Hadoop and Big Data Framework: A Technological Comparison of Various Techniques and Tools

Manika Manwal, Amit Gupta, Sonali Gupta, Shiv Ashish Dhondiyal

Abstract: ***BIG DATA is the impactful terms which we hear nowadays but the question arise what it is? So BIG DATA is considered as the data which is rapidly generating huge amount of data but the question arises from where does this colossal amount of data is being generated? The answer is that there is not only one source of data generation but multiple sources are there of colossal data generation like social media e.ginstagram, facebook etc. Big data is featured with three V's and big data can be classified into data source, content format, data stores,data staging and Data processing. This paper specifies the number of technologies which can be used in Big Data Analysis and discussion liesaround the Hadoop, itscharacteristics, and the technologiesused by Hadoop. This study specifies the comparison of all these techniques and helps the researchers to choose better techniques that can be used to data analysis***

Keywords: Hadoop, Big Data, Map Reduce, PIG, YARN, HBase Sqoop, HDFS

I. INTRODUCTION

The word"Big Data" when see in the sights ofliterature we explored that it has been described in numerousmeaningslinked to it. Out of which few are discussed and mentioned beneath:

The term Big Data defines data which ismassive, diverse,thatis been generated promptly, so this data is bagged, put in storage and then it is circulated. After its circulation, the data is bring about for analysis so that it can be of beneficial and useful[1].

Big Data is such a massiveand diverse amount of data which is normally away from the competency of any technologies to stock, control and process it competently [2].

After studying and understanding different types of literature big data can be seen as the set of methods plus technologies which has been brought together to form the unwind diverse high capacity complex and diverse type of hidden valuedatasets.[3]

Technologies that's being used in Big Data are considered as the new age technologies and latest architectures which areplanned to mined value from multi-variate bulk datasets competently by giving the high speed seizing, determining andanalyzingmeasures[4].

According to the study done big data can be stated as the technique and technology thatis effective on the colossal, diverse and data which is promptly generated, which is then processed and converted into theeffectiveinformation.

Revised Manuscript Received on March 20, 2019.

ManikaManwal, Department of CSE, GraphicEraHill University, Dehradun, India. E-mail: manikamanwal17@gmail.com

Amit Gupta, Department of CSE, Graphic EraHill University, Dehradun, India. E-mail: amitgupta7920@gmail.com

Sonali Gupta, Department of CSE, Graphic EraHill University, Dehradun, India. E-mail: m.sonaligupta15@gmail.com

Shiv ashish Dhondiyal,. Department of CSE, Graphic Era Deemed to be University, Dehradun, India.

I.1 Big Data Features

As we wentthrough the numeroustypes of research papers and literature, the greatest and most popular characteristic of big data which were talked about were the 5 V's out of which the 3 V were communal to all, basically the 3 V stands for velocity, variety, and volume other than these rest 2 V stands for value and veracitywhich are not defined in most of the studies. Figure 1 below shows these fiveBig Data characteristics.

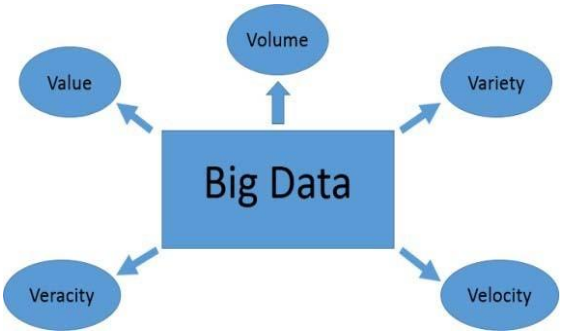

Figure 1. Features of Big Data.

The five characteristics which are mentioned in figure1 are defined below [4][5][6][7].

Volume: Can be described on the basis of examining the current scenarios. The world is turning into a digital worldas large percentage of data is being produced from multiple different sources like social networking sites, financialtrades etc.

Variety: as there are numerous organization which produces Innumerable data which when gathered byseveral means (external or internal). Due to whichinnumerableform of data sets is generated.

Velocity: In terms of big data velocity can be understood as the rapidly and frequently generating data per second. The numerous data generated is directly relative to time e.g. in banking several transaction are done in a minute.

Veracity: Itcompletely talks about the accurateness of data. It have to be firmlyexpected from the confirmeddifferent resources. It specifies that the only authentic person had better theright to use permission.

Value: This is defined aswell-informedand familiar end results which is formed after all processing steps has taken place.

I.2 Classifications of the Big Data

Big data various characteristics can know better by distributingthe data into classes.So,different classes are as follows:[1][3].

I.2.1 Data Source: is the immensely generated data which is gathered from diverse sources, like socialnetworking sites, web, machineries, sensors, financial transactions, and many more. All of these mentioned sources, generate data at distressing rate for e.g.Every secondalmost 695,000+ status are updated on Facebook, every single minute 100,000+ tweets are been posted on Twitter, various number of new cell phone user are there, 168,000,000 above mails are created and send. Thus data chunks in millions are generated by the peoples every single second using different platforms. Therefore this heavy and rapidly produced data is beenput in storage in different formats in various Datasets or Data Sources form.

I.2.2 Content Format:as diversecategory datawhich is been generatedcommencinginnumerable sources therefore their format structure will also be different from every former, in addition these various content formats are categorized into structured, unstructured and semi-structure. The structured form of data is one which comprises of the schema or structure that is primarily in structured format that is within rows and columns. Unstructured data is one and only which do not comprises of any precise format alike video files, log files, audio files and other such files plus images also falls under this category. Semi-Structured form of data is solitaryin which schema of the data is not definite the files which are considered are like .tsv, jason, e-mail, .xml, .csv, etc.

I.2.3 Data Stores: The data stores can be defined as the forms of stores in which data produced byseveral data sources is gatheredand thenstored on the storage. So this information is kept in severaltypes like column-oriented, document-oriented, key value andgraph based [1].

I.2.4 Data Staging: Data Staging can be defined as the process which comprises of the three differentforemost steps like Cleaning of Data, Normalization & Transformation of Data. In the data cleaning process, the stored data is first fetched from data storage house after that the ambiguity and other such unwanted data is removed by using the preprocessing techniques in cleansing of the data done. The cleansing process is then followed by the normalizationprocess in which the same data which has been cleaned is reorganized in the database to avoid the inconsistency and redundancy in database. In the end least but not the last comes the transformed process in which the initial data format at the time of fetching is converted into human comprehensible format or the desired format which is required for certain analytical process.

I.2.5 Processing of Data: Under data processing the enormouslyhuge and resourcefulrecords are been processed with the help of systems which has the capability of processing Real-time data and Batch data processing. According to the concept of batch processing system, various data is gathered round, than some processing is done to generate output (used by Hadoop) however in this distinct programs and algorithm are compulsory for input, process, and output like takean example of billing system. On the other hand, in context to the Real-TimeSystem, no formalrequirement for forming the different programs, the continuity is maintained in processing steps like taking the input, processing, and generating the output is done. The datamust be sort out in least possible time considering.

II. HADOOP

Hadoop stands for the Highly Archived Distributed Object Oriented Programming. It istrulyoriginated in the year 2005 which was basically the outcome of the work of Doug Cutting and MikeCafarella run-through. Dough Cutting Hadoop baptized the name Hadoop this name strike to him as it was his chap's toy elephant name. At first, Hadoop wasmodeled forassistanceof designed distributions for Nutch, that was the search engine speculation. This was the open-source software which provide the dependencyadd on to which it was accessible and also make available distributed-computing on the clusters of servers which arecost-effective [9]. Thisgenerated software is capable of handlingmassivequantity of diverse data from spatialsources like videos, sound files, pictures, files, folders, data generated from sensor,Unstructured Data, Formatted Query& conversations and any other typesof diverseformats [10]. Hadoop is composed of several kinds ofmodules likeHive, Flume, Pig, Sqoop, Oozie, Zookeeper, HBase etc.Conceptionally, Hadoop canalso bedefined as a platform which provideslocation awareness, source code, Work scheduling etc. As we know that Hadoop follows the principal of master-slave architecture therefore, master node in it encompasses of Name Node, Data Node, Task Tracker and Job Tracker while the slave node is the Data Node and Task tracker. Slave Node is the onewhich is in charge for computing of data and is stated as the worker node. The responsibility of the Job Tracker is to deal with Scheduling of various jobs. Hadoop consists of two parts namely HDFSand Map Reduce [11].The figure give below displays the Hadoop architecture in the most simpler form.

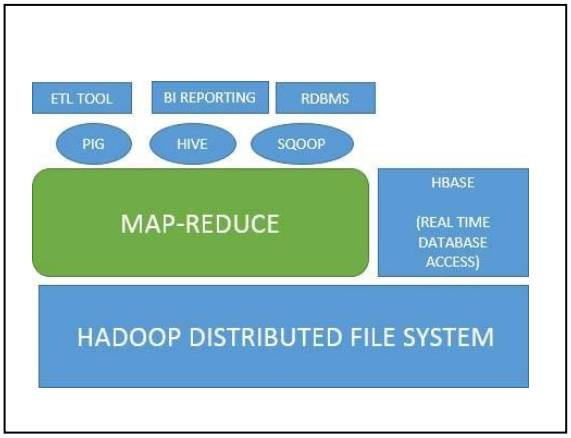

Figure 2: Hadoop Architecture

➢ **Hadoop Components:**

2.1. HDFS (Hadoop Distributed File System): The HDFS is said to be as new or latest file system whichis principally work as framework that is developed using Java programming language. HDFS is composed of two types of nodes namely Name node collection of various data nodes.HDFS executes on the principal of thearchitecture of the master-slave scenario in which the primary node behaves like a master and stores the complete information, whereas the data node works as a slave node which is controlled by name node.

The first part pf HDFS that is the Name Node,contains various information likethe location of file,the meta data, free space, attributes, active and passive node, Job Tracker, task Tracker and the data stored by them, that is data replication etc. It possesses records of entire metadata, various attributes, and thefile locations along with data block into anyof the available Data node [13]. The attributesmainly focus on the characteristicssuch as file permissions, variation, access time of a file etc. All of these files and various folders or directories are been arranged in a hierarchical fashion. The name node provides the functionality of mapping the hierarchy of the name node and block of file in the data node. If any client request to read the data from a file using HDFS, then it is the responsibility of master node (i.e. name node) to collect the required information about the various location of the data blocks which contains the files that are required by the client. Every time the master node stores the copyand periodically maintains logs of a system. Along with this it also keeps them updated in acquaintance that wherever the data blocks, various location where the file is been copiedand details related to available free blocks that are present in the system. The Data node is also said to be slave node which hold back data in additioncontains task tracker whose primary task is said to be tracking the active data node, it also tracks the various jobs upcoming from the masternode [14]. Data node holds on record of each single block it comprises of. Hourly blocks activity log is send by data node to master node or name node. Subsequently master node have tobe acquainted with updated information about the data node in which the blockcopied data is kept back in specific cluster. Whena normal operation proceeds in HDFS than data node directs it's heartbeat to name node in every ten seconds to insist on its obtainability and workingexactness. If in any situation the master node that is the name node did not receive any notification in the form of heartbeat regularly in every 10 minutes by the dta node, then, the name node starts creating replica of the data on other worker nodes. It takes delivery of thousands of heartbeats every single second however it make sure other name nodes did not get affected [13].

2.2 MR (MapReduce): It is concept which is designed to execute different processing tasks in the distributed fashion on the huge data sets which are collected in computer via various data sources. MapReduce initially isused by Google, which is capable of managing large datasets and processing tasks in parallel manner.It works very efficiently on vast variety of hardware in dependable and fault tolerant manner [15]. Map-reduce technique or algorithm is mainly used by Hadoop Map Reduce Engine which works for entire cluster and assigns the work to all the components of the cluster. These two layers namely HDFS and Map Reduce are perfectly integrated in distinctive Hadoop cluster. Map Reduce uses master slave architecture thus the worker or job tracker in the map reduce style works as the task tracker and it mainly transfer the same type of work to the slave node of the task tracker. Hence, the Master node includesthe task tracker and job tracker node where the data node and name node in the HDFS layer. In the similar fashion, the slave node is composed of data node at HDFS level.Therefore, Map Reduce incorporate job node and the task tracker node [13].

For every master single Job Tracker regulates the scheduling action of the jobs constituent's tasks on the slave node. It also witnesses the progression of the slave node and executes the failed tasks again. In disparity, the particular task tracker for each slave executes or runs the instructions as specified by the master node. The elementary functionalities of Map Reduce are divided into the map phase and the reduce phase.[13].

MapReduce is bifurcated in Map and Reduce:

2.2.1 Map step: WithinMap phase the massive process or task is received by master node as inputsubsequently the problem is separated into sub-problems and then these sub-problems are allotted to worker nodes. The multi-level tree-like hierarchal structure will be formed as worker node feasibly divide the sub-problems. Individual Data node works on each sub-problem. The alteration of a mapper is finishedbyaltering the row of input data key andits value to the output key:

a. $M_p(K_1,V_1)$ -> L_s $<K_2,V_2>$thus, Mpstances for map, Lsstances for list, Vstances for Value and K stands for Key. Therefore, for a list of input containing zero or additional pairs of (K, V)is produced**:**

i. The keys i.e. Input key as well as the output key both can differ.

ii. The output multiple entries can be established with the help of single key.

2.2.2. Reduce Step: In this phase the outputs or the answers ofevery sub-problems are bought together. This step is purely done by master node, after thisthe various outputs or the results are assimilated in previously defined fashion. Reducetransformproceedsall the obtained values and generate a fresh reduced outputlist.

i. Reduce (K_2, ls$<V_2>$) ->ls$<V_3>$.

ii. MAP: Map() is practiced by every worker node to local data to write the results to secondary storage area. The name node implements just single input copy among redundant data input copies.

iii. SHUFFLE: In this the data which is dependent on the value of output key Worker nodes reallocate data,in such a way that entire data fitting to one key is putbeneath the similarworkernode.

iv. The worker node simultaneouslyprocessevery singlecluster of output data,for eachkey.

YARN: initially the map reduce is used in Hadoop but as it takes more time in processing the data so Hadoop version 2.0 is released in which the concept of YARN is used.YARN stands for "yet another resource negotiator".The essential indication of Hadoop YARN is to deliver a generalized platform formanaging resources and supporting varied programming models, including MapReduce and MPI[21].The three basic constituents those formulate the YARN architecture are ResourceManager, NodeManager andApplicationMaster. The ResourceManager is a supreme service, that has the capability to manage various resource requests for different data and application applicationswithin the specific cluster.

Itplays role of main node in the architecture of YARN. The ResourceManager [22] againsegmented into 6constituents based upon the services presented.

YARN Architecture processing sequence is as given below.

1. The submission of a client program is done in the cluster.
2. A specified container is located by the RM to initiate the AM
3. Negotiation is done by the ApplicationMaster along with the Resource Manager foreach resource containers
4. Once theallocations of the containerisfinalized and done successfully then the ApplicationMastersend a request to Node Manager for launching the container. In YARN, the code is processed in thecontainer, after which the ApplicationMaster is retorted with theimplementation status. During singlerequestimplementation, the direct communication with ApplicationMaster is done to get thestatus, progress of jobs, any updates etc. After the completion of job, theApplicationMaster leave the container and shuts down so the container are qualified fortaking request from the resource queue.

2.3 PIG: Initially PIG was developed by Yahoo! which basically uses Hadoop for analysis purpose that works on huge datasets. It does not waste its time in writing big mapper and reducer programs. PIG provides the capability to be used in any environment thus it is being designed in such a manner so that it can be used to handle any kind of data.PIG is composed of two most important parts, namely Pig Latin (language) and runtime environment. The second part that is the runtime environment is important part of PIG as it provides a mechanism to implement the programs created in pig latin language [16]. Various steps to be followed in pig programming language are:

1. Data which is required for the Pig program needs to be uploaded from HDFS
2. Data is then transformed using certain conventional transformation processes.
3. At last, the requested data is being DUMP to display or result is STORE in a file.

2.3.1 LOAD: The most important property or entity of the Hadoop are those objects, on which certain tasks has to be performed and kept in HDFS. Therefore, the data on which the pig program is to be executed is firstly loaded from HDFS. The command to do the so is LOAD ‘data file’. In place of LAOD function, the USING function can be used in such situation where directory files are not in pig accessing format.[16]

2.3.2 TRANSFORM: It helps in performing various activities on the data which helps in interpreting or manipulating activity. Various process that can be used are FILTER to remove undesirable data, JOIN, to to concatenate or merge two or more data files, ORDER, GROUP etc.[13]

2.3.3 DUMP and STORE: The usage of the DUMP and STORE commands are to generate the results of any pig program. If these two commands are not used then the results could not be generated. The first command, DUMP is used to direct the output to be get displayed in addition to fix the errors. While the STORE command can be used to store the results on another file for future reference.

2.4 HIVE: HIVE was introduced by very popular company, namely, Facebook but afterwards it was brought by Apache. It has been used by most of the companies like Amazon, Netflix etc [13]. Basically HIVE is considered as a infrastructure for Data Warehouse, which provides the facilities like data summary, query, and analysis of data. Apache Hive facilities the storage of massive and huge datasets in HDFS. It also provides some specifications that is used for the analysis of data in temper file systems, that is, in Amazon S3 [17]. HIVE is composed of HiveQL language which is technically same as SQL language. Along with the creation of schema it also provides the mechanism for the transformation of various queries into the format of map and reduce. For the speedy processing of the queries it provides the mechanism of indexing. The meta data of hive is been deposited in Apache Derby database and also in client server related databases alike MYSQL which can be used optionally. [17]. Hive provides file formats in four different ways namely, SEQUENCEFILE, ORC, TEXTFILE and RCFILE. It provides bitmap index which enhance the performance in form of speed of the execution of various queries. It also provides the mechanism of supplementary indexes. Further, it is composed of various storage natures like ORC, RCFILE etc. in most of the RDBMS software, the time relaxation is used during the process of query checking as a popular mechanism for storage of metadata. The algorithms and techniques which are same as snappy, gzipetc are used to work on various data which is stored in the storage area of Hadoop. The various or different queries of HiveQL are circuitouslytransmitted into MapReduce jobs. Based on Structured Query Language, HiveQL do notrigorouslyfollows the whole SQL-92 standard. Furthermore, HiveQL comprises of extensions which SQL do not comprise off, it can inset multi-tables and overcan create table as select. However, it onlyoffersmeek provision castoff in indexes [18]. Hive provide a shell to run it quires which entirely a command line interface. Hadoop Cluster executes Map Reduce job which are fragments ofHQL statements.

2.5 HBASE: It is pinnacle for any project on Hadoop. It is officially work on massive data which aimed at assurances of normal language search. In the month November 2010 HBase was used by Facebook to run through its fresh messaging platform. HDFS provides platform for HBase to run its column-oriented database formats. It performsgreatest for light datasets, and in many use cases that are implemented on Big Data environment. HBase don’t practice simple structured query language (like SQL) commands. HBase at no timeis measured as relational data store. Similar to MapReduce application HBase considerJava language to inscribe an application it supports different application writing platform like Avro, Thrift and REST [13].

The HBase configuration comprises of tables sets. Eachsolitary table includes rows and columns like traditional database.

Each table contain the element like Primary Key, it is access by eachHBase tables. The object quality is stated by column in HBase[13][19].Simply, a table is made up of rows and columns. The rows hold on track of indicative logs from server, whereas each row is to be considered as a log record besidesthe column contains different log time stamp or probably name of log producing the resource. HBaseis responsible to authorizestheassemblenumerous attributes into different columnsto store all the elements of entire column family under single. HBase is not at all similar to traditional relational database where rows and columns stored together. Every new column be able to introduce at any point of time this add elasticity toschema whichturn intoused to mutablenecessities of an application. HBase also follows the master-slave architecture same as HDFS and MapReduce, likewise, in HBase the master node is able to manage the cluster and the shares of the table are stored in region servers.[19].HBase is gentletoall the damages of its masternode.

2.6. SQOOP:The import and export of massive amount of data amongst structured data stores and Hadoop is done by Sqoop. This tool is developedby Hadoop.Basically, it is a CLIapplication which is created in Java. It is scheduledmainly for transferring a colossalamount of data. It furthermoreduplicates data quickly from peripheral systems to Hadoop. Sqoop allows data to be exported from the external source of data and keep it in data warehouse in Hadoop. As data is transferred in parallel so the system utilization is at its peak due which Sqoop is considered fast. Sqoop also supplies analysis of data competently. It even differsgreat masses to various outside systems. Mainly,Sqoop is practicedwithin the Hadoop cluster. It accessesto Hadoop core. The incoming data is cut by using the mapper. Sqoop communicate with database store for attaining meta-data from relational database, the meta-data is used forinitiatingthe java class by Sqoop. JDPC API is internally used to form a JAVA class. The compiling of java class is completedby means ofJava compiler and it equates the .jar files. Sqoop yet againinstitutesassociation with storage of database, different .jar files are made in a such amanner which facilitates to determine the various divided columns that is helpful in acceleration of the process. Sqoop is used to procure data from a database. Finally, the is positioned into HDFS by Sqoop[20].

TABLE I Comparison between different technologies

	Feature	Language	Lines of Code	Code efficiency	Development	Abstraction property	Data	Used by
HADOOP	Distributed processing	Compiled	Higher lines of code	Code efficiency is high	Higher development is required	Lower level	Data processing	Real-time application
PIG	Platform for step-by-step data analysis	PIG Latin procedural language	Comparatively less lines of code	Code efficiency is less	Less development efforts	Higher level	Data access	Programmers and researcher
HIVE	Manage large datasets	Structured query type language	Less lines of code	Code efficiency is much less	Low level of development efforts are required	Higher level as compared to pig and map reduce	Data access	To generate daily reports
YARN	Cluster resource management	Batch processing language	Multiple map-reduce APIs	Efficiency is high	Shared operational services across multiple workloads	Higher level	Data distributed scheduling	Real-time streaming
SQOOP	Export data with RDBMS	RDBMS is required	Command-line interface application for code	Efficiency is high in terms of data analysis	Improves query performance in terms of development	None	Data integration	Data analysts
HBASE	Belong to DBMS fields	No SQL database	JAVA APIs are used	No code only storage	Higher maintenance of database	Higher level	Data storage	Storage manager

The above table describes about the differences between the technologies in the terms of their features, efficiency, usage etc.

III. CONCLUSION

The study above puts light on the concepts of Big Data, its, features,techniques and characteristics. This manuscriptdescribes some conceptual things and various technical facts and figures about Hadoop and its architecture. At last it specifies the comparison of these described techniques used in Hadoop so that a researcher can be benefited while selecting the required technology for his or her implementation. The overall comparison table differentiate the various technologies on the various of certain important factors.

REFERENCES

1. Manwal, M., & Gupta, A. (2017, November). Big data and hadoop—A technological survey. In *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)* (pp. 1-6). IEEE.
2. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.
3. Qian, J., Lv, P., Yue, X., Liu, C., & Jing, Z. (2015). Hierarchical attribute reduction algorithms for big data using MapReduce. *Knowledge-Based Systems*, *73*, 18-31.
4. Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.
5. Cadersaib, B. Z., Sta, H. B., &Rahimbux, B. A. G. (2018, October). Making an Interoperability approach between ERP and Big Data context. In *2018 Sixth International Conference on Enterprise Systems (ES)* (pp. 146-153). IEEE.
6. Batista, G. E., Prati, R. C., &Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20-29.
7. M. Beyer, D. Laney, The importance of big data: A definition, ID: G00235055, Retrieved from Gartner database [Online; accessed December 2013], 2012, http://www.gartner.com/id=2057415.
8. **Dhole** Poonam B, GunjalBaisa L, "Survey Paper on Traditional Hadoop and Pipelined MapReduce",InternationalJournal of Computational Engineering Research Vol 03, Issue12
9. Vera-Baquero, A., Palacios, R. C., Stantchev, V., &Molloy, O. (2015). Leveraging big-data for business process analytics. *The Learning Organization*.
10. Chu, C. T., Kim, S. K., Lin, Y. A., Yu, Y., Bradski, G., Olukotun, K., & Ng, A. Y. (2007). Map-reduce for machine learning on multicore. In *Advances in neural information processing systems* (pp. 281-288).
11. Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Zhang, N., ... & Murthy, R. (2010, March). Hive-a petabyte scale data warehouse using hadoop. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)* (pp. 996-1005). IEEE.
12. Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. *Journal of big data*, *2*(1), 8.
13. Suman Arora, Dr. Madhu Goel, "Survey Paper on Scheduling in Hadoop", International Journal of Advance Research in Computer Science and Software Engineering Volume 4, Issue 5,May2014.
14. Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences*, *275*, 314-347.
15. O'Driscoll, A., Daugelaite, J., & Sleator, R. D. (2013). 'Big data', Hadoop and cloud computing in genomics. *Journal of biomedical informatics*, *46*(5), 774-781.

16. Apache Pig. Attainedfromhttp://pig.apache.org.
17. Apache Hive. Attainedfromhttp://hive.apache.org.
18. Zhou, F., Pham, H., Yue, J., Zou, H., & Yu, W. (2015, August). Sfmapreduce: An optimized mapreduce framework for small files. In *2015 IEEE International Conference on Networking, Architecture and Storage (NAS)* (pp. 23-32). IEEE.
19. Apache HBase. Attained fromhttp://hbase.apache.org
20. Bressoud, T. C., & Tang, Q. (2016, September). Results of a model for hadoop yarn mapreduce tasks. In *2016 IEEE International Conference on Cluster Computing (CLUSTER)* (pp. 443-446). IEEE.
21. Verma, C., & Pandey, R. (2016, January). Big Data representation for grade analysis through Hadoop framework. In *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)* (pp. 312-315). IEEE.