A Survey on Big Data and its Applications using Hadoop

[1]Manpreet Kaur, [2]Dr. Vijay Dhir
[1]Reseacrh Scholar MTech, [2]Professor of CSE Department
[1]Computer Science and Engineering,
[1]Sant Baba Bhag Singh University, Jalandhar, India

Abstract : The 'Big Data' concept is very famous and useful technology in now a days. Big data can be collected from so many sources and it has many forms. It is very difficult to analyze, manage and store the big data in traditional database thus hadoop framework is used to store and manage the big data. Hadoop is an open source framework which has HDFS, Map Reduce and YARN components. By hadoop data is divided and managed in many blocks. In this paper we will discuss about big data analytics and applications of big data. So many applications of big data are useful for business, which increase the business performance. Big data technology is capable to parallely manage the data in huge amount.

***Index Terms* - Big Data, 4 V's of big data, Hadoop, YARN, HDFS, Map Reduce.**

1. INTRODUCTION

1.1 Big Data: Definition

Big data is the new topic of today that is increasing with speed. Big data is a combination of large amount of data sets. Observing the size of the increasing data, big data came into existence. In the decade of 70s and 80s it was estimated that the size of business data to be changed from MB to GB. Due to the emergence of digital technology in the late 1980s, data storage increased from GB to TB. Parallelization was presented to upgrade computer processing. late 1990 In the Internet era, unstructured and structured data production began to appear from web pages, which came into being PB storage for storing data.

Big data analysis is a very crucial stage, which is to collect the meaningful values of Big Data, based on the conclusions of them, to make a decision-making system for an organization that can compete in the market competition. There are many applications of Big Data, such as in media and communication, data is produced in audio, video, and text form. The customer's behavior in the insurance industry is analyzed by big data so that fraudsters may be detected and deal with a better customer. In Transportation you can analyze the behavior of travelling and route planning with the help of big data.

1.2 Big data comes where from?

We acquire the big data from various sources like social medias (facebook, twitter, yahoo), search engine and online shopping sites. By analyzing the customer's behavior in online shopping sites, they are categorized. The product is predicted by analyzing the purchasing capacity and purchasing range of customers, which product can be more sold.

Data may be in any form like structured data, unstructured data and semi-structured data. The structured data is stored in the table form like (rows and columns). SQL queries apply to manage structured data. Unstructured data can be in any form that is collected from web pages eg. audio, video, text, books and images etc. The structured and unstructured data together constitutes semi-structured dataeg. XML and JSON.

2. Characteristics of Big Data: Big data can be defined in 4V's. They are Volume, Velocity, Variety and Veracity.

2.1 Volume of data : The volume of data is related to the size and quantity of data. After analyzing the data based on the value of the data, it is seen that the given data can be considered as big data or not. The data has come in TB and PB over MB, it is estimated that the volume of data can come in ZB(Zeta byte) in the coming years.

2.2 Variety of data: The data produced which is not of the same category is in the traditional data, it is of heterogeneous nature eg. text, images, web pages, audio, log files and sensor information etc.

2.3 Velocity of data: Velocity is related to the generated speed of the data to calculate the speed of the data to be produced and processed. Big data is time sensitive. Social sites are generated in data milliseconds, so credit card is completed in a few seconds and stores data in the database.

2.4 Veracity of data: Veracity means the quality of data whether it is accurate and certain or uncertain. Data is uncertain due to the inconsistency and incompleteness. [11]

3. Applications of Big Data: There are various applications of big data are following:

3.1 Fraud Detection: This is an application of big data in which claims and transactions are analyzed whether these are valid or not. It is very useful for business. In this platform the identification of customer is analyzed on large scale and decided that customer is fraud or not and he/she is able to claim and perform the transaction.

3.2 Improving Healthcare: Big data is being proved a very powerful technique in medical science. DNA test can be detected by the big data technology and after studying the result of test any disease can be predicted and solution can be found. This technology is capable to monitor the premature babies and sick babies. Through this technique the heartbeat is recorded of unborn babies and with the help of algorithm, which analyze the breathing pattern, babies can be saved from forthcoming ailments.

3.3 The Role of Big Data in Medicine: Research about medical science is growing large and complex. The main motive of big data in medicine is to provide better health and make predictive models, which may better understand the ailment of patients and provide better treatment. In this case, patient is examined in routine and internal image is captured in motion and better treatment is provided. In future, wearable devices and mobile apps will be available which will analyze the symptoms and the temperature of the patient and suggest the medicine.

3.4 IT Logs Analytics: IT department generates data logs and trace data in large quantity. A lot of data remains unexamined without big data that is why the solution is exercised by it that trace data is to be used where and how? Organizations are able to find out the new methods to escape from the problems with the help of large scale patterns. The performance of organizations is increased by it.

3.5 Social Media: It is an important topic to analyze the activity of social medias. Everyone uses social media whether it may be about the page of company to be liked, chatting or the complaint of any product. It provides the facility to highlight of responses by product holder of any product.

4. Architecture of Hadoop Framework: Hadoop was developed by computer scientists Doug Cutting and Mike Cafarella in 2006**.** When Doug Cutting worked in Yahoo he named hadoop framework from his son's elephant toy. It is used for distributed processing and distributed storage of large data sets. Mapreduce and HDFS are main components of Hadoop. Firstly, HDFS is used in hadoop for distributed storage. It allows us to store and process the large amount of data into small blocks via clusters. It is capable to store the data on various servers, Secondly Mapreduce is used for parallel processing. With the help of Mapreduce a lot of data is processed simultaneously.

Hadoop is a popular framework to analyze the big data by the organization and researcher. We get Documentation, work scheduling and source code by the hadoop framework. It consists various components like Jaql, hive, pig, sqoop, flume, zookeeper and oozie etc. [11]

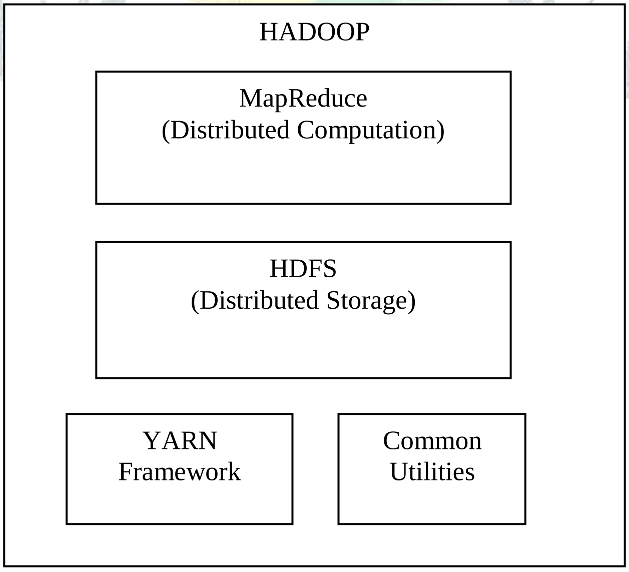

Fig. 1 Hadoop Framework [13]

4.1 Common libraries: Other hadoop modules require java libraries. It provides operating system and file system. It also provides the facility to contain the java files that are required to start hadoop. It is also called as hadoop common.

4.2 YARN framework: The full form of YARN is Yet Another Resource Negotiator. The main function of YARN is to provide the resources or devices to the each task eg. CPU, memory etc. It is also known as job scheduler. There are three components of YARN framework:

4.2.1 Resource Manager: Resource manager is called scheduler also because it decides that which resource is to allocate to which task or which applications are to be provided. Resource manager has two daemons such as scheduler and application manager.

4.2.2 Node Manager: It is second component of a YARN framework. Resource manager acquires the heartbeat from the node manager. Resource manager gets instructions from node manager that how to manage the resource? and how to launch the applications?

4.2.3 Application Master: Application master checks whether all the jobs and tasks are working properly or not. It observes the status of task.

4.3 HDFS: HDFS stands for Hadoop Distributed File System. IT is in huge demand now a days because data is increasing fastly from the social sites and shopping sites etc. Thus it is very difficult to store, manage and process the data at simultaneously. It consumes more time. To lesser the time consumption data is divided into various fixed size blocks. The size of these blocks are 64 MB or it can be increased upto 128MB according to the data. These are stored as clusters on one or more than one machine. This process is called as HDFS. It acts as master slave architecture. This architecture has name node and data node in which name node acts as a master node and another data node acts as slave node.

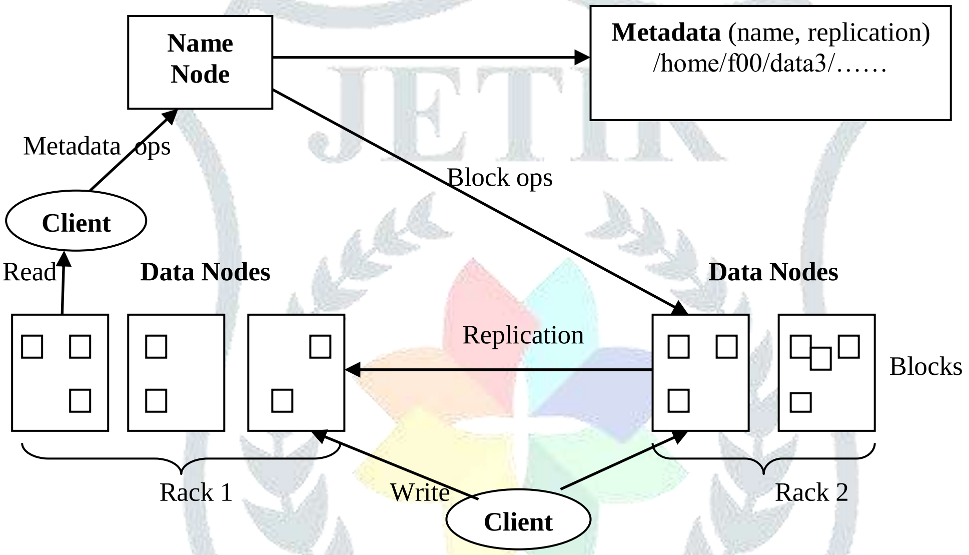

Fig 2. HDFS [12]

Name node manages the file system and executes the operations on the file as to rename the files, to open and close the files. Data stores the file with metadata as the information about the information of the files in clusters. It records about the files as the size of stored files or directories and which block is located on which location. Name node receives regularly the heartbeat from the data nodes to know that which is connected with it, whether it is alive or in running state.

The main function of data node is to read the data from the file and to write a new data in the file as per client request. It also useful to delete the file. To quote the status of data node whether it is alive and woks properly then sends the heartbeat to the name node.

HDFS provides reliability and scalability of the data. Core java language is used to write the code of HDFS.[11]

4.4 MapReduce: Google made in 2004 MapReduce framework. It was built so that large data set or huge data may break clusters to process parallel. This process uses divide and conquer method. This processing can be in the form database or file system. It is open source framework which exits on the top of hadoop. So many programming languages are used to write the code of MapReduce eg. core java, pig and hive etc. Pig language or core java is used as frontend language this means the processing is done through these languages. Hive is a data warehouse tool which is used to store the output of mapreduce.

Two functions are available in MapReduce as:

1. **Map**: The task of Map function is to convert the huge data sets such as data in the form of petabytes to another data sets where every element is broken into rows. Map function stores output data into temporary storage.
2. **Reduce**: This function acquires the output from the map function as its input and the taken input combines into very small parts. Reducer shares same key so that output may be sent to the same reducer simultaneously.

Suppose we have a vast data in the form of petabyte, we shall use mapreduce for parallel process of this data. MapReduce passes through various phases to process the data. First phase is called input phase. This phase has a provided data which we have to process. Input phase sends the data to splitting face. This phase splits the input into various data sets or blocks. Input is divided into equal input size in these blocks. After splitting the data, the splitted data is forward to mapper for mapping. Mapper phase codes the data to know that how many times the values are getting appear. Then output of previous stage moves toward the shuffling part to place the similar data into blocks. Reducer function is used to reduce the size of similar data and remove the duplicate values.

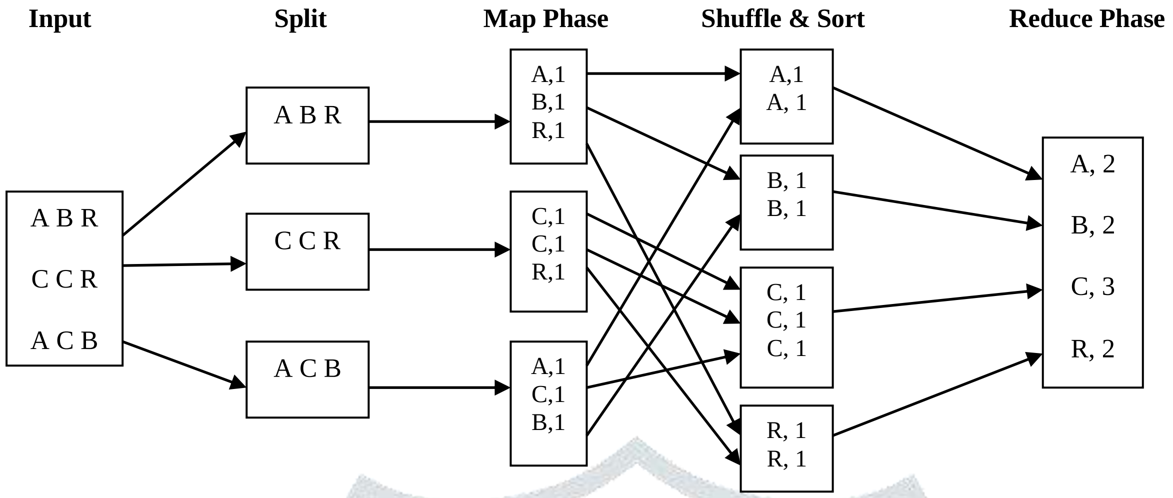

Fig 3. MapReduce [12]

5. Review of literature:

S. Annapoorani & B. Srinivasan et.al. year 2018 Presents that research on big data took place in 1970 and published full fledged in 2008. Apache hadoop permits large amount of data to fragmentize through easy programmer model, so that data may be analyzed. To provide high performance analysis hadoop performs operations on digital platform to analyze the unstructured data and controls the redundant data. The main focus of this paper is that HDFS raises the performance of data with the help of data replication strategy which increases the write throughput also. She mentioned in this paper the applications of mapreduce also like two types of benchmarks (Terasort & Test DFSIO) that provides good installation of hadoop. Terasort classifies a specified amount of non-linear or unplanned generated data. It is a co-tester of mapreduce and HDFS. Test DFSIO is a write/ read test for HDFS. It enhances the performance of data & reduces the execution time of process also. [1]

A. Antony Prakash & A. Aloysius et.al. year 2018 Illustrated that big data existed when traditional database was not able to hold unstructured data. It is a very big challenge for big data to have huge data flexible and fault tolerant. There are so many technologies to handle this data such that hadoop, HBase, Oozie, Sqoop, Avro, Chukwa and Hive etc. Hadoop allots distributed file system but its drawback is that it does not provide encryption on network level. HBase is a non relational database tool which is implemented in java language. Oozie is a web application which gathers data from various sources. Sqoop is an interface which assigns platform to convert unstructured data into relational database and hadoop. HPCC is also another tool which is used to solve the complex problems. Data model is selected according to the requirements of end user. Author has illustrated the challenges, advantages & disadvantages of big data and quoted the handling technologies of big data which improves the performance of system also. [2]

Akshay M. More, Pappu U. Rathod, Rohit H. Patil & Darshan R. Sarode et.al. year 2018 They stated that stock market has a large amount of data which works in terabyte and petabyte. Through Data mining techniques complex problems are solved and collect the meaningful data from database. So many investors are present in stock market who desire to invest the many in share purchasing. The main motive of this paper is to prepare such system who may be able to find out the stock information & stock chart on the internet which may help the users to detect the good investment strategies. Naïve bayes algorithm has been explained in this paper which uses prior knowledge of data to predict the future. The authors created NLP based modules to predict the future trends that compare the data of previous years and give result. Investors can have a right decision with the help of this system that where he/she should invest their money. [3]

Sahatiya Prashant et.al. year 2018 Projects that the use of social media has increased to a lot of extent in last years. Approximately 900 social sites are working on net are present in market now a days. The main motive of this paper is to study the techniques and tools which analyze the large amount of data of social media and various techniques for retrieving data, selecting proper tools which is very useful in business strategy. The data of social media can be analyzed via applied mathematics eg. impression post, responses, click-through rates for URLs embeds etc. Social media supply unstructured data in very huge data. For any business, social media is very important in the digital age. Big organizations use analytical technique to derive new opportunities by using given data. Big data is analyzed through data mining, text mining, predictive analytics technologies. On social media the interest of customer is examined by likes & shares your post receives, replies & comments, click your links etc. It is very big challenge for big data to keep secret the privacy of peoples on social media. [4]

Mashooque A. Memon, Safeeullah Soomro, Awais K. Jumani and Muneer A. Kartio et.al. year 2017 stated that Big data is such a massive data that is difficult to manage in a traditional database, that is why hadoop is used. Big data is handled by several techniques and tools. Big data analytics is a policy in which information about the several attributes about information are analyzed. Firms and companies apply the analytic process eg. to increase the business performance, new income opening, retail analytics and market optimization etc. Data analytics can be categorized into three parts. 1)Descriptive analytics is a simple class of investigation in which past performance is analyzed to proceed towards to future goals. 2) In Predictive analytics the data is converted into valuable data. Predictive analytics uses valuable data to finalize the future of any event. 3) Prescriptive analytics is used to take decision about the business. In this paper the applications of big data are described such as, in banking the big data secures the privacy of individual data and provides security also. In agriculture, big data analysis the plants and observes the changes in them. It provides the information to keep secure the plants from any ailment and suggests about the appropriate soil for the good cultivation of plants. In finance, big data is used to identify the fraudsters so that they may not claim illegally. Data is increasing rapidly due to web associations and phones as they produce a big data. In third-eye application the behavior of customer is analyzed and predict the future marketing of the product.[5]

Simranjot Kaur, Er. Sikander Singh Cheema et.al year 2017 Quoted that entire world is upset from the changing of the weather. Here are so many algorithms to reduce the side effects of the same so that weather can be accessed with the help of given data. To analyze the result of those techniques the data mining techniques are applied and meaningful data is retrieved from given data. It is also a predictive model. Data mining tools are used in this paper which manage the weather related data. As in agriculture field farmers have to face a lot of problems regarding unpredicted weather with it crops are damaged. Weather forecasting is dependent on the molecules which are present in the air like carbon dioxide, nitrogen dioxide and ozone etc. On this collected data incremental K means clustering algorithm is used in which new data values are examined that what are the main causes of changing the values. She introduced an algorithm in which data is collected after every hour and saved in original database. The previous data is converted in structured data by using R tool after every two hours and stored in “Structural air pollution database” (SAPD). SAPD is divided into four sub databases first region is December, January, February, secondly March, April, thirdly May, June and fourth August, September, October & November. The result of three years can be predicted by using priority based algorithm.[6]

Samiddha Mukherjee & Ravi Shaw et.al year 2016 Presents the introduction of big data, 3 V’s of big data. There are many applications of big data. In data visualization, predicting after analyzing the given data that which customer is fraud. Data scientists are needed to visualize the data to achieve the goals. Google, Facebook, ebay and Wal mart companies use this application. In integration application, multi organization are engaged to develop decision making system, which may be helpful in the growth of business. Developing organizations are growing with the growing rate of social media and mobile technology. In this way the organizations are looking for ways to increase the data so that they can increase their income. In food industry, the quality of the product is described to the customer and better product is recommended so that market strategy can be developed for better customer experience. In telecom industry, to improve the customer’s service and their fulfillment big data and machine learning techniques are implemented. Call detail records, web logs and email to the social media are accessed by telecom operators. The author has described the challenges and problems of big data in this paper such as security and encryption are big challenge for big data till up now. So many fields of big data are left where a lot of work is to be done.[7]

M. Dhavapriya & N. Yasodha et.al. year 2016 Introduced we are living on demand world where data is produced by an organization, individually and machine at very high velocity. Homogeneous and heterogeneous are two types of data. It is very needful to process so many quantities of data which never have been expanded earlier. Big data is increasing at very high speed in the form of PB & TB and a lot of data is being suppressed. To unlock the suppressed data computational tools are required. Data warehouse and traditional database are very small so that they may manage big data, that is why the data is stored in the distributed files. It is a challenge for big data which may keep the data flexible, scalable and fault tolerant. In DBMS join, indexing and graphing techniques are used to classify of data and these techniques are acquired by mapreduce.[8]

Sabitha M.S, S.Vijayalakshmi, R.M.Rathikaa Sre et.al year 2015 Express that so many changes are seen in Cloud Computing, Big Data & Internet of Things (IoT) in previous years. Forth coming years social media, IoT, medical science will produce a vast data. The main purpose of this paper is to highlight the big data tools, techniques for storage, challenges and applications. To convert raw data to meaningful data is a challenge for big data. Here, components of hadoop are described also. The applications of big data like in healthcare, to provide the better treatment to the patient by using available information. In automation, to store and analyze the sensor data which is being produced by IoT. Manufacturing industries and IoT are interrelated because automated machines work in those companies. In defence, data of satellite, aircraft & messages from various sources are stored through big data tools. [9]

Harshawardhan S. Bhosale & Devendra P. Gadekar et.al. (2014) Describes that the big data is a advance technology which is used to capture, manage and store the large amount of data at very high speed. So many resources produce big data. It is very tough to manage this data and hadoop is used to manage & store the data in structured form**.** Big data has three characteristics like volume, velocity and variety. Various problems occur in big data as privacy of the data. It is very hard to keep secret this data. Secondly it is very unsuitable to process the big data in short period. They describe about the challenges of big data also which include visualization, lack of structure and error handling. Big data exists in TB,PB which is stored in HDFS. HDFS fragments the incoming files into clusters then after generating 3 copies of every file and stores on distinct servers and for multiprocessing uses mapreduce. It is similar to ETL processing of traditional database.[10]

6. Conclusion: In this paper we have discussed about the big data technology and their tools which are used to manage the data. Big data has many applications and characteristics which are mentioned above said. To provide the privacy of data the work is still going on in big data. Big data analytics is a normal topic in these days. The future problems can be overcome by analysis process and it is very useful to predict the forthcoming problems. HDFS is also a storage framework which is used for replication. MapReduse and HDFS are main components of hadoop. Big data is a very broad technology, but a lot of work is to be done in many areas.

REFERENCES

[1] Improving Performance of Data in Hadoop Clusters using Dynamic Data Replica Placement: A Survey, "S. Annapoorani , B. Srinivasan", Feb 2018.

[2] Architecture Design for Hadoop No-SQL and Hive, "A.Antony Prakash, A. Aloysius", Feb 2018.

[3] Stock Market Prediction System using Hadoop, "Akshay M. More, Pappu U.Rathod, Rohit H. Patil & Darshan R.Sarode", March 2018.

[4] Big Data Analytics on Social Media Data: A Literature Review, "Prashant Sahatiya", Feb 2018.

[5] Big Data Analytics and Its Applications, "Mashooque A. Memon, Safeeullah Soomro, Awais K. Jumani and Muneer A. Kartio", October 2017.

[6] Big Data And Analysis Of Weather Forecasting System, "Simranjot Kaur, Sikander Cheema", August 2017.

[7] Big Data – Concepts, Applications, Challenges and Future Scope, "Samiddha Mukherjee, Ravi Shaw", Feb 2016.

[8] Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table, " M. Dhavapriya, N. Yasodha", Feb 2016.

[9] Big Data – Literature Survey, "Sabitha M.S, S.Vijayalakshmi, R.M.Rathikaa Sre, December 2015.

[10] A Review Paper on Big Data and Hadoop, "Harshawardhan S. Bhosale, Devendra P. Gadekar", October 2014.

[11] A Review Paper on Big Data and Hadoop, "Manpreet Kaur, Harpreet Kaur", May 2018.

[12] Big Data and Hadoop book, V.K.Jain.

[13] Big Data Analytics, www.tutorialpoint.com

[14]Big Data and Hadoop : A Review Paper, "Rahul Beakta", 2015.

[15] Internet of Things: Vision, Challenges and Future Scope, "Balwinder Kaur, Vijay Dhir", May 2017.

[16] Big Data Analytics: A Literature Review Paper, "Nada Elgendy and Ahmed Elragal", ICDM, 2014.

[17] Review Paper on Big Data:Challenges and Applications,"Priya Parhate, Gaurav Ghogle, Jyoti Bhange, Ashwini Ingle", Jan 2017.

[18] Grid Job Scheduling - A Detailed Study, "Rattan k datta Vijay Dhir", 2013.

[19] TBSD:A Defend Against Sybil Attack in Wireless Sensor Networks, "Jatinder Singh Bal", 2016.

[20] Research Paper on Big Data and Hadoop, "Iqbaldeep Kaur, Navneet Kaur, Amandeep Ummat, Jaspreet Kaur, Navjot Kaur", December 2016.

[21] Big Data and Hadoop - A Study In Security Perspectives, "B. Saraladevia, N. Pazhanirajaa, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvanc", 2016.

[22] Exploring Cluster Analysis, "Mini Singh Ahuja, Jatinder Singh Bal", 2014.

[23] Big Data: A Review, "Seref Sagiroglu and Duygu Sinanc", 2013.

[24] The Rise of "Big Data" on Cloud Computing: Review and Open Research Issues, "Ibrahim Abaker Targio Hashem , Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, Samee Ullah Khan", 2015.

[25] Issues, Challenges, And Solutions: Big Data Mining, "Jaseena K.U. and Julie M. David", 2014.

[26] A Review on Big Data and Methodology, "Shilpa, Manjeet Kaur", LPU, Phagwara, India.

[27] Statically Analysis on Big Data Using Hadoop, "Jyoti Kumari, Mr. Surender", June 2017.