Challenges, Research Issues Open and State of Art of Big Data – A Study

Mathi Vanan. P [1], Mohana Priya. D [2], Nagamany Abirami. D [3]

[1]*Department of Information Technology, Manakula Vinayagar Institute of Technology, Puducherry, India*
[2, 3] *Department of Computer Science and Engineering, Manakula Vinayagar Institute of Technology, Puducherry, India*

Abstract: **- The term 'Big Data', refers to data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to capture, manage, process or analyzed. Big data is a collection of large data sets that include different types such as structured, unstructured and semi-structured data. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. In this paper, we have proposed a recommendation system for the large amount of data available on the web in the form of ratings, reviews, opinions, complaints, remarks, feedback, and comments about any item (product, event, individual and services) using Hadoop Framework. The main objective of the proposed study is to provide a better and significant research perspective and an overview of data analysis techniques which are referred to the papers found on the web which will be quite helpful for the future research prospective of this domain.**

Keywords:-**Big Data Analysis, Hadoop Framework, Recommendation System, Mahout, HDFS.**

I. INTRODUCTION

Big Data is as a collection of large dataset that cannot be processed using traditional computing techniques .Big Data is not merely a data rather it has become a complete subject which involve various tools, techniques and framework. The biggest phenomenon that has captured the attention of the modern computing industry today since the "Internet" is "Big Data". The fundamental reason why "Big Data" is popular today is because the technology platforms that have emerged along with it provide the capability to process data of multiple formats and structures without worrying about the constraints associated with traditional systems and database platforms. The need of big data generated from the large companies like facebook, yahoo, Google, YouTube etc for the purpose of analysis of enormous amount of data also Google contains the large amount of information

Big Data can be defined as volumes of data available in varying degrees of complexity, generated at different velocities and varying degrees of ambiguity, that cannot be processed using traditional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions. Data defined as Big Data includes machine-generated data from sensor networks, nuclear plants, X-ray and scanning devices, and airplane engines, and consumer-driven data from social media. Big Data producers that exist within organizations include legal, sales, marketing, procurement, finance, and human resources departments.

Big Data is a term that refers to dataset whose volume (size), complexity and rate of growth (velocity) make them to difficult to captured, managed, processed or analyzed by conventional technology and tools such as relational databases. There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. The data in it will be of three types. Structured data: Relational data. Semi Structured data: XML data. Unstructured data: Word, PDF, Text, Media Logs. The data analysis concept of Big Data gives various analytical methods which can be applicable for analyzing traditional datasets which includes various analytical architecture, software requirement for exploration of big data. Data investigation is one of the most essential stages of the big data value chain where the main objective is to extract the meaningful information and providing suggestions and decisions. Different types of possible and gravitational values can be produced through the several stages of analysis in different fields. Data Analysis is considered to be a very broad area where the environment is so complex and associated with the use of various complex methods, architectures, and tools.

1.1 CHARACTERISTICS OF BIG DATA

As the data is too big and comes from various sources in different form, it is characterized by the following five components:

VARIETY

Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail, documents, sensor devices data both from active passive devices. All this data is totally different consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems.

VOLUME

The Big word in Big data itself defines the volume. At present the data existing is in peta bytes and is supposed to increase to zeta bytes in nearby future. The social networking sites existing are themselves producing data in order of terabytes everyday and this amount of data is definitely difficult to be handled using the existing traditional systems.

VELOCITY

Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example, the data from the sensor devices would be constantly moving to the database store and this amount won't be small enough.

Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.

VARIABILITY

Variability considers the inconsistencies of the data flow. Data loads become challenging to be maintained especially with the increase in usage of the social media which generally causes peak in data loads with certain events occurring.

COMPLEXITY

It is quite an undertaking to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control.

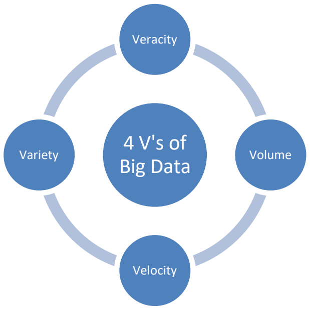


Fig. 1: Characteristics of Big Data

II. LITERATURE REVIEW

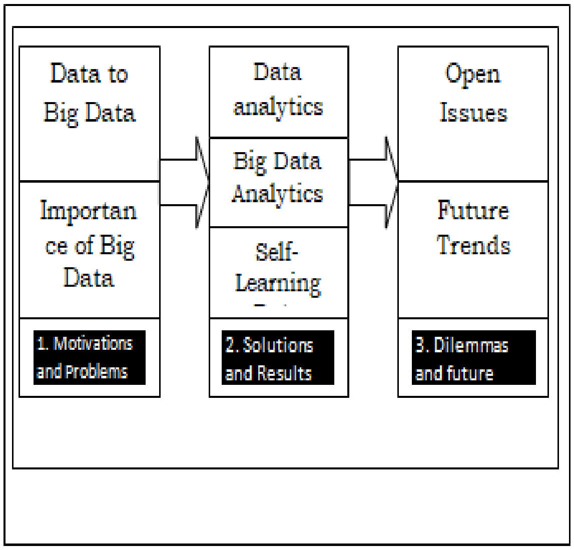
The authors [5] pointed out that handling of huge data using earlier RDBMS tools is little bit complex, hence feels the necessity of alternate tools that can handle such a huge data which is usually referred to as "big data**".** In this, the authors argued that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They illustrated the hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data systems. The authors also focused on the challenges that need to be faced by enterprises when handling big data: - data privacy, search analysis, etc. The authors [5] discussed the analysis of big data and they stated that Data is generated through many sources like business processes, transactions, social networking sites, web servers, etc. and remains in structured as well as unstructured form Processing or analyzing the huge amount of data or extracting meaningful information is a challenging task. The term "Big data" is used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many peta bytes of data in a single data set.

Difficulties include capture, storage, search, sharing, analytics and visualizing. The Authors [5] **have** done a lot of experiment on the big data problem. At last he found that the hadoop cluster, Hadoop Distributed File System (HDFS) for storage and map reduce method for parallel processing on a large volume of data. The Authors [5] emphasizes on a prominent data processing tool Map Reduce survey which will help in understanding various technical aspects of the Map Reduce framework. In this survey, the author expresses different views on Map Reduce framework and introduces its optimization strategies. Author also hands a challenge on parallel data analysis with Map Reduce framework. The Authors [5] defines big data Problem using Hadoop and Map Reduce" reports the experimental research on the Big data problems in various domains. The authors have briefly discussed about HDFS and Map Reduce technology to process massive data sets and records. The Authors [5] Stated that streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge, allowing organizations to react quickly when problem appear or detect to improve performance. Huge amount of data is created everyday termed as "big data". The tools used for mining big data are apache Hadoop, apache pig, cascading, scribe, storm, apache Hbase, apache mahout, MOA, etc. Thus, he instructed that our ability to handle many Exabyte's of data mainly dependent on existence of rich variety dataset, technique, software framework.

Case to manage and different data structure using map reduce method. The advantage of this paper is improves the performance of large scale Hadoop clusters.

III. DATA ANALYTICS

To make the whole process of knowledge discovery in databases (KDD) more clear, Fayyad and his colleagues summarized the KDD process by a few operations in [10], which are selection, preprocessing, transformation, data mining, and interpretation/ evaluation. As shown in Fig. 2, with these operators at hand we will be able to build a complete data analytics system to gather data first and then find information from the data and display the knowledge to the user. According to our observation, the number of research articles and technical reports that focus on data mining is typically more than the number focusing on other operators, but it does not mean that the other operators of KDD are unimportant. The other operators also play the vital roles in KDD process because they will strongly impact the final result of KDD. To make the discussions on the main operators of KDD process more concise, the following sections will focus on those depicted in Fig. 2, which were simplified to three parts (input, data analytics, and output) and seven operators (gathering, selection, preprocessing, transformation, data mining, evaluation, and interpretation).



The preprocessing operator plays a different role in dealing with the input data which is aimed at detecting, cleaning, and filtering the unnecessary, inconsistent, and incomplete data to make them the useful data. After the selection and preprocessing operators, the characteristics of the secondary data still may be in a number of different data formats; therefore, the KDD process needs to transform them into a data-mining-capable format which is performed by the transformation operator.

The methods for reducing the complexity and downsizing the data scale to make the data useful for data analysis part are usually employed in the transformation, such as dimensional reduction, sampling, coding, or transformation.

The data extraction, data cleaning, data integration, data transformation, and data reduction operators can be regarded as the preprocessing processes of data analysis [10] which attempts to extract useful data from the raw data (also called the primary data) and refine them so that they can be used by the following data analyses. If the data are a duplicate copy, incomplete, inconsistent, noisy, or outliers, then these operators have to clean them up. If the data are too complex or too large to be handled, these operators will also try to reduce them. If the raw data have errors or omissions, the roles of these operators are to identify them and make them consistent. It can be expected that these operators may affect the analytics result of KDD, be it positive or negative.

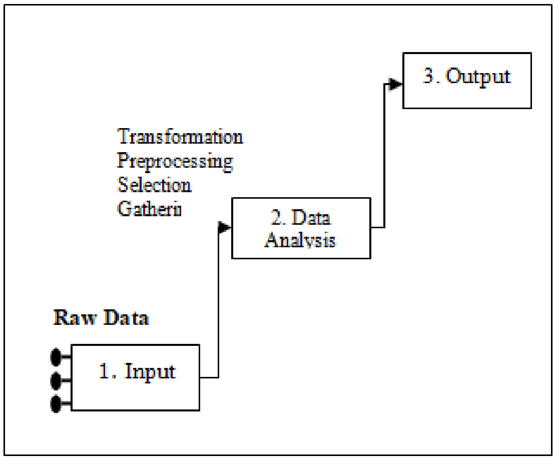


Fig: 3 The Processing of Databases

In summary, the systematic solutions are usually to reduce the complexity of data to accelerate the computation time of KDD and to improve the accuracy of the analytics result.

3.1 DATA ANALYSIS

The data analysis (as shown in Fig. 3) in KDD is responsible for finding the hid-den patterns/rules/information from the data, most researchers in this field use the term data mining to describe how they refine the “ground” (i.e, raw data) into “gold nugget” (i.e., information or knowledge). The data mining methods [10] are not limited to data problem specific methods.

In fact, other technologies (e.g., statistical or machine learning technologies) have also been used to analyze the data for many years. In the early stages of data analysis, the statistical methods were used for analyzing the data to help us understand the situation we are facing, such as public opinion poll or TV program me rating. Like the statistical analysis, the problem specific methods for data mining also attempted to understand the meaning from the collected data.

After the data mining problem was presented, some of the domain specific algorithms are also developed. An example is the apriori algorithm [10] which is one of the useful algorithms designed for the association rules problem. Although most definitions of data mining problems are simple, the computation costs are quite high.

To speed up the response time of a data mining operator, machine learning [10], meta heuristic algorithms [10], and distributed computing [10] were used alone or combined with the traditional data mining algorithms to provide more efficient ways for solving the data mining problem. One of the well-known combinations can be found in [10], Krishna and Murty attempted to combine genetic algorithm and *k*-means to get better clustering result than *k*-means alone does.

As Fig. 4 shows, most data mining algorithms contain the initialization, data input and output, data scan, rules construction, and rules update operators [10].

```
Input data D
Initialize candidate solutions r
While the termination criterion is not met
d = Scan(D)
v = Construct(d, r, o)
r = Update(v)
End
Output rules r
```

represents the raw data, *d* the data from the scan operator, *r* the rules, *o* the predefined measurement, and *v* the candidate rules. The scan, construct, and update operators will be performed repeatedly until the termination criterion is met. The timing to employ the scan operator depends on the design of the data mining algorithm; thus, it can be considered as an optional operator. Most of the data algorithms can be described by Fig. 4 in which it also shows that the representative algorithms—*clustering*, *classification*, *association rules*, and *sequential patterns*—will apply these operators to find the hidden information from the raw data. Thus, modifying these operators will be one of the possible ways for enhancing the performance of the data analysis.

Clustering is one of the well-known data mining problems because it can be used to understand the “new” input data. The basic idea of this problem [10] is to separate a set of unlabeled input data[2] to *k* different groups, e.g., such as *k*-means [28].

3.2 OUTPUT THE RESULT

Evaluation and interpretation are two vital operators of the output. Evaluation typically plays the role of measuring the results. It can also be one of the operators for the data mining algorithm, such as the sum of squared errors which was used by the selection operator of the genetic algorithm for the clustering problem [10].

To solve the data mining problems that attempt to classify the input data, two of the major goals are: (1) cohesion—the distance between each data and the centroid (mean) of its cluster should be as small as possible, and (2) coupling—the distance between data which belong to different clusters should be as large as possible. In most studies of data clustering or classification problems, the sum of squared errors (SSE), which was used to measure the cohesion of the data mining results, can be defined as

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{ni} D(xij - ci)$$

$$Ci = \frac{1}{ni}\sum_{i=1}^{n} Xij$$

where *k* is the number of clusters which is typically given by the user; ni the number ofdata in the *i*th cluster; xij the *j*th datum in the *i*th cluster; ci is the mean of the *i*th cluster;

For solving different data mining problems, the distance measurement D(pi , pj) can be the Manhattan distance, the Minkowski distance, or even the cosine similarity between two different documents.

Accuracy (ACC) is another well-known measurement which is defined as

$$\textbf{ACC} = \frac{\textbf{Number of cases correctly classified}}{\textbf{Total number of test cases}}$$

To evaluate the classification results, precision (*p*), recall (*r*), and *F*-measure can be used to measure how many data that do not belong to group *A* are incorrectly classified into group *A*; and how many data that belong to group *A* are not classified into group *A*. A simple confusion matrix of a classifier as given in Table 1 can be used to cover all the situations of the classification results. In Table 1, TP and TN indicate the numbers of positive examples and negative examples that are correctly classified, respectively; FN and FP indicate the numbers of positive examples and negative examples that are incor-rectly classified, respectively. With the confusion matrix at hand, it is much easier to describe the meaning of precision (*p*), which is defined as

$$p = \frac{TP}{TP + FP'}$$

and the meaning of recall (*r*), which is defined as

$$r = \frac{TP}{TP + FN}$$

The *F*-measure can then be computed as

$$F = \frac{2pr}{p + r}$$

In addition to the above-mentioned measurements for evaluating the data mining results, the computation cost and response time are another two well-known measurements.

When two different mining algorithms can find the same or similar results, of course, how fast they can get the final mining results will become the most important research topic.

Problem	Method
Clustering	BIRCH DBSCAN Incremental DBSCAN RKM
Classification	TKM SLIQ TLAESA FastNN SFFS GPU-based SVM
Association rules	CLOSET FP-tree CHARM MAFIA FAST
Sequential patterns	SPADE CloSpan PrefixSpan SPAM ISE

3.3 BIG DATA ANALYTICS

Nowadays, the data that need to be analyzed are not just large, but they are composed of various data types, and even including streaming data [10]. Since big data has the unique features of "massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous," which may change the statistical and data analysis approaches. Although it seems that big data makes it possible for us to collect more data to find more useful information, the truth is that more data do not necessarily mean more useful information. It may contain more ambiguous or abnormal data. For instance, a user may have multiple accounts, or an account may be used by multiple users, which may degrade the accuracy of the mining results . Therefore, several new issues for data analytics come up, such as privacy, security, storage, fault tolerance, and quality of data.

3.3.1 Big data input

The problem of handling a vast quantity of data that the system is unable to process is not a brand-new research issue; in fact, it appeared in several early approaches [2, 2, 4], e.g., marketing analysis, network flow monitor, gene expression analysis, weather forecast, and even astronomy analysis. This problem still exists in big data analytics today; thus, preprocessing is an important task to make the computer, platform, and analysis algorithm be able to handle the input data. The traditional data preprocessing methods (e.g., compression, sampling, feature selection, and so on) are expected to be able to operate effectively in the big data age.

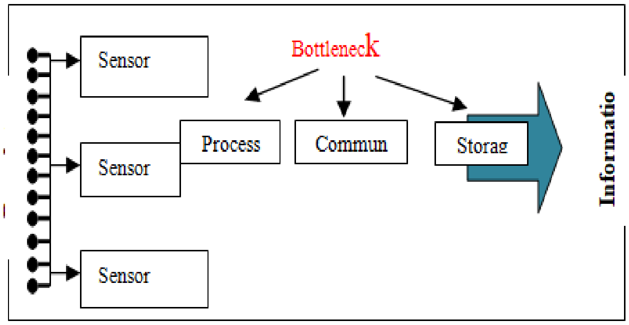


Fig:4 The Sensing system of data deluge

Sampling and compression are two representative data reduction methods for big data analytics because reducing the size of data makes the data analytics computationally less expensive, thus faster, especially for the data coming to the system rapidly.

3.4 BIG DATA ANALYSIS FRAMEWORKS AND PLATFORMS

Various solutions have been presented for the big data analytics which can be divided [10] into (1) Processing/Compute: Hadoop [10], Nvidia CUDA [10], or Twitter Storm [10], (2) Storage: Titan or HDFS, and (3) Analytics: MLPACK [10] or Mahout [10]. Although there exist commercial products for data analysis , most of the studies on the traditional data analysis are focused on the design and development of efficient and/or effective "ways" to find the useful things from the data. But when we enter the age of big data, most of the current computer systems will not be able to handle the whole dataset all at once; thus, how to design a good data analytics framework or plat-form3 and how to design analysis methods are both important things for the data analysis process.

3.5 RESEARCHES IN FRAMEWORKS AND PLATFORMS

To date, we can easily find tools and platforms presented by well-known organizations. The cloud computing technologies are widely used on these platforms and frameworks to satisfy the large demands of computing power and storage. As shown in Fig. 7, most of the works on KDD for big data can be moved to cloud system to speed up the response time or to increase the memory space. With the advance of these works, handling and analyzing big data within a reasonable time has become not so far away. Since the foundation functions to handle and manage the big data were developed gradually; thus, the data scientists nowadays do not have to take care of everything, from the raw data gathering to data analysis, by themselves if they use the existing platforms or technologies to handle and manage the data.

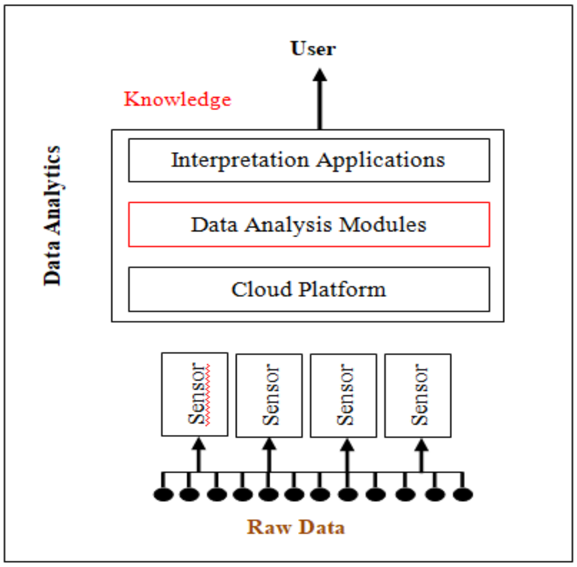


Fig:5 : Big Data Analytics

The data scientists nowadays can pay more attention to finding out the useful information from the data even thought this task is typically like looking for a needle in a haystack. That is why several recent studies tried to present efficient and effective framework to analyze the big data, especially on find out the useful things.

3.6 BIG DATA ANALYSIS ALGORITHMS

Mining algorithms for specific problem

The big data and big data mining almost appearing at the same time explained that finding something from big data will be one of the major tasks in this research domain. Data mining algorithms for data analysis also play the vital role in the big data analysis, in terms of the computation cost, memory requirement, and accuracy of the end results. In this section, we will give a brief discussion from the perspective of analysis and search algorithms to explain its importance for big data analytics.

Clustering algorithms

Clustering algorithms In the big data age, traditional clustering algorithms will become even more limited than before because they typically require that all the data be in the same format and be loaded into the same machine so as to find some useful things from the whole data. Although the problem [10] of analyzing large-scale and high-dimensional dataset has attracted many researchers from various disciplines in the last century, and several solutions [2, 109] have been presented in recent years, the characteristics of big data still brought up several new challenges for the data clustering issues. Among them, how to reduce the data complexity is one of the important issues for big data clustering. The big data clustering into two categories: single-machine clustering (i.e., sampling and dimension reduction solutions), and multiple-machine clustering (parallel and MapReduce solutions). This means that traditional reduction solutions can also be used in the big data age because the complexity and memory space needed for the process of data analysis will be decreased by using sampling and dimension reduction methods. More precisely, sampling can be regarded as reducing the "amount of data" entered into a data analyzing process while dimension reduction can be regarded as "downsizing the whole dataset" because irrelevant dimensions will be discarded before the data analyzing process is carried out.

Classification algorithms

Classification algorithms Similar to the clustering algorithm for big data mining, sev-eral studies also attempted to modify the traditional classification algorithms to make them work on a parallel computing environment or to develop new classification algorithms which work naturally on a parallel computing environment. In [10], the design of classification algorithm took into account the input data that are gathered by distributed data sources and they will be processed by a heterogeneous set of learners.

IV. OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Big data analytics and data science are becoming the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction from data. Applications of big data and data science include information science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing. Effective integration of technologies and analysis will result in predicting the future drift of events. Main focus of this section is to discuss open research issues in big data analytics. The research issues pertaining to big data analysis are classified into three broad categories namely internet of things (IoT), cloud computing, bio inspired computing, and quantum computing. However it is not limited to these issues. More research issues related to health care big data can be found in Husing Kuo et al. paper .

A. IoT for Big Data Analytics

Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Currently, machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things (IoT). Thus, appliances are becoming the user of the internet, just like humans with the web browsers. Internet of Things is attracting the attention of recent researchers for its most promising opportunities and challenges. It has an imperative economic and societal impact for the future construction of information, network and communication technology. The new regulation of future will be eventually; everything will be connected and intelligently controlled. The concept of IoT is becoming more pertinent to the realistic world due to the development of mobile de-vices, embedded and ubiquitous communication technologies, cloud computing, and data analytics. Moreover, IoT presents challenges in combinations of volume, velocity and variety. Knowledge acquisition from IoT data is the biggest challenge that big data professional are facing. Therefore, it is essential to develop infrastructure to analyze the IoT data. An IoT device generates continuous streams of data and the re-searchers can develop tools to extract meaningful information from these data using machine learning techniques. Under-standing these streams of data generated from IoT devices and analyzing them to get meaningful information is a challenging issue and it leads to big data analytics. Machine learning algorithms and computational intelligence techniques is the only solution to handle big data from IoT prospective.

Knowledge exploration system have originated from theories of human information processing such as frames, rules, tagging, and semantic networks. In general, it consists of four segments such as knowledge acquisition, knowledge base, knowledge dissemination, and knowledge application.

In knowledge acquisition phase, knowledge is discovered by using various traditional and computational intelligence techniques. The discovered knowledge is stored in knowledge bases and expert systems are generally designed based on the discovered knowledge.

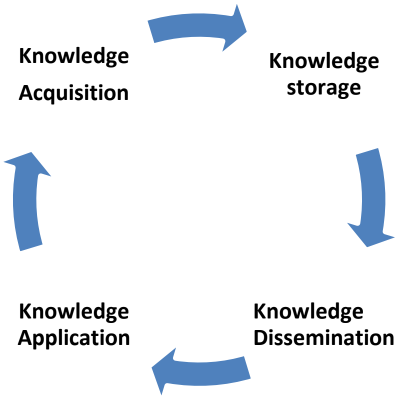


Fig. 6: IoT Knowledge Exploration System

Knowledge dissemination is important for obtaining meaningful information from the knowledge base. Knowledge extraction is a process that searches documents, knowledge within documents as well as knowledge bases. The final phase is to apply discovered knowledge in various applications. It is the ultimate goal of knowledge discovery. The knowledge exploration system is necessarily iterative with the judgment of knowledge application.

B. Cloud Computing for Big Data Analytics

The development of virtualization technologies have made supercomputing more accessible and affordable. Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processors, disk space, memory, and operating system. The use of these virtual computers is known as cloud computing which has been one of the most robust big data techniques.

Big Data and cloud computing technologies are developed with the importance of developing a scalable and on demand availability of resources and data. Cloud computing harmonize massive data by on demand access to configurable computing resources through virtualization techniques. The benefits of utilizing the Cloud computing include offering resources when there is a demand and pay only for the resources which is needed to develop the product. Simultaneously, it improves availability and cost reduction.

Open challenges and research issues of big data and cloud computing are discussed in detail by many researchers which highlights the challenges in data management, data variety and velocity, data storage, data processing, and resource management [1], [2]. So Cloud computing helps in developing a business model for all varieties of applications with infrastructure and tools. Big data application using cloud computing should support data analytic and development. In addition to this, cloud computing should also enable scaling of tools from virtual technologies into new technologies like spark, R, and other types of big data processing techniques.

Big data forms a framework for discussing cloud computing options. Depending on special need, user can go to the marketplace and buy infrastructure services from cloud service providers such as Google, Amazon, IBM, software as a service (SaaS) from a whole crew of companies such as NetSuite, Cloud9, Job science etc.

Another advantage of cloud computing is cloud storage which provides a possible way for storing big data. The obvious one is the time and cost that are needed to upload and download big data in the cloud environment. Else, it becomes difficult to control the distribution of computation and the underlying hardware. But, the major issues are privacy concerns relating to the hosting of data on public servers, and the storage of data from human studies. All these issues will take big data and cloud computing to a high level of development.

C. Bio-inspired Computing for Big Data Analytics

Bio-inspired computing is a technique inspired nature to address complex real world problems. Biological systems are self organized without a central control. A bio-inspired cost minimization mechanism search and find the optimal data service solution on considering cost of data management and service maintenance. These techniques are developed by biological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and processing of data. A significant feature of such computing is that it integrates biologically derived materials to perform computational functions and receive intelligent performance. These systems are more suitable for big data applications.

D. Quantum Computing for Big Data Analysis

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously [4]. This exponential improvement in computer systems might be possible. If a real quantum computer is available now, it could have solved problems that are exceptionally difficult on recent computers, of course today's big data problems. The main technical difficulty in building quantum computer could soon be possible.

Quantum computing provides a way to merge the quantum mechanics to process the information. In traditional computer, information is presented by long strings of bits which encode either a zero or a one. On the other hand a quantum computer uses quantum bits or qubits[6]. The difference between qubit and bit is that, a qubit is a quantum system that encodes the zero and the one into two distinguishable quantum states. Therefore, it can be capitalized on the phenomena of superposition and entanglement. It is because qubits behave quantum.

V. TOOLS AND TECHNOLOGIES

For the purpose of processing the large amount of data, the big data requires exceptional technologies. The various techniques and technologies have been introduced for manipulating, analyzing and visualizing the big data . There are many solutions to handle the Big Data, but the Hadoop is one of the most widely used technologies [3].

A. HADOOP

Using the solution provided by Google, Doug Cutting and his team developed an Open Source Project called HADOOP. Hadoop runs applications using the Map Reduce algorithm, where the data is processed in parallel with others. Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data. Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models.

The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput. Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines.

Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop perform. Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M).These files are then distributed across various cluster nodes for further processing. HDFS, being on top of the local file system, supervises the processing [8]. Blocks are replicated for handling hardware failure. Checking that the code was executed successfully, performing the sort that takes place between the map and reduce stages, sending the sorted data to a certain computer, writing the debugging logs for each job.

Hadoop has two major layers namely:

1. Processing/Computation layer (Map Reduce)

2. Storage layer (Hadoop Distributed File System).

B. MAP REDUCE

Map Reduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multiterabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

The Map Reduce program runs on Hadoop which is an Apache open-source framework [4]. It is a processing technique and a program model for distributed computing based on java[5]. The Map Reduce algorithm contains two important tasks, namely Map and Reduce.

The major advantage of Map Reduce is that it is easy to scale data processing over multiple computing nodes. Under the Map Reduce model, the data processing primitives are called mappers and reducers.

Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the Map Reduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the Map Reduce model.

The stages of Map Reduce Program

Generally Map Reduce paradigm is based on sending the computer to where the data resides! Map Reduce program executes in two stages, namely map stage and reduce stage.

1. *Map stage: The map or mapper"s job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.*
2. *Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer"s job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS*

C. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.

HDFS follows the master-slave architecture and it has the following elements.

NAME NODE

The name node is the commodity hardware that contains the GNU/Linux operating system and the name node software. It is software that can be run on commodity hardware. The system having the name node acts as the master server and it does the following tasks: Manages the file system namespace. Regulates clients access to files and It also executes file system operations such as renaming, closing, and opening files and directories.

DATA NODE

The data node is a commodity hardware having the GNU/Linux operating system and data node software. For every node (Commodity hardware/System) in a cluster, there will be a data node. These nodes manage the data storage of their system. Data nodes perform read-write operations on the file systems, as per client request[5]. They also perform operations such as block creation, deletion, and replication according to the instructions of the name node.

BLOCK

Generally the user data is stored in the files of HDFS. The file in a file system will be divided into one or more segments and/or stored in individual data nodes. These file segments are called as blocks. In other words, the minimum amount of data that HDFS can read or write is called a Block[5]. The default block size is 64MB, but it can be increased as per the need to change in HDFS configuration.

OTHER DIFFERENT COMPONENTS OF HADOOP ARE:

Apache Mahout

Apache mahout aims to provide scalable and commercial machine learning techniques for large scale and intelligent data analysis applications. Core algorithms of mahout including clustering, classification, pattern mining, regression, dimensionality reduction, evolutionary algorithms, and batch based collaborative filtering run on top of Hadoop platform through map reduce framework. The goal of mahout is to build a vibrant, responsive, diverse community to facilitate discussions on the project and potential use cases. The basic objective of Apache mahout is to provide a tool for elevating big challenges. The different companies those who have implemented scalable machine learning algorithms are Google, IBM, Amazon, Yahoo, Twitter, and face book [6].

D. Dryad

It is another popular programming model for implementing parallel and distributed programs for handling large context bases on dataflow graph. It consists of a cluster of computing nodes, and an user use the resources of a computer cluster to run their program in a distributed way. Indeed, a dryad user use thousands of machines, each of them with multiple processors or cores. The major advantage is that users do not need to know anything about concurrent programming.

E. Storm

Storm is a distributed and fault tolerant real time computation system for processing large streaming data. It is specially designed for real time processing in contrasts with Hadoop which is for batch processing. Additionally, it is also easy to set up and operate, scalable, fault-tolerant to provide competitive performances.

The storm cluster is apparently similar to Hadoop cluster. On storm cluster users run different topologies for different storm tasks whereas Hadoop platform implements map reduce jobs for corresponding applications.

F. Apache Drill

Apache drill is another distributed system for interactive analysis of big data. It has more flexibility to support many types of query languages, data formats, and data sources. It is also specially designed to exploit nested data. Also it has an objective to scale up on 10,000 servers or more and reaches the capability to process patabytes of data and trillions of records in seconds. Drill use HDFS for storage and map reduce to perform batch analysis.

G. Splunk

In recent years a lot of data are generated through machine from business industries. Splunk is a real-time and intelligent platform developed for exploiting machine generated big data. It combines the up-to-the-moment cloud technologies and big data. In turn it helps user to search, monitor, and analyze their

The most important objective of Splunk is to provide metrics for much application, diagnose problems for system and information technology infrastructures, and intelligent support for business operations.

H. Apache Spark

Apache spark is an open source big data processing framework built for speed processing, and sophisticated analytics.

It is easy to use and was originally developed in 2009 in UC Berkeleys AMPLab. It was open sourced in 2010 as an Apache project. Spark lets you quickly write applications in java, scala, or python. In addition to map reduce operations, it supports SQL queries, streaming data, machine learning, and graph data processing. Spark runs on top of existing hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes. The driver program serves as the starting point of execution of an application on the spark cluster.

Figure 6 depicts the architecture diagram of Apache Spark. The various features of Apache Spark are listed below:

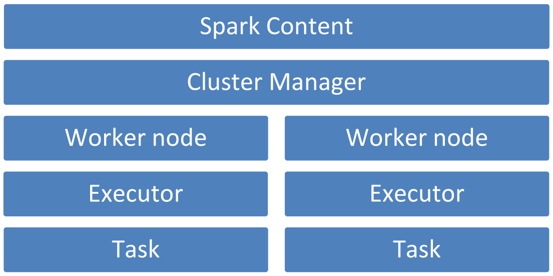


Fig.6: Architecture of Apache Spark

1. *The prime focus of spark includes resilient distributed datasets (RDD), which store data in-memory and provide fault tolerance without replication. It supports iterative computation, improves speed and resource utilization.*
2. *The foremost advantage is that in addition to MapReduce, it also supports streaming data, machine learning, and graph algorithms.*
3. *Another advantage is that, a user can run the application program in different languages such as Java, R, Python, or Scala. This is possible as it comes with higher-level libraries for advanced analytics. These standard libraries increase developer productivity and can be seamlessly combined to create complex workflows.*
4. *Spark helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk. It is possible because of the reduction in number of read or write operations to disk.*
5. *It is written in scala programming language and runs on java virtual machine (JVM) environment. Additionally, it supports java, python and R for developing applications using Spark.*

The Hadoop Ecosystem

1. HBases

Hbase is distributed column oriented database where as HDFS is file system. But it is built on top of HDFS system. HBase is a management system that is open-source, versioned, and distributed based on the BigTable of Google. It is Non-relational, distributed database system written in Java. It runs on the top of HDFS. It can serve as the input and output for the MapReduce. For example, read and write operations involve all rows but only a small subset of all columns.

2. Avro:

Avro is data serialization format which brings data interoperability among multiple components of apache hadoop. Most of the components in hadoop started supporting Avro data format. It works with basic premise of data produced by component should be readily.

3. Pig:

Pig is platform for big data analysis and processing. Pig adds one more level abstraction in data processing and it makes writing and maintaining data processing jobs very easy. Pig. can process tera bytes of data with half dozen lines of code.down and plan necessary communication protocol around node failure.

4. Hive:

Hive is a dataware housing framework on top of Hadoop. Hive allows writing SQL like queries to process and analyzing the big data stored in HDFS. It is Data warehousing application that provides the SQL interface and relational model. Hive infrastructure is built on the top of Hadoop that help in providing summarization, query and analysis.

5. Sqoop:

Sqoop is tool which can be used to transfer the data from relational database environments like oracle, mysql and postgresql into hadoop environment Sqoop is a command-line interface platform that is used for transferring data between relational databases and Hadoop.

6. Zookeeper:

Zookeeper is a distributed coordination and governing service for hadoop cluster It is a centralized service that provides distributed synchronization and providing group services and maintains the configuration information etc. In hadoop this will be useful to track if particular node is down and plan necessary communication protocol around node failure.

VI. CONCLUSION AND FUTURE WORK

Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. In this paper we discussed Hadoop and mahout tool for Big data in detail. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes using clustering algorithm. We also discussed some hadoop components which are used to support the processing of large data sets in distributed computing environments. So, it is sure that this paper helps the researches to understand the basic concepts of Big data, Hadoop and MapReduce to move further In future we can use some clustering techniques and check the performance by implementing it in hadoop.

REFERENCES

[1]. Varsha B.Bobade "Survey Paper on Big Data and Hadoop" Volume: 03 Issue: 01 | Jan-2016 , e-ISSN: 2395-0056 p-ISSN: 2395-0072

[2]. Rajeshwari.D "State of the Art of Big Data Analytics: A Survey", *International Journal of Computer Applications (0975 – 8887).*

[3]. D. P. Acharjya, Kauser Ahmed P," A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, *(IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 7, No. 2, 2016*

[4]. Chun-Wei Tsai[1], Chin-Feng Lai[2], Han-Chieh Chao[1,3,4] and Athanasios V. Vasilakos[5] "Big data analytics: a survey ",Tsai *et al. Journal of Big Data (2015) 2:21* DOI 10.1186/s40537-015-0030-3

[5]. Rotsnarani Sethy, Mrutyunjaya Panda," Big Data Analysis using Hadoop: A Survey", Volume 5, Issue 7, July 2015.

[6]. Jai Prakash Verma, Bankim Patel, Ph D, Atul Patel, Ph D," Big Data Analysis: Recommendation System with Hadoop Framework", *2015 IEEE International Conference on Computational Intelligence & Communication Technology, 978-1-4799-6023-1/15 $31.00 © 2015 IEEE DOI 10.1109/CICT.2015.86.*

[7]. *D. P. Acharjya, Kauser Ahmed P,"* A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools", *(IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 7, No. 2, 2016.*

[8]. D. Usha, A. P. S. Aslin Jenil, "A Survey of Big Data Processing in Perspective of Hadoop and Mapreduce", *International Journal of Current Engineering and Technology, Vol.4, No.2, April 2014.*

[9]. Cannataro M, Congiusta A, Pugliese A, Talia D, Trunfio P. *Distributed data mining on grids: services, tools, and applications. IEEE Trans Syst Man Cyber Part B Cyber. 2004;34(6):2451–65.*

[10]. Demirkan H, Delen D. *Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud. Decision Support Syst. 2013;55(1):412–21.*

[11]. Huai Y, Lee R, Zhang S, Xia CH, Zhang X. DOT: a matrix model for analyzing, optimizing and deploying software for big data analytics in distributed systems. *In: Proceedings of the ACM Symposium on Cloud Computing, 2011. pp 4:1–4:14.*

[12]. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, "Shared disk big data analytics with Apache Hadoop", 2012, 18-22

[13]. Aditya B. Patel, Manashvi Birla, Ushma Nair, (6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce".

[14]. kranthi Kiran B, Babu AV. A comparative study of issues in big data clustering algorithm with constraint based genetic algorithm for associative clustering. Int J Innov Res Comp Commun Eng 2014;

[15]. Bu Y, Borkar VR, Carey MJ, Rosen J, Polyzotis N, Condie T, Weimer M, Ramakrishnan R. Scaling datalog for machine learning on big data, *CoRR*, vol. abs/1203.0160, 2012. *Volume 120 – No.22, June 2015*