A Technical Report on Big Data and Its Challenges

Rajesh Kumar Kaushal
Chitkara University Institute of Engineering and Technology,
Chitkara University, Punjab, India

Naveen Kumar*
Chitkara University Institute of Engineering and Technology,
Chitkara University, Punjab, India

Surya Narayan Panda
Chitkara University Institute of Engineering and Technology
Chitkara University, Punjab, India

Abstract-During the last decade, there has been immense growth in the amount of data. Companies like Google, Facebook, Instagram and Pinterest are gathering huge amounts of data and this data is utilized for data analytics using different tools and technologies. As the amount of data grows, there is a deficiency of techniques and tools that can handle it as the development of these techniques or tools is not fast enough to match the growth of data. Analyzing Big Data is not that easy and there are various challenges that are incurred when dealing with Big Data. Hadoop is a framework which is used widely to analyze Big Data. It is fault tolerant and provides scalability and parallel processing environment. This paper discusses the characteristics of Big Data, Hadoop and its architecture, Hadoop Distributed File System i.e. HDFS and its various technologies, the working of Map-Reduce. The challenges presented by Big Data have been discussed in detail. The present work found that security, analysis, distributed storage and staging are the key challenges due to rapid growth of data and these challenges need to be addressed in the future.

Keywords – Big Data, Hadoop, Map-Reduce, Big Data Challenges

I. INTRODUCTION

As the human race embraces the digital world that we live in today, huge amounts of data is being generated every second by millions of users throughout the world. Estimated 4.4 zettabytes of data is generated in a year and there is estimation that this amount will be increased to 44 zettabytes by the end of year 2020. Facebook, Instagram, YouTube and other social media websites generate almost half the data of the amount specified above. Around 267 million transactions are recorded per day by Walmart from the sales at their stores worldwide [1]. A total of 3.5 billion requests are processed by Google per day and they store 10 Exabyte of data. The data is expected to grow more due to increasing trend of the Internet of Things (IoT). The International Data Corporation (IDC) predicts that there will be 28 billion sensors by the end of 2020. The analytical market is growing by 10.3% annually [2]. This huge amount of data, which may consist of both structured and unstructured data, is known as Big data. The Big data is characterized by its high volume, variety and velocity. This data is very large and in unstructured format and it is growing rapidly. The traditional tools and resources that were used to handle smaller amounts of data cannot be used for handling Big data. A specialized tools are required for processing and storing such huge and complex data as Big data finds applications in fields like medicine, health and engineering. So, the demand for processing huge amounts of data and to find fine grained information through analysis is the need of the hour for the companies [3]. These tools are necessary to work on large datasets, in order to mine those datasets and perform data pre-processing on them which will give us analyzed results.

Big data is very large and has high dimensionality and these two features lead to accumulation of noise, spurious correlations and algorithmic instability which need to be catered by better tools that are designed in order to handle such complexities [4].

To analyze this data, Hadoop MapReduce is used extensively. Hadoop is an open source, Java based programming framework and it is a part of Apache Project. The MapReduce framework is used to provide parallelization and scalability during the processing of Big Data. This framework consists of a Distributed File System (DFS) where the map and reduce functions are applied on data which is in the form of a key-value pair. These functions may run a single time or as many times as wanted.

Big data is a collection of huge amount of data that consists of either structured or unstructured data or both. This data, due to its sheer size, is difficult to store, manage, process and analyze. The complexity of these datasets is high so the usual methods of analyzing data cannot be applied on these datasets.

The term “Big Data” was used by John Mashey in 1990’s and then it was defined by Gartner using the 3 Vs which are Volume, Velocity and Variety. Gartner defines Big Data as Big Data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.” We can define Big Data as large amount of data that can be analyzed using different tools and techniques, in order to find important patterns that helps us to gain knowledge for strategic planning and have a better insight about the data.

As mentioned above, Gartner proposed a definition of Big Data based on the three characteristics of Big Data called 3 Vs. But in the recent years, a wider definition of Big Data is becoming popular and it consists of 5Vs namely- Variety, Velocity, Volume, Veracity and Value. The characteristics of Big Data have been discussed below [5-8].

Volume is the data that is collected in huge amounts from various sources. The growth of data has been exponential over the years and the amount of data that is collected today is in zettabytes. Variety refers to different types of data that is generated and then collected from various sources is the reason for this characteristic. As the domain of data collection widens, the varieties of data increase. The data collected may be structured, unstructured or a mix of both. Velocity is the speed with which the data is generated defines this characteristic. The processing speed should be fast enough for this data as the data becomes historical at a fast rate and it also becomes historical due to this evolution. In case the data being analyzed is inaccurate, then the results obtained by it will be inaccurate too. Veracity pertains to validation of data which indicates that whether the data that is being processed is reliable or not. Value can be defined as those hidden patterns or relationships that ought to be analyzed to gain a proper insight about the dataset. A bigger dataset means that it will take more time to establish a relationship in the different attributes or values. So, value is closely related to data volume and variety. The other two Vs that are defined these days are visualization and variability and are described below. Data needs to be in a form such that it is easy to read and understand it and this is known as visualization. The value of data keeps on changing due to high evolution speed i.e. there is variance in the data then this is known as variability [9].

Because Big Data has huge volume and has various varieties of data, so it has some classes like storing, sourcing, formatting, staging and processing. Sourcing consists of various data sources from which data is collected. Format consists of different types of data. Storing and organization consists of various ways of storing data and extracting, transforming and loading processes. There are large number of applications of Big Data. These are described below.

1. Analytical tools for E-Science Data Management

   The data that has been published should be accessible to the fellow scientists and it should also be secure enough. The paradigm change in modern E- Science has helped in achieving this [5]. Quite a lot of data has been automated and re-purposed to carry out researches in different scientific domains.

2. Tools for improvement in healthcare facilities

   Many researches are underway in the Healthcare Industry and Big Data is being used in order to analyze the general patterns in large datasets [10]. It is being used in researches related to DNA and genomes [6][11]. Bigger healthcare facilities are analyzing the data collected from numerous patients in order to gain a detailed insight into the data. The smart fitness devices like bands and watches also help us in monitoring our heart beat, fitness level and calories by analyzing the collected data for better lifestyle choices.

3. Optimizing Business performance

   Walmart acquired Inkiru Inc. and this has helped Walmart in increasing sales as it analyzes trends before launching a product, provides a personalized experience to each user by customized recommendations for products. It also analyzes the sales of various products according to regions or areas.

4. To analyze user demands and preferences

   In 2012, retail giant Walmart started using a 250 node cluster and as it started analyzing Big Data to make decisions and to analyze customer preferences, and this led to a considerable increase in the online sales of Walmart. As the data needs to be processed in real time, so Walmart uses Mupd8 (Map Update

Application) in order to handle faster data. Mupd8 is able to tackle huge load and data-distribution across multiple CPU cores.

5. Tools for optimizing performance in sports

   Germany’s big win in FIFA World Cup 2014 was due to its partnership with SAP wherein they analyzed huge amounts of player performance data using a tool called match insights and helped every player to work on his weaknesses [12]. This helped in optimizing the performance of the players.

6. Improvement in Science and research facilities

   NASA generates huge amounts of data because of the numerous ongoing missions. This data is then used to generate meaningful knowledge which leads to new findings and discoveries. The Large Hadron Collider at CERN, Geneva also generates data that can’t be handled by traditional techniques. So a specially designed analytical tool ROOT is used in order to analyze this complex data that is related to particles and their collisions.

The present paper discusses the MapReduce framework based on Hadoop along with the challenges presented by Big Data. The paper is arranged as follows: The section II gives the literature review and the section V describes the architecture of Hadoop followed by description of HDFS architecture in section VI. The MapReduce is described in section VII followed by performance analysis of Hadoop in section VIII. The section IX and X lists down the challenges and open research challenges respectively.

II. LITERATURE REVIEW

Bryant et al., [1] focused on how advances in different technologies like sensors, networks, data storage, computer clusters and data algorithms had led to the rise of big data computing. With an increase in its popularity, certain challenges were faced in its implementation like cluster programming, extending the reach of cloud computing, data analysis techniques, security and privacy. The paper also elaborated long term and short term actions that a company should take so as to implement big data computing in a successful way.

Abouzeid et al., [2] presented the idea that how parallel databases and MapReduce based systems did not meet certain properties that were desired for performing data analysis and these properties were performance, fault-tolerance, flexible query interface and ability to run on heterogeneous systems. The authors also discussed shortfalls of present data analysis approaches. They also described the HadoopDB wherein the implementation of Hadoop background and its architecture was also discussed. The benchmarked systems were also mentioned along with summarized results.

Some researchers elaborated the definition of big data [3] and summarized how it is used by big companies and presented how to collect useful data by using cloud computing techniques. Thereafter the authors focussed on the challenges presented in this field like defining the constructs for access control and auditing in terms of data platforms, devising a querying technique for approximate results, building an environment that enables exploration of data for deep analytics, reconsidering optimization of query for data parallel platforms, defining a model for performance.

In another study, the autors shared the challenges in analysis of big data [4]. The authors started discussion with the two main advantages of big data which were finding a hidden pattern in big data as well as finding common features among a variety of individual features. Thereafter, the challenges in big data mining were mentioned which were spurious correlations, issues regarding heterogeneity and heavy computational costs plus algorithmic instability. To handle these challenges, various paradigm shifts were adopted. The authors also pointed on the characteristics of big data and the various experiments that were carried out. At last, the authors presented its statistical impact on data.

One of the studies [5] pointed out the challenges on the present and future Scientific Data Infrastructure (SDI). The authors defined requirements and different access control and security measures for management of data. Scientific Data Lifecycle Management (SDLM) was also introduced along with all the stages associated with it. The paradigm changes for big data and cloud based services for e-data were also included so as to provide an interoperable data for using it with modern technologies and for best practices.

In another study, the authors introduced 5 V’s of big data followed by description of technical view of big data which includes sourcing, storing, formatting, organizing, processing and querying [6]. The major research areas in big data were identified like application of ontology, security, storage and transport, accessibility, mobility and inconsistencies along with technical challenges that loom over big data namely failure handling, data heterogeneity and data quality.

Gandomi & Haider [9] emphasized on the neglected areas or dimensions of big data. The authors started the discussion with defining big data and then by defining its 3 V’s. The text analytics techniques like information extraction, summarization, question answering techniques and opinion mining followed by audio, video and social media analytics were also defined. It also described predictive analytics at the end.

Kumar et al., [11] presented the applications of big data for the digital India scheme. The authors of this paper also discussed 5 V’s and then stated some figures about the growing popularity of big data. It then discussed the various application areas like predictive model for understanding end users behavior and preference, optimization models for businesses or individual quantification etc. The authors then mentioned the challenges that usually occurs in different phases of big followed by the recommendations for handling these challenges.

In the study [13] the authors highlighted about big data and how Hadoop is used for analysis of unstructured datasets using HDFS and MapReduce processing. The supporting tools of Hadoop like core, MapReduce, HDFS, HBase, Pig, ZooKeeper, Hive, Chukwa were also discussed. The architecture of Hadoop which consists of data nodes, name nodes, resource manager, secondary name nodes and node manager were discussed in detail.

Patel, Birla, & Nair [14] primarily focused on defining big data and then mentioned that still there are problems in handling huge amounts of data using relational database or parallel processing with respect to scalability, unstructured data and fault tolerance. The automation techniques being applied to big data like Multiple Parallel Processing and data mining grids etc. were also mentioned. Hadoop-DFS and MapReduce were discussed along with the working of MapReduce and the architecture of Hadoop. After that, experimental setup had been stated along with the discussion on results.

Bhosale & Gadekar [15] defined big data and presented 3 V’s along with the problems related to big data processing which were like heterogeneity, incompleteness, scale, timeliness and privacy. Hadoop and its various components along with HDFS architecture and the MapReduce architecture were also discussed.

One of the studies [16] differentiated between SQL and NoSQL and then moved on to discuss the architecture of Hadoop along with HDFS and the MapReduce. Thereafter, the authors pointed out the benchmarks while considering performance consideration and capacity planning.

Another group of researchers presented the comparative study of distributed and MapReduce methodologies that are commonly used for big data mining [17]. The authors evaluated large datasets so as to compare these two technologies and to find out the accuracy that one gets when baseline comparison is done according to these two procedures using SVM (Support Vector Machine). It also mentioned that the distributed data mining is manual while the data mining done using MapReduce is automatic. The paper also compared four sets of data and provides a result by mining the data in different forms and suggested that MapReduce doesn't cater to class imbalance dataset.

The study [18] suggested the different ways in which we could carry out big data processing in a cloud environment as there has been fast growth of data over various platforms and huge amount of data has to be analysed. This leads to challenges in management of data which can be solved by using cloud data management. The architecture and platform of cloud were also discussed followed by discussion on MapReduce paradigm and the ways in which we can optimize it using methods like Map-Join-Reduce, Iterative Optimization and MapReduce Online.

Agrawal, Das, & El Abbadi [19] presented an analysis of scalable data management solutions using cloud computing. Both update heavy applications and decision support systems were covered in the paper. Increasing popularity of key and value stores due to single tenant systems and the rise of Hadoop had also been referred to. The design principles and problems in cloud data management were also addressed in the paper.

W. Fan & Bifet [20] emphasized on the explosion of data which led to the birth of ‘Big Data’ and the 3 V’s associated with it along with the introduction of variability and value. The example of global pulse were taken, an initiative of the US, which uses Big Data for development purposes. The authors then described the controversies in big data along with the general issues related with it followed by reference to associated tools.

Hashem et al., [21] focused on 4V’s of big data followed by the application of cloud computing on big data and the relationship between them. It also discussed how the classification of big data can be done on five aspects. The various ways of big data storage like NAS, DAS, SAN were also mentioned along with the HDFS and MapReduce. The research challenges like scalability, privacy, data integrity and quality etc. were also described along with open research issues like data staging, data security, data analysis and distributed storage system.

III. OBJECTIVE

The objective of this research work is to discuss the Big Data characteristics, Hadoop and its architecture, HDFS and its various technologies and MapReduce framework along with the open challenges incurred due to Big Data.

IV. METHODOLOGY

To achieve the objective, related published research work was searched on popular databases like Google Scholar and Science Direct. The keyword descriptors like “big data challenges”, “hadoop architecture”, “big data and mapReduce”, “big data and hdfs” etc. were used to searched the articles. The past published studies which were not directly related to the objective of this study were excluded. Thereafter, technological information and innovations were retrieved and presented along with the challenges.

V. ARCHITECTURE OF HADOOP

Hadoop is a batch processing framework where large datasets are processed in a distributed environment. It is an open source framework developed by Apache and it has two components namely MapReduce programming framework and HDFS (Hadoop Distributed File System) [22]. It has been written in Java and we can work with huge amounts of data [13]. It was inspired by Google File System (GFS) and MapReduce. Hadoop can be run on different platforms and it is fault tolerant as it is able to detect and handle failures. It provides scalability. Hadoop is used by Facebook to handle its 2.5 petabyte data warehouse.

Hadoop architecture (see Fig.1) consists of a certain number of nodes and each node has a well-defined role in a Hadoop framework. The components of Hadoop are Name Node, Data Node, Node Manager, Resource Manager and Application Master [13]. The system that has the Name Node is the master node and it is responsible for managing the file system, regulation of access and file system operations. There can be only one Name Node in each cluster.

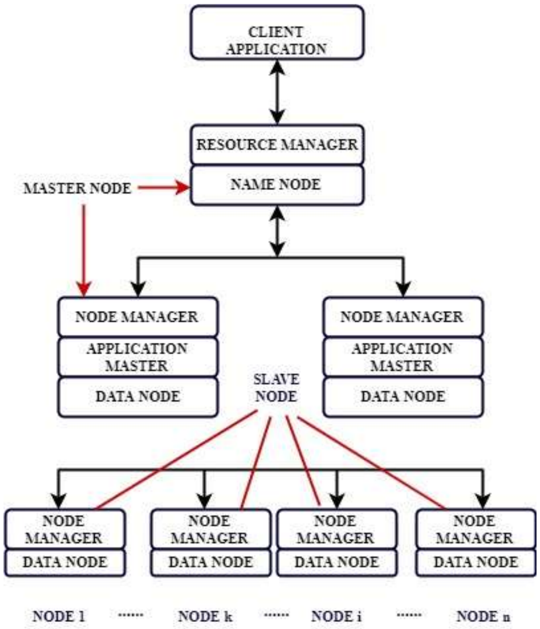

Figure 1.Architecture of Hadoop

VI. HDFS ARCHITECTURE

HDFS is the Hadoop Distributed File System which is used to store large datasets. These huge datasets are broken into chunks and stored across various servers so as to provide fault tolerance and parallelized processing [20].

1. Data Node

   The system has one Data Node for each node in the cluster and this node is the slave node. This node is responsible for storing data, creation and replication of blocks.

2. Node Manager

   It manages the node and is responsible for keeping the Resource Manager updated.

3. Resource Manager

   It determines the resources that are available and manages the distributed applications on YARN.

4. Application Master

   It manages the application lifecycle and interacts with both the Resource Manager and Node Manager.

Hadoop has a distributed file system that is used for storing large volumes of data. This system is fault tolerant and scalable as it divides huge datasets into smaller blocks and then processes them. HDFS consists of the data node and the name node wherein Name Node (or master node) maintains all the metadata and has the details about the various Data Nodes while Data Node (or slave node) maintains the records about the stored and replicated data and reports to the Name Node. The huge amount of data is divided into blocks which are further replicated three times and stored in different data nodes and the name node keeps a tab about all the data nodes. The block size in HDFS is 64 MB and this reduces the metadata storage that is required for each file and thus increases the streaming process. The HDFS architecture is depicted in the Fig. 2. The working of HDFS is as follows[13]:

1. As HDFS uses Java so the client can access (using read operation) the HDFS file by using standard file input stream which is processed in the background.
2. The Name Node takes decision about granting permission to let the client access the file. In case the permission is granted, the block IDs of the file to be accessed are retrieved and all the Data Nodes that store those blocks are also retrieved. This information is sent to the client.
3. Now the closest Data Node is accessed by the client and the required block ID is requested and this block is then delivered to the client.
4. In case the client wants to write in a HDFS file, the standard file output stream is used.
5. The stream is processed in the background and each 64 MB block is broken into smaller block of 64 KB data and then these blocks are queued in FIFO manner.
6. Another thread takes care of de-queuing these blocks from FIFO and works with the Name Node to get block IDs and to send blocks to the Data Nodes.

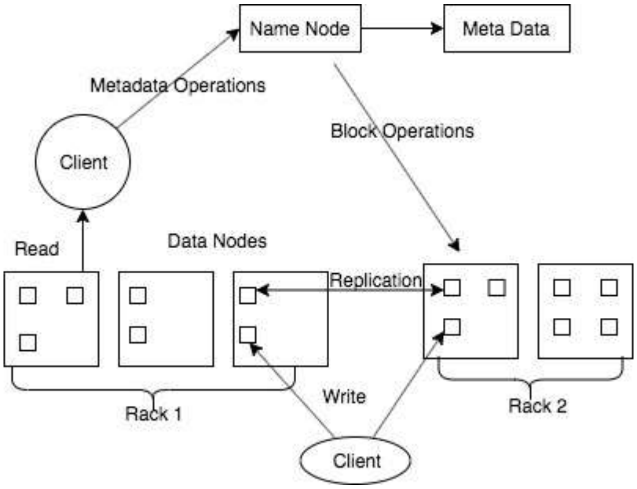

Figure 2. HDFS Architecture

VII. HADOOP MAPREDUCE

MapReduce is a parallel processing framework introduced by Google and it consists of three stages namely mapping stage, the middle stage followed by reducing stage. The idea of MapReduce is to split the file into several files and then into sub files so as to perform mapping on them followed by sorting and shuffling the files and then reducing the result to gain the final output [13][15]. This framework uses key-value pairs (See Fig. 3) where key is a unique identifier for some data item and value is the data that is identified or located. MapReduce provides flexibility, scalability and fault tolerance as it is schema-free and index-free. The various stages of MapReduce are described below [14] [16].

1. Mapping Stage/Function

   This is the first stage and the data is transformed into key-value pairs by applying mapping it. The original input data is not modified by it and only a new output is generated by it. This function provides a sorted key-value pair as the input for the next stage.

2. The Middle Stage

   This stage consists of shuffling, sorting and aggregation wherein the list that is obtained from the mapping stage after parallel processing of chunks of data is shuffled and then aggregated so as to get some new patterns from the list and then the list is sent to the reducing stage.

3. Reducing Stage/Function

   The different patterns that are obtained after the second phase are analyzed and merged in this phase so as to obtain the final result.

VIII. PERFORMANCE ANALYSIS OF HADOOP

HadoopDB consists of multiple single node databases which use Hadoop as a coordinating and network communication layer. It should perform efficiently, be fault tolerant, have flexibility in its query interface and it should be able to run in a heterogeneous environment.

The four components of Hadoop DB are DB Connector, Catalogue, Data-Loader and SQL to Map-Reduce to SQL Planner(SMS).

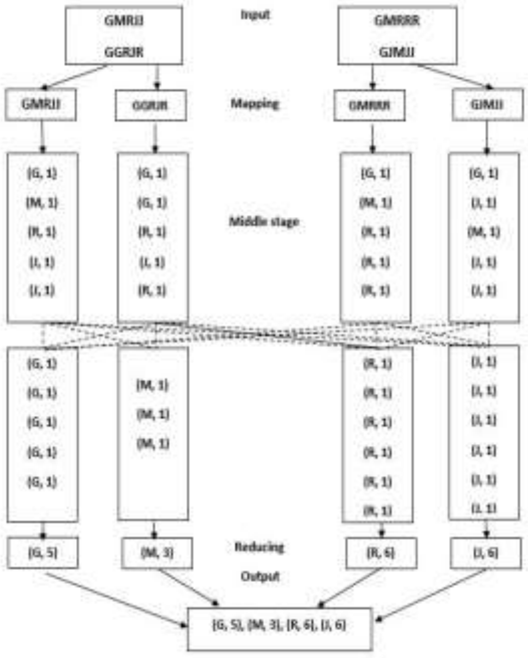

Figure 3. Workflow of MapReduce

Abouzeid et al., [2] compared the performance of Hadoop, Hadoop DB and two commercial parallel processing databases namely Vertica and DBMS-X. In case of data loading, as the nodes are increased, Hadoop DB and Vertica scaled up but DBMS-X took the longest time. But when pre-processing is done, the load time is increased for Hadoop and Hadoop DB. In case of fault tolerance, there were lesser slowdowns for Hadoop and HadoopDB and in case of node failures (i.e. n=10) these have higher fault tolerance but they also pointed out that Vertica's total query processing time was low. So, it is better for smaller node systems. For systems with higher nodes, robustness of Hadoop and HadoopDB is really useful due to the use of low granular chunks.

Patel, Birla, & Nair [14] analysed text processing application wherein the processing time was decreased with increase in the number of nodes and increased with increased size of dataset. They also performed analysis on earthquake data where the processing time decreases with increase in nodes but as the number of days increased, the execution time of data using four nodes was lesser than the time taken for execution by two nodes.

Tsai, Lin, & Ke [17] used Support Vector Machine (SVM) technique for baseline procedure, distributed procedure and MapReduce based procedure. Baseline procedure was carried out on a single machine in a centralized manner. The distributed procedure was divided into n-sets for n computer nodes where n=10 to 50. For MapReduce based procedure, cloud computing environment was simulated. Four datasets were used to analyse using the above mentioned methods. The dataset use for breast cancer was class imbalance dataset and SVM was acquired using baseline and distributed procedures had more classification accuracy than MapReduce based procedure. In case of protein homology dataset, classification accuracy was increased for a single node while it decreased as the nodes increased for MapReduce procedure. The “Cover-type” dataset had high classification accuracy for Map-Reduce. For person activity dataset, MapReduce classification accuracy increased to around 90%.

The study [18] emphasized on Hadoop++ which provides better performance when working with analytical queries tool and Hadoop was integrated in Ricardo. MapDupReducer detects duplicates in big datasets which helps in removing unwanted duplicate values. Map-Reduce-Merge adds a merge phase to avoid data transfer bottleneck which provides optimization of results. Map-Join-Merge improves runtime framework. HaLoop adds loop control which is much needed in Hadoop. To increase the performance efficiency of Hadoop, new tools are being developed. These tools focus on increasing the performance efficiency of Hadoop [19]. For mining big graphs, tools named Pegasus and GraphLab are being used while for Big Data Mining, R, MOA and Vowpal Wabbit are widely used [20]. Some widely used tools related to Hadoop have been mentioned in Table 1.

TABLE I. TOOLS RELATED TO HADOOP

Tool	Description
Hive	It is a warehouse-infrastructure that summarizes, analyses and enables ad-hoc querying by using HiveQL which is a language similar to SQL. Hive is used by Facebook.
Pig	Pig is a platform that is used to process and analyze large datasets that have a high level language. It optimizes and analyses the datasets using parallel processing techniques and it is used by Yahoo.
Spark	Spark is a cluster computing platform that has been built on top of MapReduce and it helps in performing fast computations on datasets. It uses Hadoop for storage purposes and can carry out both batch and real time processing.
Yarn	YARN manages the resources and schedules jobs. It consists of Resource Manager and Application Master. The Resource Manager and the Node Manager are the two components of the computational framework. The Resource Manager manages the resources that are allocated in the system. The Node Manager takes care of the resource usage and reports back to the Resource Manager.
Storm	Storm is a distributed platform which supports real time processing of datasets. It is faster than Hadoop as it repartitions the streams and it requires less development effort. It provides scalability and fault tolerance along with being simple.
Flink	Flink supports stream processing and it was designed for large clusters. It is fault tolerant and it can continuously process infinite datasets. It is faster than Hadoop but it requires more memory.
Samza	Samza is a durable platform that has a simple API and is used for distributed stream processing. It uses Kafka and Hadoop YARN for messaging and resource management respectively.
Beam	Beam is programming model which uses a unified approach as it can carry out both- the stream and batch data processing. It can be used for those problems that can be executed in smaller chunks in an independent way.
Kafka	Kafka is distributed streaming framework that is used for building data pipelines that transform streams of data. It has records which consist of a key-value pair and a timestamp.
Flume	Flume is a reliable and robust platform that is used for collection and aggregation of log data. It uses a simple extensible data model for processing the data.
Sqoop	Sqoop is used for transferring huge amounts of data between Hadoop and relational database. The data is imported and transformed in MapReduce and then it is exported back to the relational database.
Apex	Apex is a batch processing platform which consists of events that are defined according to timestamps. Just like beam, it combines both the stream and batch processing techniques. It is highly scalable and fault tolerant along with being secure. It is easy to use.

IX. CHALLENGES IN HANDLING BIG DATA

HadoopDB consists of multiple single node databases which use Hadoop as a coordinating and network communication layer. It should perform efficiently, be fault tolerant, have flexibility in its query interface and it should be able to run in a heterogeneous environment.

As the amount of data grows, the challenges related to its handling and processing also grow. The major challenges presented by Big Data are discussed in this section [6][21].

1. Scalability

   If we are able to scale up or scale down according to the increasing or decreasing demand, then the system is scalable. As the data grows rapidly, so we should be able to scale up with such an efficiency that we don’t stop the process of analyzing data and while it grows. Due to this NoSQL is used widely as it is consistent and schema free. Hashem, Ibrahim Hashem et al., [21] mentioned a technique called HaCube. This technique is used for analysis of data cube and it is an extension of MapReduce. In the experimental setup, HaCube was faster than Hadoop in terms of view maintenance.

2. Spurious correlation

   Random variables tend to have high sample correlations in data sets of high dimensionality. Due to spurious correlations, we may infer a wrong outcome for a given dataset which can ultimately downgrade the quality of decisions. J. Fan et al., [4] use coefficient vector β which is the sparse vector as $x = Y \beta + \varepsilon$ and var (ε) =σ2 ıd. Using these equations, the author illustrates the maximum absolute sample correlation coefficient between two variables. The illustration also shows the maximum absolute sample correlation coefficient between one variable and closest linear projection of any four members to the other variable. The author pointed out, based on 1000 simulations, that distinguishing a gene from the other four genes was very difficult even though these four genes were not that relevant.

3. Privacy

   As the amount of data rises, an individual prefers to store his or her data on a cloud. When different applications access that data, crucial information can be leaked and that leads to loss of an individual’s private data as this data needs to be analyzed by different platforms to gain an insight about the data. Block Encryption is used to protect the data on clouds.

Hashem et al., [21] highlighted various ways that the privacy of a user could be protected. If the intermediate clouds have to be encrypted, then the time utilization and cost of providing privacy can increase. If the user’s search privileges are controlled then, the task of preserving privacy becomes easier for the cloud owner. Three tier architecture that provides several layers of privacy to cloud users were elaborated.

1. Data quality

   Data quality has been a concern even for traditional analytical techniques [5]. In terms of Big Data, dirty data i.e. the data that has errors or is incorrect is a big problem as on analyzing it we’ll get unreliable outputs which can affect the decision of an organization. Data that is collected from heterogeneous sources may be inconsistent that is the data may not have the same format in a database. This will result in inaccurate outputs as the quality of data is deteriorated.

2. Heterogeneity

   As the data is generated and collected from various resources, so the data is a collection of structured, unstructured and semi- structured data. To operate on this data, we need to find techniques that will be able to work on such a huge collection of various types of data [4].

Heterogeneity plays a big role in Big Data. The data has a large sample size using which we can discover an association between covariates and develop a better understanding about the data.

X. OPEN RESEARCH CHALLENGES

HadoopDB consists of multiple single node databases which use Hadoop as a coordinating and network communication layer. It should perform efficiently, be fault tolerant, have flexibility in its query interface and it should be able to run in a heterogeneous environment.

Data is growing exponentially but the development of processing mechanisms that are efficient to handle this data is not at a rate at par with the growth of data [6]. There are certain problems that are posed by rapid growth of data and there is a need to address these problems by carrying out research in these areas. [6][21].

1. Security

   Security of Big Data is also an issue as the huge amount of data needs to be encrypted along with the application of user authentication and access regulation techniques so as to save the data from unauthorized access.

2. Analysis

   To analyze large datasets that have data from various sources, we need specials tools. These tools should be able to provide an accurate outcome irrespective of the source and type of data.

3. Distributed Storage

   Storing and retrieving big data is a huge task. The storage devices that are available are incapable of storing this data as the maximum size of a hard disk available till date is just 60 TB and this is not enough to handle huge datasets [18].

4. Staging

   Before transformation of data, the data that has to be processed is stored in a single place. This is known as the staging area. The data collected is of heterogeneous nature in case of Big Data. So the data may be structured, unstructured or semi-structured. This data needs to be converted into a simplified form so as to procure useful information from it.

XI. CONCLUSION

The data, as predicted, is growing at a fast rate and the various organizations are trying to work at its pace in order to yield the best results out of this growth. The organizations are striving to use this data to the best of their potential so as to help them understand and gain an insight into their key demographics and also to help them in making a better and informed decision. Hadoop provides a framework to handle datasets that are large and complex with MapReduce being one of the most popular ones. It also provides other specialized tools like Spark, Pig, and Hive etc. to operate on these datasets. Hadoop is an easy to use platform with scalability and parallelization as its main features. But there are various other challenges that need to be worked on in order to provide a better experience in working with Big Data thus helping the organizations in optimizing their services. The present work highlighted key challenges of big data i.e. scalability, spurious correlation, privacy, data quality, heterogeneity, integrity and noise, accumulation and incidental endogeneity which need to be addressed in the future. The paper also pointed out some open research challenges like security, data analysis, distributed storage and staging which are crucial and must be addressed in the future research.

REFERENCES

[1] R. Bryant, R. H. Katz, and E. D. Lazowska, “Big-data computing: creating revolutionary breakthroughs in commerce, science and society.” December, 2008.

[2] A. Abouzeid, K. Bajda-Pawlikowski, D. J. Abadi, A. Silberschatz, and A. Rasin, “HadoopDB: An architectural hybrid of MapReduce and DBMS technologies,” 2009.

[3] S. Chaudhuri, “What next?: a half-dozen data management research goals for big data and the cloud,” in Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems, 2012, pp. 1–4.

[4] Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National science review, 1(2), 293-314.

[5] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, “Addressing big data issues in scientific data infrastructure,” in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 2013, pp. 48–55.

[6] S. J. Samuel, K. Rvp, K. Sashidhar, and C. R. Bharathi, “A survey on big data and its research challenges,” *ARPN J. Eng. Appl. Sci*, vol. 10, no. 8, pp. 3343–3347, 2015.

[7] R. Kitchin and G. McArdle, “What makes Big Data, Big Data? Exploring the ontological

characteristics of 26 datasets,” *Big Data Soc.*, vol. 3, no. 1, p. 2053951716631130, 2016.

[8] Anuradha, J. (2015). A brief introduction on Big Data 5Vs characteristics and Hadoop technology. Procedia computer science, 48, 319-324.

[9] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), 137-144.

[10] P. Groves, B. Kayyali, D. Knott, and S. Van Kuiken, “The’big data'revolution in healthcare: Accelerating value and innovation,” 2016.

[11] V. Kumar, A. Chaturvedi, and P. Verma, “Applications of Big Data in the Digital India: Opportunities and Challenges,” IRA-International J. Technol. Eng. (ISSN 2455-4480), vol. 3, no. 3, 2016.

[12] R. Rein and D. Memmert, “Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science,” *Springerplus*, vol. 5, no. 1, p. 1410, 2016.

[13] Padhy, Rabi Prasad. "Big data processing with Hadoop-MapReduce in cloud systems." International Journal of Cloud Computing and Services Science 2.1 (2013): 16.

[14] A. B. Patel, M. Birla, and U. Nair, “Addressing big data problem using Hadoop and Map Reduce,” in 2012 Nirma University International Conference on Engineering (NUiCONE), 2012, pp. 1–5.

[15] H. S. Bhosale and D. P. Gadekar, “A review paper on big data and hadoop,” *Int. J. Sci. Res. Publ.*, vol. 4, no. 10, pp. 1–7, 2014.

[16] K. Bakshi, “Considerations for big data: Architecture and approach,” in *2012 IEEE Aerospace Conference*, 2012, pp. 1–7.

[17] C.-F. Tsai, W.-C. Lin, and S.-W. Ke, “Big data mining with parallel computing: A comparison of distributed and MapReduce methodologies,” *J. Syst. Softw.*, vol. 122, pp. 83–92, 2016.

[18] C. Ji, Y. Li, W. Qiu, U. Awada, and K. Li, “Big data processing in cloud computing environments,” in 2012 12th international symposium on pervasive systems, algorithms and networks, 2012, pp. 17–23.

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

[20] Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. ACM SIGKDD explorations newsletter, 14(2), 1-5.

[21] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information systems, 47, 98-115.

[22] Ghazi, M. R., & Gangodkar, D. (2015). Hadoop, MapReduce and HDFS: a developers perspective. Procedia Computer Science, 48(C), 45-50.