International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 10 | Oct -2017 www.irjet.net p-ISSN: 2395-0072

A Review paper on Big Data: Technologies, Tools and Trends

Anurag Agrahari[1], Prof D.T.V. Dharmaji Rao[2]

[1]M.tech Student, Dept. Of Computer Sci & Engg, AITAM College, Tekkali, Srikakulam, Andhra Pradesh, India.
[2]Professor, Dept of Computer Sci & Engg, AITAM College, Tekkali, Srikakulam, Andhra Pradesh, India.

Abstract - *In the past few years, tremendous changes are happening in Cloud Computing, Big Data, Communication technology and Internet of things, we are flooded with huge amount of data. By 2020, 50 billion devices are expected to be connected to the Internet. With the enhancement of internet related technology, networking and cost reducing storage technology, The world are producing huge amount of data which is not in same nature. The term of big data is referring big data set coming from various sources. In many area of science such Biology, Astrology, Physics, Business etc are producing huge amount of data which is in the form of structure and non structured data. They need to cater to find out the useful information from the massive, noisy data. The core purpose of this paper is to discuss state of art of Big Data technology. We will also cover technology and tools, suspected challenges future trends domains.*

***KeyWords*: BigData, BigData Technologies, Hadoop, Tools, Big Data Analysis, Application.**.

1. INTRODUCTION

The international population is increasing day by day. The current population of world have crossed the landmark of 7.6 billion[1] and By the end of 2016, 3.9 billion people are connect to Internet and 7 billion people are using the various mobile devices[2]. Day by day enhancement of computer and electronic technology used in the various field producing the huge amount of row data which is going to be 44 zettabytes, or 44 trillion gigabytes by 2020.Today, people and systems overload the web with an exponential generation of huge amount of data. It is doubling in size every two years. The word"Big data" is used by sociologist Mr. Charles Tilly in his article. later on the word "Big Data" is used by CNN in year of 2001 in news article.

2. BIG DATA

2.1 History of Big Data

The first major data project is created in 1937 and was ordered by the Franklin D. Roosevelt's administration in the USA. The first data-processing machine appeared in 1943 and was developed by the British to decipher Nazi codes during World War II. This device, named Colossus, searched for patterns in intercepted messages at a rate of 5.000 characters per second [3]. As of the 90s the creation of data is spurred as more and more devices are connected to the Internet. After the first super computer ware built, it was not able the process the data which are different in nature of storage, size and format. Big data creating a new challenge in handing the data and producing useful information out it. In the year of 2005, when the Hadoop ware developed by the yahoo which is built on top of Google's MapReduce. The goal was to indexing the entire World Wide Web. Today, the open source Hadoop is used by a lot of organization to tackle the huge amount of data. Many government organizations is using the big data to find out useful decision support for the betterment of societies out of it. In 2009 the Indian government has started the project named "AADHAAR "to take an iris scan, fingerprint and photograph of all of its 1.32 billion inhabitants. All this data is stored in the largest biometric database in the world. Recently, the Indian government has started big data project to find out the income tax defaulter using social media(facebook data and twitter) in year 2017.

2.2 Definition of Big Data

The growth in the volume of structured and unstructured data, the speed at which it is created and collected, and the scope of how many data points are covered. Big data often comes from multiple sources, and arrives in multiple formats.

In 2010, Apache Hadoop defined big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope". According to Wikipedia, "Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them". In addition to this, NIST has defined the term big data as "Big Data consists of extensive datasets –primarily in the characteristics of volume, variety, velocity, and variability– that require a scalable architecture for efficient storage, manipulation, and analysis."[4]. we define the term of big data as "A set of data–set refer to large amount of data which is not in same nature. It could be sensor data, Gps data, web related data and mobile generated data".

2.3 Characteristic Of Big Data

NIST has defined the characteristic of big data in four v's i.e. Value, Variety, Velocity and Value. But recent development in industry and research, one more v is added i.e. Veracity. Some other author and researcher added some more characteristic to big data, making it 7 v's (Volume, Velocity, Variety, Variability, Veracity, Visualization, and Value)[5].Kirk Borne put forward "10V ": Volume, Variety, Velocity, Veracity, Validity, Value, Variability, Venue, Vocabulary, and Vagueness[16]. Furthermore, Tom Shafer has define the 42 v's of Big Data[17].Those 42 are Vagueness, Validity, Valor, Value, Vane, Vanilla, Vantage, Variability, Variety, Varifocal, Varnish, Varmint, Vastness, Veer, Veil, Vaticination, Vault, Velocity, Venue, Veracity, Verdict ,Versed, Viral, Version Control, Vet, Vexed, Viability, Vibrant, Victual, Virtuosity, Viscosity, Visibility, Visualization, Vivify, Vocabulary, Vogue, Voice, Volatility, Volume, Voodoo, Voyage and Vulpine.

In this section We briefly discussed on 7'v of Big Data.

1) VALUME: Valume refer to the data we have. So many technologies adding the data day by day. For example, The IoT (Internet of Things) is creating exponential growth in data. It refers to the vast amounts of data generated every second. The valume of data double itself in last past 2 years. Just think of all the emails, twitter messages, photos, video clips, sensor data etc. Facebook itself registering more than 34,722 likes/sec, more than 100 terabytes of data are uploaded daily, having more than 1.7 billion users etc [6].

2) VELOCITY: Velocity refers to the speed at which new data is generated and the speed at which data moves around. Report publish by[2] clearly show that the speed of Internet have much higher than the developing country.

3) 3) VARIETY: Variety is a measure of the richness of the data representation text, images, video, audio, and sensor data. In fact, 75 percent of world data are unstructured. It can be unstructured and it can include so many different types of data from XML to video to SMS. The main task to organized the useless data into meaningful data is big task for the analysis.

4) VALUE: As it name implies, The Value factor of Big Data define as related to a size which is enormous. Size of data plays very crucial role in determining value out of data.

5) VARIABILITY: This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data. Continues change of data and its value will impact on decision making.

6) VISUALIZATION: Visualization, such as charts, graphs and other displays of the data. Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spreadsheets and reports [7].

7) VERACITY: Data being stored in different database may have anatomies, noise and unfiltered. This is most important task of big data analyst then compare to other task.

2.4 Big Data Generation

In this section, we try to explore the key role actor that generate the flood of data.

1) IoT (INTERNET OF THINGS): The phrase "Internet of Things"which is also shortly well-known as IoT is coined from the two words i.e. the first word is "Internet"and the second word is "Things". The best Definition of IoT would be "An open and comprehensive network of intelligent objects that have the capacity to auto-organize, share information, data and resources, reacting and acting in face of situations and changes in the environment "[8]. Indian government has plans to allocate 20 million USD for five internet-based, sensor-driven projects within its "IoT and 100 Smart Cities" development program by 2020. The estimated 26 billion units to be installed within the Internet of Things by 2020, according to Gartner, Inc., where in PCs, tablets and smart phones were not included. These excluded items are predicted to reach 7.3 billion units by 2020. Cisco and Intel estimate 50 billion connected devices by 2020[9].

2) BIOMEDICAL DATA AND MEDICAL DATA: Bio and Medical related field such as Bio informatics, Clinical informatics, Health related field developing the multidimensional Dataset are another source of Big Data generation. Clinical related project, Genome related project, real time health monitoring system are adding huge data. In addition to this medical imaging software such as (ct scam,MRI etc) are producing the vast amount of data even with more complex feature. A project named ProteomicsDB was started by the Swiss government to handle the genes which is size of 5.17 TB[10].

3) SOCIAL MEDIA: Nowadays, Social media (facebook, youtube, twitter, instagram etc) are producing vast amount of data generation. Insta gram having 7.1 Million of active user,34.7 billion of active photo share, average of 1650 Million like per day[11].In Youtube,300 hours of video being upload per second, more then 3.25 billion

hours video being watched by one month[12]. Same wise, the twitter having 115 million active user every month, average of 58 million tweet per day[13].More than 5 billion people worldwide call, text, tweet, and browse on mobile devices [14]. The amount of e-mail accounts created worldwide is expected to increase from 3.3 billion in 2012 to over 4.3 billion by late 2016 at an average annual rate of 6 % over the next four years. In 2012, a total of 89 billion e-mails were sent and received daily, and this value is expected to increase at an average annual rate of 13 % over the next four years to exceed 143 billion by the end of 2016[15].

4) OTHER SCIENTIFIC DATA: some other scientific Area like astronomy, Earth Related, ocean related project are generating vast amount of data. DPOSS (The Palomar Digital Sky Survey) having 3 –TB, 2MASS (The Two Micron All- Sky Survey) having 10–TB, GBT (Green Bank Telescope) having 20–PB, GALEX (The Galaxy Evolution Explorer) having 30– TB, SDSS (The Sloan Digital Sky Survey) having 40–TB,SkyMapper Southern Sky Survey having 500 TB, Pan STARRS (The Panoramic Survey Telescope and Rapid Response System) having more then 40–PB expected, LSST (The Large Synoptic Survey Telescope) having more then 200 –PB expected, SKA (The Square Kilometer Array) having more then 4.6 –EB expected[18].

2.5 Categories Of Big Data

The Big data can be categorized into three main section i.e structured, un–structured and semi–structured.

Structured data can be stored, accessed and processed in the form of fixed format is termed as a 'structured 'data. Banking related data which can be stored in the row and column format.

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in Big data analysis. In fact, world 80 % of data are unstructured.

Semi-structured data can contain both the forms of data. semi-structured data as a structured in form but it is actually not defined.

3. BIG DATA MANAGEMENT

Question comes to mind that how to manage and develop the Big Data related project. What architecture should we fallow to manage all component of Big data .Architecture of Big Data must be synchronized with support infrastructure of the institution or company. Data are generating different source which is noisy, faulty and messy. In this section, we will briefly discuss data storage tools, Hadoop and other management tools.

Figure 1 show the simple architecture of Big Data project. Another detailed life cycle of Big Data ware presented in [19], in which author has covered the detailed aspect of Big Data Architecture.

The infrastructure layer manages the system-to system interactions, stores data, network related device that is external to the system and caters to requests for data

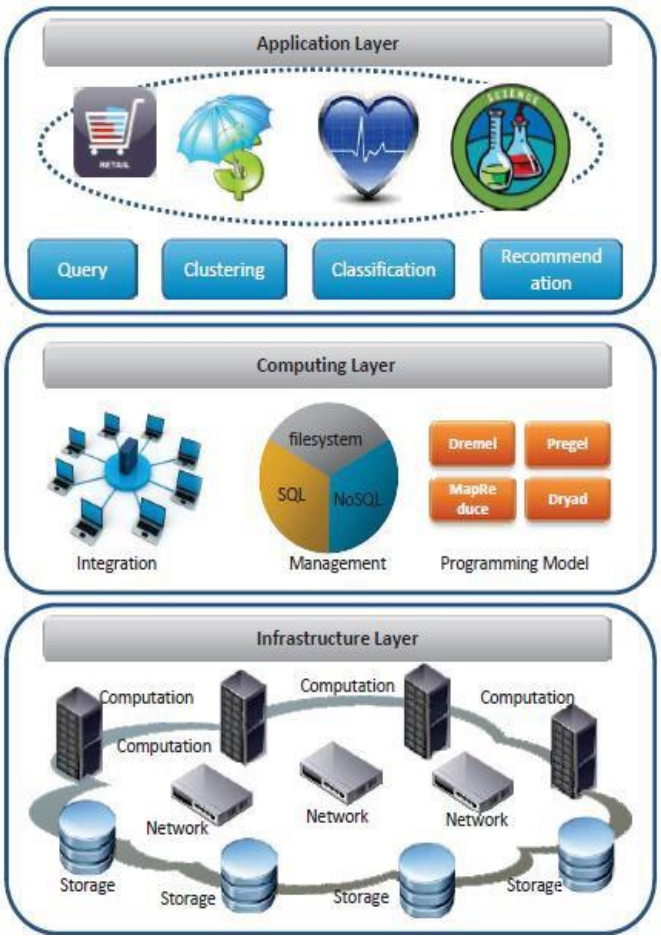

Fig. 1. A layered Big Data Architecture

Retrieval received from the other layers such as the computing layer. This layered also know as physical layer of Big Data.

This Mid Layer provides an abstraction over the physical data layer and offers the core functionalities such as data organization, access and retrieval of the data. This layer indexes the data and organizes them over the distributed store devices. The data is partitioned into blocks and organized onto the multiple repositories [42].

Analytics or Application layer consists tools and techniques, logic for developing the domain specific analytics. This layer is also known as Logical layer.

3.1 Management/Storage Tools Of Big Data

With the enhancement of computing technology, huge data can be managed without supercomputer and high cost. Data can be saved for the over the network. Many tools and techniques are available for storage management. some of them are Google Big Table, Simple DB, NoSQL, MemcacheDB[20].

3.2 Hadoop

A project ware started by Mike Cafarella and Doug Cutting to indexing nearly 1 billion page for their search engine project. In year 2003, Google has introduced the concept of Google File system known as GFS. Later on in year of 2004, the Google has given architecture of Map Reduce, which become the foundation of the framework know as Hadoop. In simple language the core of Hadoop system are Mapreduce and HDFC(Hadoop Distributed File System).In this section we will briefly discuss the component of Hadoop.

3.2.1 HDFS: HDFS is a Java-based file system that provides scalable and reliable data storage, and it was designed to span large clusters of commodity servers. Cluster contains two types of nodes. The first node is a name-node that acts as a master node. The second node type is a data node that acts as slave node. HDFS stores files in blocks, default block size of 64MB.THose files are replicated in multiples to facilitate the parallel processing of large amounts of data.

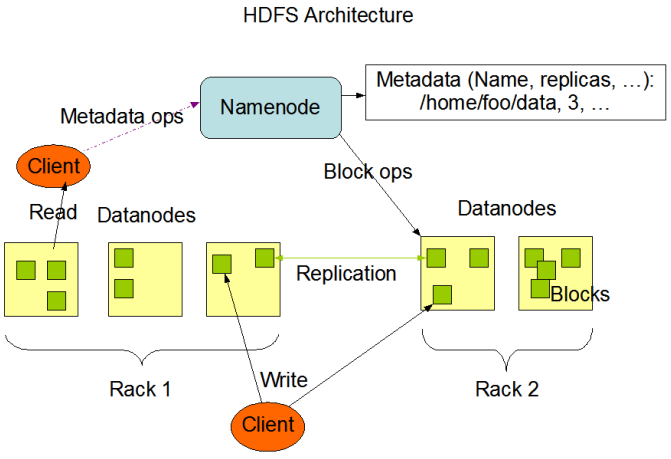

Fig 2. HDFS Architecture [35]

HDFS stores huge data and for storing such huge data, the files are stored across multiple machines. These files are stored in redundant manner so that it can prevent the system from possible data losses in case of failure. HDFS provides parallel processing also. This architecture consists of a master server and a single Name-Node that handles the file system namespace and regulates access to files by clients. In this architecture, there is one Data Node in each cluster. This manages storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files. In HDFS internally, a file is split into one or more blocks. These blocks are stored in a set of Data Nodes. The Name-Node performs file system namespace operations like closing, opening and renaming directories and files. It also controls the mapping of blocks to Data Nodes.

3.2.2 MapReduce Frameworks

Map Reduce is a program model for distributed computing based on java. It is a processing technique. The Map Reduce algorithm includes two important tasks, namely Map and Reduce. The term Map Reduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map Reduce implies, the reduce job is always performed after the map job.

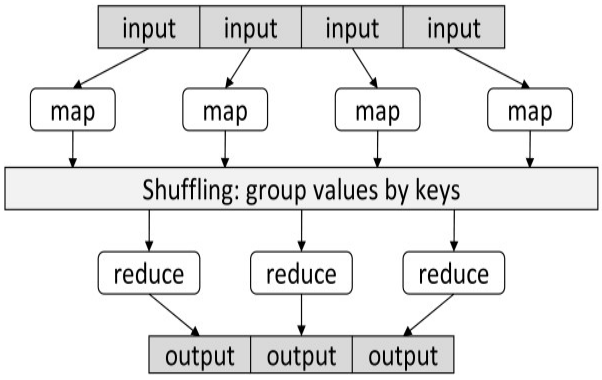

Fig. 3. MapReduce Architecture [35]

In MapReduce, it is easy to scale data processing over multiple computing nodes. Under this model, the data processing primitives are called mappers and reducers. Writing the application in the MapReduce form is not so easy but once it is written it allows scaling the application to run over thousands or even tens of thousands of machines in a cluster.

In MapReduce, it is easy to scale data processing over multiple computing nodes. Under this model, the data processing primitives are called mappers and reducers.MapReduce program runs in three stages, as shown in Fig. 3, namely map stage, shuffle stage, and reduce stage.

Map Phase The input data is stored in the form of files in the Hadoop file system (HDFS). This input file is then passed to the mapper function line by line. These data is then processed by the mapper and several small chunks of data is created.

Suffule Phase Worker nodes redistribute data based on the output keys (produced by the "map()" function), such that all data belonging to one key is located on the same worker node.

Reduce Phase Worker nodes now process each group of output data, per key, in parallel. The Reducer takes output of mapper and the process. It and generates new set of output which will be stored in HDFS.

3.2.3 Hadoop Ecosystem

Hadoop ecosystem can have multiple tools which can be categories in four main sectors according to their working. In this section we tried to briefly cover all the available tools.

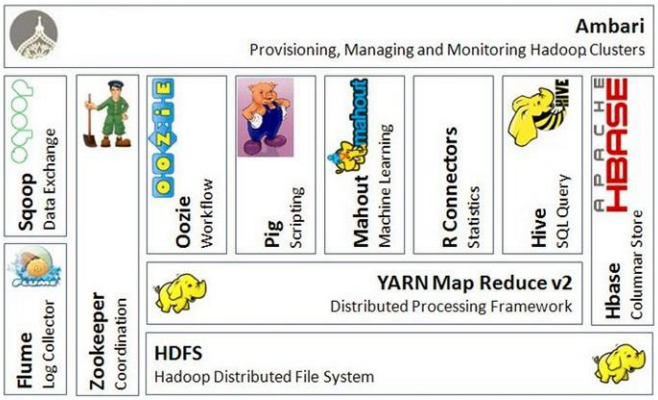

Fig. 4. Hadoop Ecosystem[21]

YARN: YARN is stand for Yet Another Resource Negotiator, is a technology used for cluster management, introduced in Hadoop version 2.YARN is now considered as a largescale, distributed operating system for big data applications.It is originally labelled by Apache as a redesigned resource manager. The YARN Infrastructure is responsible for providing the computational resources for example, CPUs, memory, etc, needed for application executions.it is divided into two sub part known as Resource manager and Node manager.

The Resource Manager (one per cluster) is the master. It knows where the slaves are located (Rack Awareness) and how many resources they have. It runs several services, the most important is the Resource Scheduler which decides how to assign the resources.

The Node Manager (many per cluster) is the slave of the infrastructure. When it starts, it announces himself to the Resource Manager. Periodically, it sends an heartbeat to the Resource Manager[22].

Hbase: Apache HBase provides random, real time access to your data in Hadoop. It was created for hosting very large tables, making it a great choice to store multistructured or sparse data. HBase is accessible through application programming interfaces (APIs) such as Thrift, Java, and representational state transfer (REST). These APIs do not have their own query or scripting languages. By default, HBase depends completely on a ZooKeeper instance[19].The characteristic of Hbase are Fault tolerant,usable and Fast then other technology[23].

ZooKeeper: ZooKeeper is a distributed, open-source coordination service for distributed applications. It contains master and slave nodes and stores configuration information. It is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications[24].

HCatalog: HCatalog is a table and storage management layer for Hadoop that enables users with different data processing tools like Pig, MapReduce,enable it to more easily read and write data on the grid. HCatalog ensures that users need not to worry about where or in what format their data is stored.

Pig: Apache Pig allows Apache Hadoop users to write complex MapReduce transformations using a simple scripting language called Pig Latin. Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The characteristic of Pig are extensibility, easy programmed and self optimizing.

Mahout: Apache Mahout is an open source project that is primarily used in producing scalable machine learning algorithms. It implements popular machine learning techniques such as: Recommendation, Classification, Clustering[25]. It is divided into four main groups: collective, filtering, categorization, clustering, and mining of parallel frequent patterns. The Mahout library belongs to the subset that can be executed in a distributed mode and can be executed by MapReduce[19].

Hive: Hive was developed by Facebook initially. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. A command line tool and JDBC driver are provided to connect users to Hive. Hive is a sub platform in the Hadoop ecosystem and produces its own query language, known as HiveQL. Feature of Hive are fast, scalable and compatibility.

Oozie: Apache Oozie is a Java Web application used to schedule Apache Hadoop jobs.Oozie combines multiple jobs sequentially into one logical unit of work.Oozie can also schedule jobs specific to a system, like Java programs or shell scripts.it is a scalable, reliable and extensible system.

There are two basic types of Oozie jobs[26]: i.e Oozie Workflow and Oozie coordinator.

Oozie Workflow jobs are Directed Acyclical Graphs (DAGs), specifying a sequence of actions to execute.

Oozie Coordinator jobs are recurrent Oozie Workflow jobs that are triggered by time and data availability.

Avro: Avro is a remote procedure call and data serialization framework developed within Apache's Hadoop project. It uses JSON for defining data types and protocols, and serializes data in a compact binary format. Avro data is stored in a file, its schema is stored with it, so that files may be processed later by any program.

Chukwa: Apache Chukwa is an open source data collection system for monitoring large distributed systems. Apache Chukwa is built on top of the Hadoop Distributed File System (HDFS) and Map/Reduce framework and inherits Hadoop's scalability and robustness.

Flume: Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS) and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. Feature of Flume of stream data, Insulate systems, Guarantee data delivery, Scale horizontally [27].

3.3 Other Tools

Falcon: Falcon allows an enterprise to process a single massive dataset stored in HDFS in multiple ways for batch, interactive and streaming applications.it is basically data management tool to provide a framework to define, deploy, and manage data pipelines.

Atlas: The ATLAS Distributed Data Management system requires accounting of its contents at the metadata layer. Atlas is a scalable and extensible set of core foundational governance services which provide Data Classification, Centralized Auditing, Search Lineage, Security and Policy Engine[28].Apache Atlas provides scalable governance for Enterprise Hadoop that is driven by metadata. The ATLAS Distributed Data Management system requires accounting of its contents at the metadata layer.

Tez: Tez improves the MapReduce paradigm by dramatically improving its speed, while maintaining MapReduce's ability to scale to petabytes of data. It is a DAG (Directed acyclic graph) architecture.

Sqoop: Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. It has two main role i.e Sqoop import and Sqoop export.

The import tool imports individual tables from RDBMS to HDFS. Each row in a table is treated as a record in HDFS.

The export tool exports a set of files from HDFS back to an RDBMS. The files given as input to Sqoop contain records, which are called as rows in table. Apart from this, it has other function like load balancing, efficient data analysis etc.

Kafka: Apache Kafka was originated at LinkedIn and Later became an open sourced Apache project in 2011, then First-class Apache project in 2012. Kafka is written in Scala and Java[29].It support wide range of use cases as a general purpose messaging system for scenarios where high throughput,
reliable delivery, and horizontal scalability are important. Stream Processing, Website Activity Tracking, Metrics Collection and Monitoring, Log Aggregation are the common uses of Kafka.

Accumulo: Accumulo was originally developed at the National Security Agency.Apache Accumulo is a key/value store based on the design of Google's BigTable. Accumulo stores its data in Apache Hadoop's HDFS and uses Apache Zookeeper for consensus. The common feature of Accumulo are Table design and configuration, Integrity and availability, Performance, Data Management [30].

Storm: Storm was originally created by Nathan Marz and team at BackType. BackType is a social analytics company. Later, Storm was acquired and open-sourced by Twitter. Storm has many use cases: real-time analytics, online machine learning, continuous computation, distributed RPC, ETL, and more. Storm is fast: a benchmark clocked it at over a million tuples processed per second per node.

Solr: It is indexing tools which do indexing via XML, JSON, CSV or binary over HTTP.Solr is an open-source search platform which is used to build search applications. It was built on top of Lucene (full text search engine).It was Yonik Seely who has created Solr in 2004 in order to add search capabilities to the company website of CNET Networks. In Jan 2006, It was made an open-source project under Apache Software Foundation.

Spark: Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application.

Ranger: Using the Apache Ranger console, security administrators can easily manage policies for access to files, folders, databases, tables, or column.The Ranger Key Management Service (Ranger KMS) provides a scalable cryptographic key management service for HDFS "data at rest "encryption.

Knox: Apache Knox provides a configuration driven method of adding new routing services. The Apache Knox Gateway ("Knox") provides perimeter security so that the enterprise can confidently extend Hadoop access to more of those new users while also maintaining compliance with enterprise security policies.

Ambari: A completely open source management platform for provisioning, managing, monitoring and securing Apache Hadoop clusters[31]. Ambari enables System Administrators to:

Provision a Hadoop Cluster Ambari provides a step-by-step wizard for installing Hadoop services across any number of hosts. It handles configuration of Hadoop services for the cluster.

Manage a Hadoop Cluster Ambari provides central management for starting, stopping, and reconfiguring Hadoop services across the entire cluster.

Monitor a Hadoop Cluster Ambari provides a dashboard for monitoring health and status of the Hadoop cluster. Ambari leverages Ambari Metrics System for metrics collection. Ambari leverages Ambari Alert Framework for system alerting and will notify you when your attention is needed (e.g., a node goes down, remaining disk space is low, etc).

Apart from this, Ambari is very useful in large scale cluster setup because it provide consistent, secure platform for operational control. It is Simplified Installation and Configuration and Management, Centralized Security Setup, Full Visibility into Cluster Health and Highly Extensible and Customizable.

Phoenix: Apache Phoenix abstract away the underlying data store by enable you to query the data with standard SQL via JDBC driver. Apache Phoenix provides features such as secondary indexes to help you speed up the queries without relying on specific row key designs.

NiFi: Apache NiFi is an integrated data logistics platform for automating the movement of data between disparate systems. NiFi was built to automate the flow of data between systems. It provides real-time control that makes it easy to manage the movement of data between any source and any destination. It is data source agnostic, supporting disparate and distributed sources of differing formats, schemas, protocols, speeds and sizes such as machines, geo location devices, click streams, files, social feeds, log files and videos and more.

HAWQ: HAWQ is a Hadoop native SQL query engine that combines the key technological advantages of MPP database with the scalability and convenience of Hadoop. HAWQ reads data from and writes data to HDFS natively. Move and analyze entire workloads, while simplifying management and expanding the breadth of data access and analytics, all natively in Hadoop.

Zeppelin: Apache Zeppelin is a new and incubating multipurposed web-based notebook which brings data ingestion, data exploration, visualization, sharing and collaboration features to Hadoop and Spark. Interactive browser-based notebooks enable data engineers, data analysts and data scientists to be more productive by developing, organizing, executing, and sharing data code and visualizing results without referring to the command line or needing the cluster details.

DRUID: Druid is an open-source analytics data store designed for business intelligence (OLAP) queries on event data. Druid provides low latency (real-time) data ingestion, flexible data exploration, and fast data aggregation.

Slider: Apache Slider is a application to deploy existing distributed applications on an Apache Hadoop YARN cluster, monitor them and make them larger or smaller as desired –even while the application is running.It is a YARN-based framework for Long-running Applications In Hadoop. Slider "slides" these long-running services (like Apache HBase, Apache Accumulo and Apache Storm) onto YARN, so that they have enough resources to handle changing amounts of data, without tying up more processing resources than they need.

Metron: Apache Metron is a big data cyber security application framework that enables a single view of diverse, streaming security data at scale to aid security operations centers in rapidly detecting and responding to threats. It is a next generation SOC (security operations center) data analytics and response application that integrates a variety of open source big data technologies into a centralized tool for security monitoring and analysis[32].

Cloudbreak: A tool for provisioning and managing Apache Hadoop clusters in the cloud.Cloudbreak, as part of the Hortonworks Data Platform, makes it easy to provision, configure and elastically grow HDP clusters on cloud infrastructure. Cloudbreak can be used to provision Hadoop across cloud infrastructure providers including Amazon Web Services, Microsoft Azure, Google Cloud Platform and OpenStack[33].

Impala: Cloudera Impala is an addition to tools available for querying big data. Impala does not replace the batch processing frameworks built on MapReduce such as Hive. Hive and other frameworks built on MapReduce are best suited for long running batch jobs, such as those involving batch processing of Extract, Transform, and Load (ETL) type jobs[34].

4. BIG DATA ANALYSIS

Big Data analysis refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases using data analysis techniques. Big Data analysis can be applied to special types of data. Big Data Analytics can be defined as the use of advanced analytic techniques on big data [36].Analysis of Big Data involves various data mining techniques to find the objectives. In this section, we will briefly discuss various technique of Data mining which is frequently used in Big Data Analysis.

Machine Learning: .Machine learning is a mature and well-recognized research area of computer science, mainly concerned with the discovery of models, patterns, and other regularities in data. Machine learning to bring computer to learn complex patterns and make intelligent decisions based on it [10].

Cluster Analysis: Clustering is an unsupervised technique used to classify large datasets in to correlative groups. No predefined class label exists for the data points or instances. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups and the groups are called as clusters. Clustering can be classified into Partitioning clustering, Hierarchical clustering, Density based clustering, Model based clustering, Grid based clustering[37].

Correlation Analysis: Correlation is a technique for investigating the relationship between two quantitative, continuous variables.Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables. In other words, it determining the low of relation among variables.

Statistical Analysis: A collection of automated or semi automated techniques for discovering previously unknown patterns in data, including relationships that can be used for prediction of user-relevant quantities. There are two computational barriers for big data analysis: 1) the data can be too big to hold in a computers memory; and 2) the computing task can take too long to wait for the results[38]. These barriers can be approached either with newly developed statistical methodologies and/or computational methodologies.

Regression Analysis: Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor).Regression analysis is an important tool for modeling and analyzing data. Seven regression techniques i.e Linear Regression, Logistic Regression, Polynomial Regression, Stepwise Regression, Ridge Regression, Lasso Regression, ElasticNet Regression are used in Big Data analysis. Most frequent techniques for Big Data analysis are linear analysis and polynomial analysis

5. BIG DATA APPLICATION

Big Data Analysis provide the useful attributes via suggestion, judgment, decision and support form huge amount of data. In this section, we briefly discuss the application of Big Data.

5.1 Text Data Analysis

Another name for text analytics is text mining. A good reason for using text analytics might be to extract additional data about customers from unstructured data sources. Text analytics involve statistical analysis, computational linguistics, and machine learning. Text analytics enable businesses to convert large volumes of human generated text into meaningful summaries, which support evidence-based decision-making. For example, text analytics can be used to predict stock market based on information extracted from financial news[39].Most text mining techniques are based on Natural Language Processing(NLP).NLP allows computers to analyze, interpret, and even generate text. Some NLP-based techniques have been applied to text mining, including information extraction, topic models, text summarization, classification, clustering, question answering, and opinion mining.

5.2. Social Media Analysis

Social network analysis refers to methods used to analyze social networks, social structures made up of individuals or organizations, which are connected by one or more specific types of interdependency, such as friendship, common interest, financial exchange, or relationships of beliefs, etc. A social network is a set composed of nodes and links between each two node[40].a social structure is some mode of stable relationships and is always expressed as a network formed by a series of nodes (actors) and links that represent relationships among nodes. There are three basic substances of social network structures: the actor, relation, and network.The social media service can be divided into categories i.e. Link based analysis and content based analysis[41]. link-based structural analysis has always been committed on link prediction, community discovery, social network evolution, and social influence analysis, etc. SNS may be visualized as graphs, in which every vertex corresponds to a user and edges correspond to the

correlations among users. Content-based analysis in SNS is also known as social media analyses include text, multimedia, positioning, and comments.

5.3 Mobile Data Analysis

Mobile analytics is the practice of collecting user behavior data, determining intent from those metrics and taking action to drive retention, engagement, and conversion. The progress in wireless sensor, mobile communication technology, and stream processing add a new research area. With it, the developer can develop the health related and business related application there are so many other Application Area of Big Data such as multimedia data analysis, surveillance analysis, weather forecasting etc.

6. CHALLENGES AND OPEN ISSUE

The success of Big Data in the enterprises requires biggest cultural and technological change. Enterprise wise strategy required to derive the business value by integrating the available traditional data. Biggest challenges in Big Data analysis are Heterogeneity and Incompleteness, Scalability, Timeliness and security of the data. Privacy is one of the major concerns for the outsourced data. Policies have to be deployed and rule violators to identify for avoiding the misuse of data. Data integrity is a challenge for the data available in cloud platform.

7. CONCLUSION

In this paper concept of Big Data and various technologies has been surveyed which are used handle the big data. This paper discussed architecture of Big Data using Hadoop HDFS distributed data storage under which its different components are also explained. One area that sees a lot of potential in big data is the mining industry. For an industry that does trillions of dollars in business every year, big data is not seen as a luxury but as a necessity. The main objective of this paper was to make a survey of various Big Data architecture, its handling techniques which handle a huge amount of data from different sources and improves overall performance of systems and its applications It's no secret that big data has led to major changes within the business world. The uses of big data are many and can apply to areas that many might not have thought of before. One area that sees a lot of potential in big data is the mining industry. For an industry that does trillions of dollars in business every year, big data is not seen as a luxury but as a necessity. Researchers are continuously working on the algorithms to mine big data efficiently and quickly. Furthermore, some of tools which are used in the analytics and management are discussed.

REFERENCES

1) Worldmeters, "Real time world Statistics", 2017. http://www.worldometers.info/world-population/ access on 3/09/2017.

2) "ICT Facts Figures: The World in 2015"[https://www.itu.int/en/ITUD/Statistics/Documents/facts/ICTFactsFigures2015.pdf]February 2016

3) https://datafloq.com/read/big-data-history/239 access on 3/9/2017.

4) https://www.nist.gov/publications/nist-big-data-interoperabilityframework-volume-1-definitions. access on 3/9/2017.

5) https://www.impactradius.com/blog/7-vs-big-data/ access on 02/09/2017.

6) Facebook statistc, http://www.statisticbrain.com/facebook-statistics/ access on 12/09/2017.

7) https://www.impactradius.com/blog/7-vs-big-data/ access on 5/09/2017.

8) Somayya Madakam, R. Ramaswamy, Siddharth Tripathi,"Internet of Things (IoT): A Literature Review",Journal of Computer and Communications,2015, 3, 164-173.

9) http://www.tikitoki.com/timeline/entry/438056/Internet-of-Things-Timeline/ access on 01/09/2017.

10) Jake Luo, Min Wu, Deepika Gopukumar, and Yiqing Zhao,Big Data Application in Biomedical Research and Health Care: A Literature Review.Biomed Inform Insights. 2016.

11) instagram statistc ,http://www.statisticbrain.com/facebook-statistics/ access on 12/09/2017.

12) youtube statistc, http://www.statisticbrain.com/facebook-statistics/ access on 12/09/2017.

13) Twitter statistc http://www.statisticbrain.com/facebook-statistics/ access on 12/09/2017.

14) P. Russom, "Big data analytics," TDWI Best Practices Report, Fourth Quarter, 2011.

15) S. Radicati and Q. Hoang, Email Statistics Report, 2012?2016, The Radicati Group, London, UK, 2012.

16) Kirk Borne,http://www.mapr.com/blog/top-10-big-data-challenges-%E2%80%93-serious-look-10-big-data-v%E2%80%99s

17) TomShafer,https://www.elderresearch.com/company/blog/42-v-of-bigdata. access on 15-09-2017.

18) Yanxia Zhang,Yongheng Zhao,Astronomy in the Big DataEra,https://datascience.codata.org/articles/10.5334/dsj-2015-011/print access on 05-08-2017.

19) Nawsher Khan,et.al,Big Data: Survey, Technologies, Opportunities, and Challenges,Hindawi Publishing Corporation,the Scientific World Journal,Volume 2014, Article ID 712826, 18 pages.

20) Min Chen Shiwen Mao Yunhao Liu,Big Data: A Survey,Mobile Networks and Application 19:171:209,2014

21) Hadoop Ecosystem, https://hortonworks.com/ecosystems/ access on 05-09-2017.

22) http://ercoppa.github.io/HadoopInternals/Hadoop Architecture Overview.access on 3/09/2017.

23) Hbase, https://hortonworks.com/apache/hbase/ access on 3/09/2017.

24) ZooKeeper, https://zookeeper.apache.org/ access on 3/09/2017.

25) Mahout,https://www.tutorialspoint.com/mahout/mahout-introduction.access on 3-09-2017.

26) Oozie,https://hortonworks.com/apache/oozie/ access on 3/09/2017.

27) Flume,https://hortonworks.com/apache/flume/ access on 2/09/2017.

28) Atlas,http://atlas.apache.org/ access on 3/09/2017.

29) kafka,https://www.tutorialspoint.com/apache-kafka/ access on 3/09/2017.

30) Accumulo,https://hortonworks.com/apache/accumulo/ access on 03/08/2017.

31) Ambari,https://ambari.apache.org/ access on 2/08/2017.

32) Metron,https://hortonworks.com/apache/metron/ access on 3/07/2017.

33) Cloudbreak,https://hortonworks.com/opensource/cloudbreak/ access on4/09/2017.

34) Impala,https://www.cloudera.com/documentation/enterprise/5 3 x/topics/ impala intro.html access on 04/08/2017.

35) MapReduce,https://hadoop.apache.org/docs/r1.2.1/hdfs design.html

36) P. Russom, et al. Big data analytics, TDWI Best Practices Report, FourthQuarter.

37) Keshav Sanse, Meena Sharma,Clustering methods for Big data analysis,(IJARCET) Volume 4 Issue 3, March 2015.

38) Chun Wang, Ming-Hui Chen, Elizabeth Schifano, Jing Wu, and Jun Yan,Statistical Methods and Computing for Big Data.arXiv:1502.07989.

39) Chung, W. (2014). BizPro: Extracting and categorizing business intelligence fac-tors from textual news articles. International Journal of Information Management,34(2), 272284.

40) R. Albert, A. L. Barabsi, Statistical mechanics of complex networks. Reviews of Modern Physics, vol. 74, 2002, pp. 47-97.

41) Aggarwal CC (2011) An introduction to social network data analytics. Springer.