- Volume: Organization collects data from different sources including various business transactions, social media. At present the existence of data is in petabytes and is assumed to increase to more than zettabytes in coming era [2] .The social networking sites are producing data in order of terabytes and this amount of data is definitely difficult to be handled using the existing traditional systems.
- Velocity: Velocity basically deals with the rate at which data arrives from various nodes to one server. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows [3].
- Variety: Data that is being produced does not include only a single category of data but it also includes semi structured data in addition to traditional data from various sources. All this data is totally different that consist of raw, structured, semi-structured and even unstructured data which is difficult to be handled by existing traditional analytics systems.

CATEGORIES OF BIG DATA

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

- ✓ **Structured data**: 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.
- ✓ **Unstructured Data:** 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:** Semi structured data can contain both the forms of data. semi-structured data as a structured in form but it is actually not defined.

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.

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[19]. In simple language the core of Hadoop system are Mapreduce and HDFC(Hadoop Distributed File System)

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 [20].

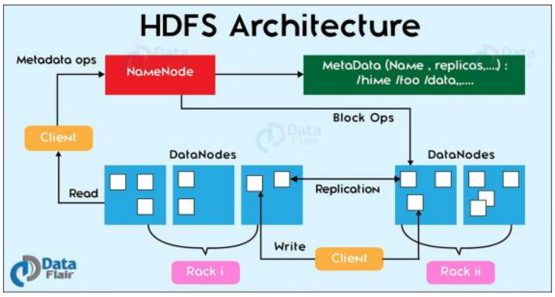

Fig2: HDFS Architecture

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.

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[21]. 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 jobs the output from a map as input and combines those data tuples into a smaller set of sequence name reduce implies, the reduce job is always performed after the map job.

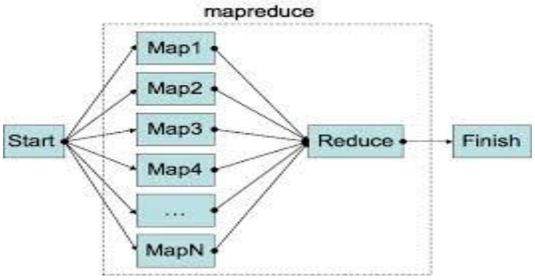

Fig3: Map reduce Framework

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 [19].

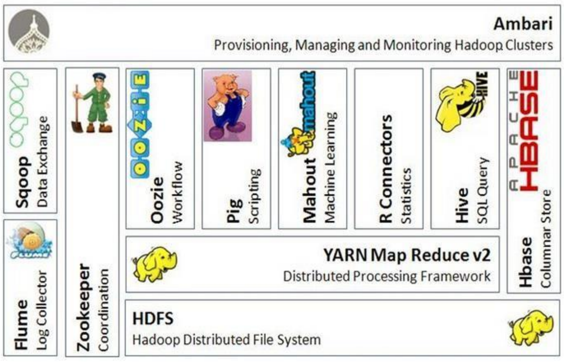

Fig 4: Hadoop Ecosystem

- **YARN:** YARN is stand for Yet Another Resource Negotiator, is a technology used for cluster management, introduced in Hadoop. The YARN Infrastructure is responsible for providing the computational resources for example, CPUs, memory, etc, needed for application executions.
- **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.
- **ZooKeeper:** ZooKeeper is a distributed, open-source coordination service for distributed applications. It contains master and slave nodes and stores configuration information
- **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.
- **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.
- **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.
- **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.
- **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.
- **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.
- **Ambari:** A completely open source management platform for provisioning, managing, monitoring and securing Apache Hadoop clusters.

BIG DATA ANALYSIS

Big Data Analytics can be defined as the use of advanced analytic techniques on big data Analysis of Big Data involves various data mining techniques to find the objectives.

- **Machine Learning:** .Machine learningmainly 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
- **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.
- **Correlation Analysis:** Correlation is a technique for investigating the relationship between two quantitative, continuous variables.
- **Statistical Analysis:** A collection of automated or semi automated techniques for discovering previously unknown patterns in data.
- **Regression Analysis:** 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.

BIG DATA APPLICATION

Big Data Analysis provides 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.

- **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.
- **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.

- **Mobile Data Analysis**

Mobile analytics is the practice of collecting user behaviour 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.

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.

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. https://datafloq.com/read/big-data history/239 access on 3/9/2017.
3. https://www.nist.gov/publications/nist-big-datainteroperabilityframework-volume-1-definitions. access on 3/9/2017.
4. https://www.impactradius.com/blog/7-vs-bigdata/ access on 02/09/2017.
5. Facebook statistc, http://www.statisticbrain.com/facebook-statistics/ access on 12/09/2017.
6. https://www.impactradius.com/blog/7-vs-bigdata/ access on 5/09/2017.
7. Somayya Madakam, R. Ramaswamy, Siddharth Tripathi,"Internet of Things (IoT): A Literature Review",Journal of Computer and Communications,2015, 3, 164-173.
8. http://www.tikitoki.com/timeline/entry/438056/I nternet-of-Things-Timeline/ access on 01/09/2017.
9. 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.
10. instagram statistc http://www.statisticbrain.com/facebookstatistics/ access on 12/09/2017.
11. youtube statistc, http://www.statisticbrain.com/facebook-statistics/ access on 12/09/2017.
12. Twitter statistc http://www.statisticbrain.com/facebook-statistics/ access on 12/09/2017.
13. P. Russom, "Big data analytics," TDWI Best Practices Report, Fourth Quarter, 2011.
14. S. Radicati and Q. Hoang, Email Statistics Report, 2012? 2016, The Radicati Group, London, UK, 2012.
15. KirkBorne,http://www.mapr.com/blog/top-10big-data-challenges-%E2%80%93-serious-look-10 big-data v%E2%80%99s
16. Tom Shafer, https://www.elderresearch.com/compa ny/blog/42-v-of-bigdata. access on15-09-2017.
17. 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.
18. Min Chen Shiwen Mao Yunhao Liu,Big Data: A Survey,Mobile Networks and Application 19:171:209,2014
19. Hadoop Ecosystem, https://hortonworks.com/ecosystems/ access on 05-09-2017.
20. Hbase,https://hortonworks.com/apache/hbase/ access on 3/09/2017.
21. MapReduce,https://hadoop.apache.org/docs/r1.2.1 /hdfs design.html