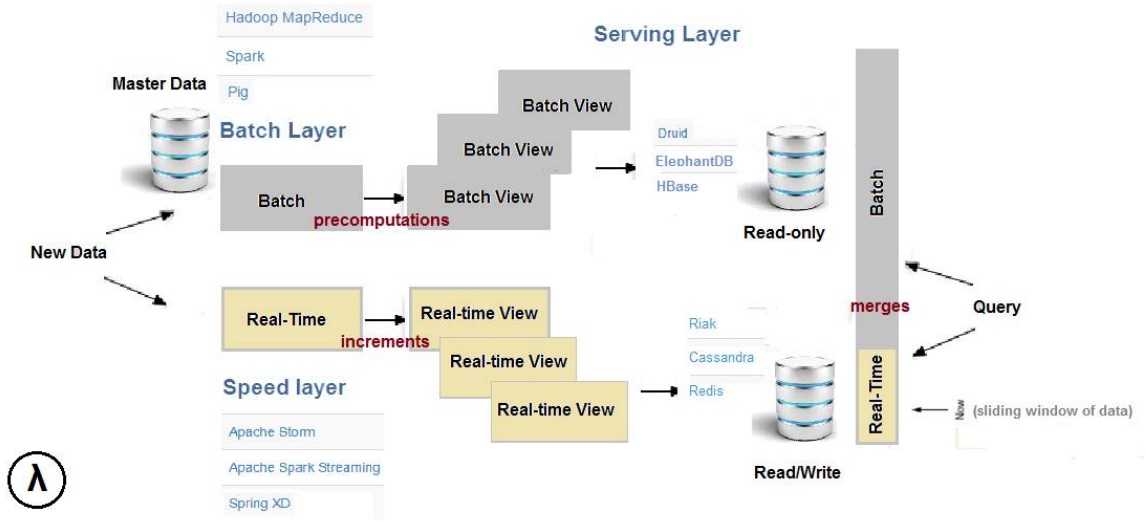
(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

the temporary layers. The technologies used in this layer include Apache Spark, Apache Storm and SQL Stream and the output is stored on NoSQL databases [8].

Serving layer: The serving layer stores and subsequently indexes the data or the output received from the batch layer and the speed layer. When a query is invoked by the user, the serving layer responds correspondingly either by displaying the views as such or processing these views to provide appropriate results. The technologies used in the serving layer are, Druid (handles output from batch and speed layer), Apache Cassandra or HBase (as storage media) and Elephant DB or Cloudera Impala (as batch layer output) [8].



II. BIG DATA SURVEY OVERVIEW

Industrial survey report: According to the survey report, 2014, published by Forbes.com, 87% of organizations believe that within the next three years, 'Big data' will play a vital role in redefining the strategies of their industries for competitive growth. According to Wikibon, it is anticipated that by 2020, enterprises worldwide will be spending nearly $500 B on Industrial Internet (including software, hardware, and services). In 2015, Industrial Internet Insights Report concluded that 73% of the companies invest more than 20% of their technology budget on 'Big data'. The 'Big data' analytics become prominent in the sectors like the wind (45%), aviation (61%), and manufacturing (42%) companies. The following graph shows the descending level of importance of 'Big data' analytics in various sectors.

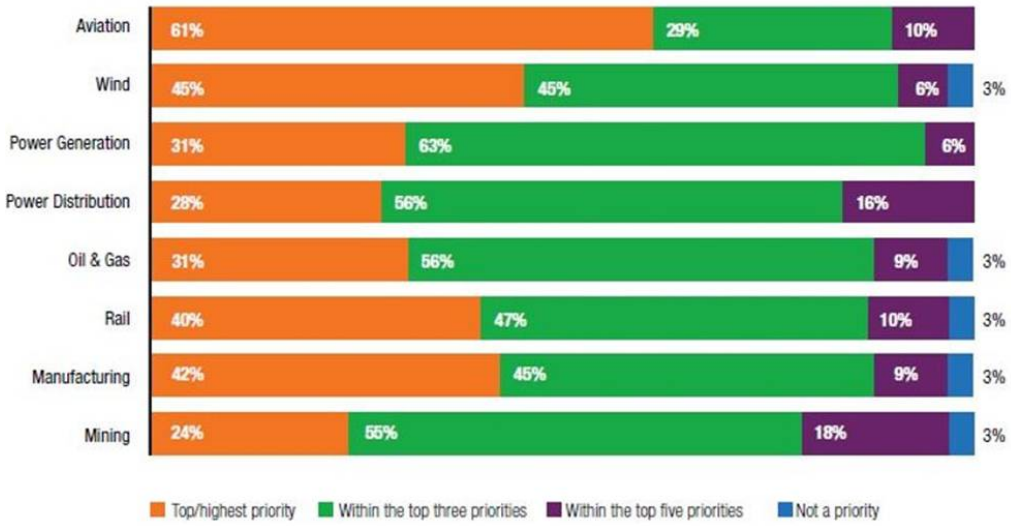
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

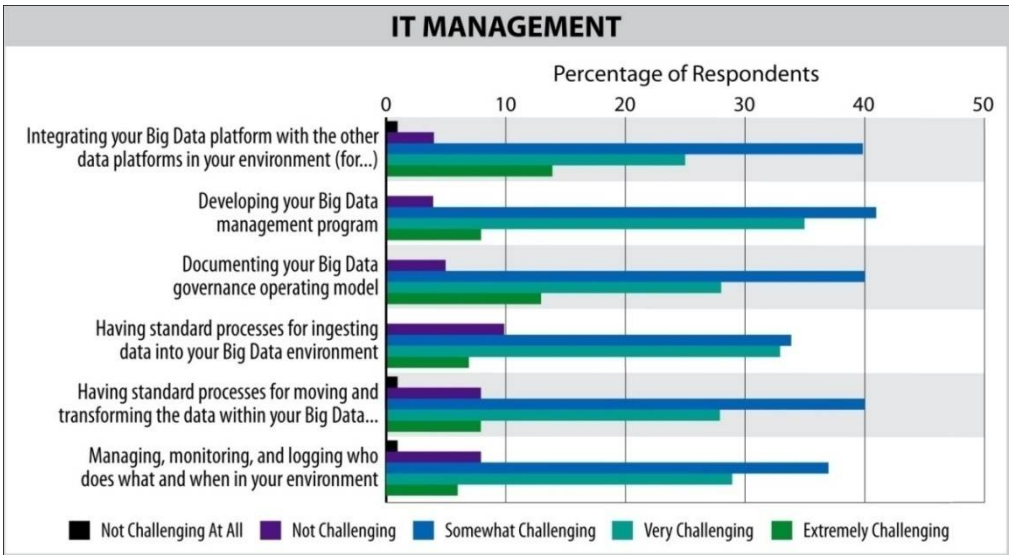
Website: www.ijircce.com

Vol. 5, Issue 4, April 2017



According to Gartner Survey,2012, almost 58% of the enterprises are already investing or intend to invest in 'Big data' analytics in the next two years but according to a similar survey conducted in 2015, there was a considerable increase of more than 75% in the number of such companies [12]. Another survey was conducted by The Knowledgent from 2012 to 2015, namely, 'The Big data survey' targeting the IT practitioners having knowledge about big data [12]. The key findings were:

- On average, 75% of the respondents believe that managing and implementing Big data framework is somewhat challenging.
- The most difficult part is to develop the overall program.
- The simplest task is to control access and authentications.



A.R. Guess et al. Teena Maddox quoted that according to Tech Pro survey, 2016, 'Big data' implementation largely depends on the size of the organization [13]. The companies with the work force of at least 1000 employees have implemented big data nearly 1.5 times more often than small scale companies. Also, almost one-fifth of small scale companies are using big data tools and three-quarters of these companies are not using 'Big data' solution. Further, she

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

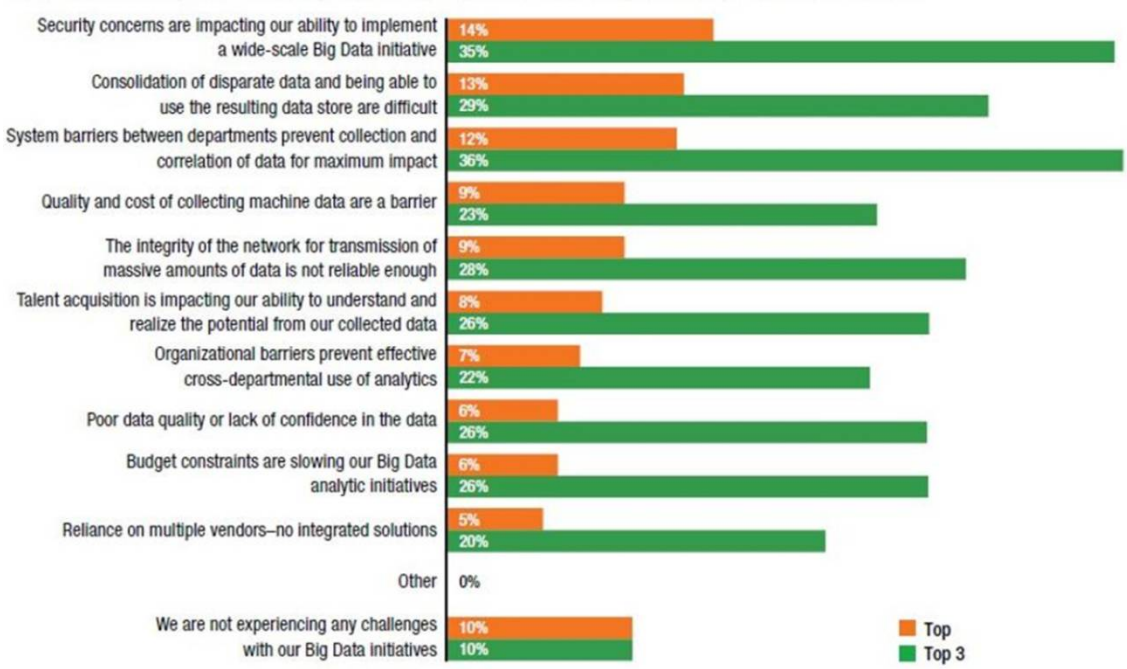
stated that Asia-Pacific region has undertaken big data implementation at a large scale followed by Europe which is second highest in the ranking. On the contrary, big data is rarely implemented in Central/South America.

Limitations in the implementation of Big Data:

According to the survey report, 2014, published by Forbes, the main challenges enterprises are facing main in implementing 'Big data' are [11]:

- The cluster of heterogeneous systems established between different departments barriers the collection and correlation of the data for maximum insights (36%).
- The enterprises have security concerns regarding the implementation of 'Big data' on a large scale, thus, preventing them from taking initiative (35%).
- Storing and categorizing data available in different formats and using the resulting data store (29%).

The following graph represents the challenges in implementation states by different companies:



According to Gartner Survey, 2015, almost 55% of the respondents admitted that the biggest challenge in the implementation of 'Big Data' is finding the experienced resources with specific skills. On the contrary, the easiest task is to gain buy-in from business stakeholders. According to Tech Pro survey, 2016, one of the most challenging tasks is to find skilled people to handle the big data as shown below [12]:

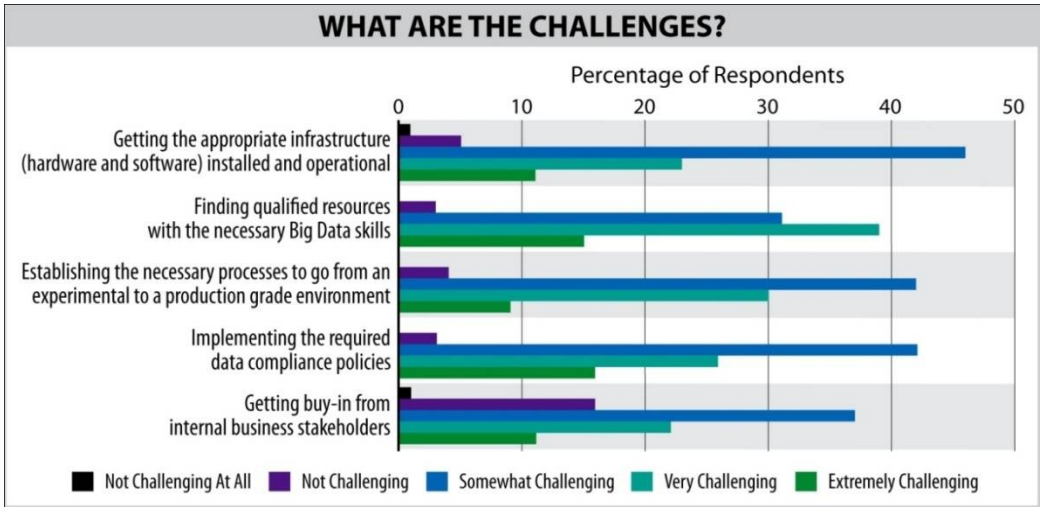
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017



'Big data' consists of piles of unsorted data including the web traffic log, location information, and financial data and also spam data. Companies have to employ specialists to sort and clean-up the data and to extract quality data efficiently. Segregating the relevant data or valuable information is the crucial step for the companies in the process of formulating profitable strategies. Sometimes, the companies have to reinvent the big data tools to store and analyze the various types and format of data sets. This is one of the most time-consuming process and the costs are directly relative to the volume of data. In addition to this, most of the companies try to sort and analyze the data by using less skilled employees, which leads to inaccurate data gathering and faulty results. Moreover, data privacy and security is one of the major concerns among the clients for implementing 'Big data'. As most of the companies use the public cloud to store their data, therefore, the sensitive or important information like perspective projects, financial data, and clients list of the company becomes more prone to the security risks. The complexity of software and hardware infrastructure is directly proportional to the complexity of data [20]. As the companies can't afford the loss of data so they have to spend heavily on the management of the data as the cost of the implementation increases with the increase in the volumes of data. Another challenging issue for the companies is to maintain the continuity of service to the clients and uninterrupted processing of data even during system malfunctioning [21].

III. HADOOP: SOLUTION FOR BIG DATA PROCESSING

The basic version of Hadoop was developed by Doug Cutting in 2004 and in few years it had become a top level software foundation project. Apache Hadoop is an open source programming framework basically written in Java used for distributed processing of 'Big data' across the Linux clusters of computer systems. It is implemented by the whopping companies like Google, Facebook, and Amazon. Hadoop ecosystem consists of a number of components like Hadoop Distributed File system (HDFS), MapReduce, Avro, Hbases, Hive, Zookeeper, Pig, Sqoop, and Mahout. The main advantages of using Hadoop framework are its low-cost resourceful scheduling technique and fault tolerance; as it continues job even if one node is crashed without losing data or disrupting work, by shifting it to the remaining machines in the cluster. It also provides an efficient way for implementing cloud storage system.

Hadoop Distributed File System (HDFS): HDFS is based on the Master/Slave architecture. It consists of ordinary hardware like an interconnected cluster of nodes where we store our directories and files. It helps in storing huge amounts of information by scaling up incrementally as and when required. HDFS stores three copies of each file to three different servers. The architecture of HDFS cluster consists of mainly three nodes: Name Node, Data Node, and HDFS Clients or Edge node.

Name Node or master node is a central node which contains all the information (location, attributes etc.) about the data blocks in each node. Moreover, it also provides information about the newly added, updated and removed data nodes and its corresponding data blocks.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

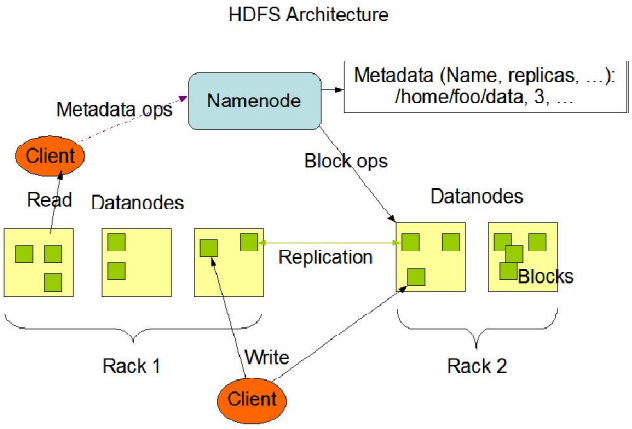
(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Data Node or slave node are the nodes where data as blocks are stored within a file and can be hold or fetched as and when requested by the user for read and write operations. These data nodes can be scale up or down according to the work load, capacity or performance requirements. The two main functions of data node are to store a block in HDFS and to provide a platform for running jobs.
HDFS Clients/Edge node acts as a linker between name node and data nodes.

The following figure defines the HDFS Architecture [17]:



IV. TECHNOLOGIES FOR BIG DATA

MapReduce: MapReduce is the most common parallel and distributed programming technique used to handle Big Data [14]. One of the main characteristics of MapReduce is its adaptability to work with different data sources and its ability to assemble multiple systems for large scale processing. MapReduce consists of two procedures: map() and reduce(). Map() inputs the key/value pair to give a set of intermediate key/value pairs. These transformed records thus obtained need not be of the same type as the input records. An input pair can map from zero to many output pairs [18]. The two basic tasks of MapReduce can be represented as [14]:

Map: (k1, v1) list (k2, v2)

Reduce: (k2, list (v2)) list (v2)

The intermediate and processed data are kept in HDFS (Hadoop Distributed File System).

Hive: It is a data warehousing framework built on the top of Hadoop. Developed by the Facebook, it provides the relational model and SQL interface that allows the applications to analyze the data, summarized it and run the queries. It is easy to use and understand. The disadvantage of using Hive is high latency on executed queries.

HBase: HBase is an open source non-relational (NoSQL) database that runs on the top of HDFS. It is used to handle Big Data by providing real- time access to read/write operations on the large data sets. HBase works seamlessly along with Hadoop compatibly and helps in combining data sources of different structures and schemas. Basically, it provides input and output to the MapReduce [19].

Pig: Pig provides a platform for analyzing and processing large data sets. It adds additional level abstraction in data processing, thus making the writing and maintaining of the data processing jobs very easy [19]. It is an execution framework for parallel processing tera bytes of data with few lines of code and acts as a bridge between SQL and MapReduce. Therefore, the reuse of the intermediate results saves execution time and increases the efficiency of the system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Sqoop: It is a command line interface used to import the data from relational databases like oracle and MySQL to Hadoop environment and export the data from Hadoop file system to the database servers. It is a tool that allows transfer of data and interaction between Hadoop environment and database servers [19].

Zookeeper: Zookeeper is an open source centralized distributed synchronization and governing service for Hadoop framework. It also maintains the distributed configuration service which is useful for tracking the working of a particular node [19]. The performance of Zookeeper is high with read operations are more as compared to write operations.

Mahout: Mahout is an open source library for producing machine-learning algorithms and data mining. It implements four main machine learning techniques: collective filtering, classification, clustering, and mining of parallel frequent patterns [19]. As it is compatible to work in the distributed environment, we can use Hadoop library to scale the cloud effectively.

V. CONCLUSION

The paper highlights the basic concepts of Big Data and Hadoop architecture along with the problems in its implementation. It also focuses on the storage, maintenance, and processing of the big data. Though big data technology has come a long way but still we face the challenge to overcome some of the basic technical issues in its implementation like security, scaling, flexibility in handling heterogeneous data, error handling, performance optimization and lack of structure. These are the common issues across a large variety of application domains so it's not competent to focus in the context of a single domain only. Companies are concentrating on finding out a more consistent solution for handling the ever growing large and complex data sets. Though open source software are used in big data, still a lot of time and money is spend on hiring specialist to handle the incoming data and to extract valuable information from it. Hadoop framework is a proficient solution for big data but a lot of work is still required to make this technique more constructive and advantageous to the business and the scientific world.

REFERENCES

1. Big data available at: https://en.wikipedia.org/wiki/Big_data
2. Concepts of Big data available at: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html.
3. 3 V's of Big data available at: http://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data- volume-velocity-and-variety-day-2-of-21/.
4. Big data definition available at: http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data
5. Big data available at: https://datafloq.com/read/3vs-sufficient-describe-big-data/166
6. Big data available at: http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/
7. Lambda Architecture available at: http://lambda-architecture.net/
8. Lambda Architecture available at: https://en.wikipedia.org/wiki/Lambda_architecture
9. Big data architecture available at: http://www.datasciencecentral.com/profiles/blogs/lambda-architecture-for-big-data-systems
10. Lambda architecture available at: https://tsicilian.wordpress.com/2015/01/05/lambda-architecture-for-big-data/
11. Big data analysis available at: http://www.forbes.com/sites/louiscolumbus/2014/10/19/84-of-enterprises-see-big-data-analytics-changing-their-industries-competitive-landscahjpes-in-the-next-year/#20498ff32502
12. Big data challenges available at: https://knowledgent.com/whitepaper/2015-big-data-survey-current-implementation-challenges/
13. Big data solutions available at: http://www.dataversity.net/survey-finds-49-of-large-companies-implementing-big-data-solutions/
14. Cloud Computing: A Comprehensive View, Priyanka Thakur and Dr. Pawan Thakur, International Journal of Electrical, Electronics and Computer Engineering
15. Hbase available at: https://hortonworks.com/apache/hbase/
16. Sqoop available at: https://www.tutorialspoint.com/sqoop/
17. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, International Journal of Scientific and Research Publications, Volume 4, "A Review Paper on Big Data and Hadoop"
18. Hadoop architecture available at: http://hadoop.apache.org/docs/r2.2.0/api/org/apache/hadoop/mapreduce/package-summary.html
19. Varsha B.Bobade, International Research Journal of Engineering and Technology, Survey Paper on Big Data and Hadoop
20. Chen Zhang, Hans De Sterck, Ashraf Aboulnaga, Haig Djambazian, and Rob Sladek, Case Study of Scientific Data Processing on a Cloud Using Hadoop
21. Alexandru Adrian TOLE, Big Data Challenges

BIOGRAPHY

Priyanka Rana is an Assistant Professor in the KC Group of Institutions, Una. She received Master of Computer Application (MCA) degree in 2010 from the Institute of Engineering & Technology, Bhaddal, Punjab, India. She is currently working on a project in Big Data and Hadoop. She has published a research paper in International Journal titled "Cloud Computing: A Comprehensive View".