ISSN 2347 - 3983
Volume 8. No. 9, September 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter137892020.pdf
https://doi.org/10.30534/ijeter/2020/137892020

Transforming Unstructured data to Structured data using Map Reduce and HBase

Dr. R. Vijaya Kumar Reddy[1], G. Venugopal[2], Girijaswi Rajanala [3], V. Sambasivarao [4], N. Harshavardhan[5], T. Ajay[6]

[1] AssistantProfessor, Department of InformationTechnology, Prasad V Polturi Siddhartha Institute of Technology, Vijayawada., A.P, India. vijayakumarr285@ gmail.com,

[2] AssistantProfessor, Department-of InformationTechnology, Prasad V Polturi Siddhartha Institute of Technology, Vijayawada., A.P, India. .venugopal.gaddam@ gmail.com.

[3] Assistant Systems Engineer, TCS, Kolkata,West Bengal, India. girijaswi.rajanala1210@gmail.com

[4] Assistant Systems Engineer, TCS, Hyderabad, Telangana, India. vsamba16@gmail.com

[5] Assistant Systems Engineer, TCS, Hyderabad, Telangana, India. harshanallapati@gmail.com

[6] Assistant Analyst, Veetechnologies, Hyderabad, Telangana, India. ajaykumarajaykumar33@gmail.com

ABSTRACT

Unstructured data is defined as a pre-defined data model or unorganized in a pre-defined mode. It is classically text-bulk, but it can hold data like dates, numbers, and facts. This outcome in irregularity and ambiguity which make it hard to comprehend with long-established programs when compared to stored data in particular location in data bases. This type of data can be altered into structured format using Hadoop Distributed File System. By using Big Data Analytics technique, a Hadoop ecosystem tools like Map Reduce and HBase, data that is unstructured can be formatted into structured data and then results display in HBase.

Key words : Unstructured data, Hadoop Distributed File System, structured data.

1. INTRODUCTION

Big Data Analytics refers to the methodologies that can be used for convert uncooked data into significant data, which helps in commerce analysis and forms a decision support arrangement for the executive in the society. Volume, Velocity, Variety includes dissimilar kinds of data present like structured semi-structured and un-structured data in big data. Hadoop is the uncomplicated java programming skill that provides framework for giving out distributed data sets across distributed computer groups called Hadoop distributed file system. HDFS is a extremely scalable and distributed file system. It fragmented the file into blocks, which is assigned to the dissimilar nodes in the group of Hadoop framework where the contribution of initial data is processed with the help of Map Reduce programming and then outcome is written again in HDFS as shown in Figure 1.

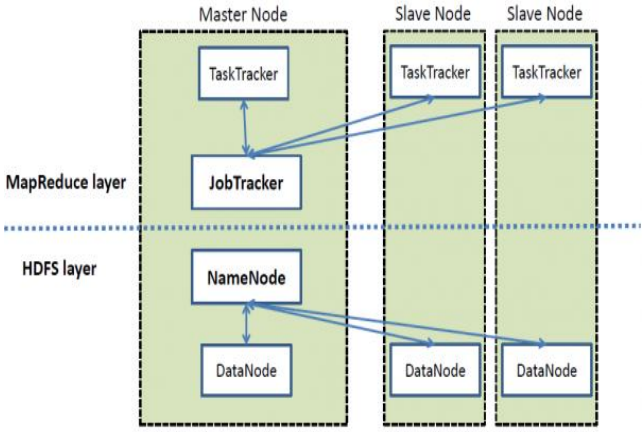

Figure 1: Architecture of Hadoop

It is observed that using the earlier versions and giving out the prearranged data mark-up is the world communal verbal communication format which gives search engines recognize and facilitate users the best outcome who are thorough for connected data. The structured data that is taken back from the unstructured data in past methodologies are giving solutions to more complicated issues with the help of java and RDBMS. By using Hadoop ecosystem tools like Map Reduce and HBase we can convert Unstructured Data to Structured Data and the search of any type of information in less time and to understand meaning in an easily when data in this format and reuse the already converted data for the future purpose.

The paper follows background work in section two. The third section deal with methodology and section four results and analysis. Finally conclusion end of the paper.

2. BACKGROUND WORK

Mapper is a database management system. It is a software tool for help end-users contribute to computer power in a system. Users are able to build up their own applications and process them instantaneously. A mapper function usually divides the input data-sets into independent chunks that are processed by the map jobs in an entirely matching method. The input data sorts out the framework, based on the matching of input data. Typically, the data is stored as an input in a file-system. The setting up jobs framework examines, monitor them and re-executes the aborted jobs.In Hadoop, it is observed that the process of transferring intermediate output from mappers to the reducer is named as shuffling. The keys that are produced by the mapper are mechanically sorted out by MapReduce Framework. It means that before the start of reducer, all intermediate key-value pairs in MapReduce which are produced by mapper get sorted out by key and not by value. In this process, values acknowledged to each reducer are not sorted out. Sorting in Hadoop helps reduce to easily differentiate whenever a new reduced task starts and saves time for the reducer. Reducer starts an innovative reduce method when the next key in the sorted input data is unlike than the past. Every reduce job takes key-value pairs as input that generate key-value pair as outcome.

In the event of generating the outcome, reducer takes the outcome of the Mapper process of each step. The outcome of the reducer is considered the final outcome, that is piled up in HDFS. Hadoop Reducer receives a set of a middle key-value pair that are generated by the Mapper considering it as the input and moves a reducer function on every one of them. In order to get a broad range of processing, the data is aggregated, filtered and combined in a number of ways. First, the intermediate values are processed by reducer for particular key that can be produced by the map function and then gives final the output.

3. METHODOLOGY

Apache Hadoop eco system is a framework that is used for reduce big data organization issues. The Hadoop core affords laying up of un-structured data using the HDFS along with Map Reduce programming replica to evaluate data in parallel method, and then it is stored in the distributed system. This system helps in retrieval of necessary data in a short span of time and more accurate outcomes are obtained. Generally, the typical start-up and shut-down draft requires protected shell to be arranged between nodes and cluster. It comprises OS level Group of data like Map-Reduce Engine and Hadoop Distributed File System(HDFS). HDFS is a file System which is written in Java Hadoop Framework. The progress of Hadoop has a major development in the field of big data. Hadoop holds up the structure of Big Data since it is an equal programming stage [1] [2].

A MapReduce work for the most part separates the data information into self-governing chunks that are processed by the map jobs in a totally parallel way. The structure sifts through the result of the guides, that help the contribution to the decrease occupations. Traditionally both info and yield of the undertaking are presrved in a document framework. The system looks at plan of occupations, screen them and re-executes the bombed positions. It is very common that the process nodes and the capacity nodes are the equivalent. The Map Reduce system and the HDFS are running on the comparable arrangement of nodes. This arrangement concedes the structure to plan occupations well on the nodes where information is now present, resultant in very far above the ground aggregate bandwidth across the group. The Map Reduce structure includes a solitary master Job Tracker and one slave Task Tracker for every group nodes. The master is the in-control to plan the positions part on the slaves, screen them and re-execute the useless errands. The slaves at that point complete the positions as focussed by the master. The joiner sticks to the joining of the moderate outcomes from the guide occupations and reducer subtask that is utilized for completing total. The final product is put away in HDFS soon after the map and reduce jobs [3].

As it is a big data processing, it facilitates people to run applications on systems that involve thousands of nodes having terabytes of data. Hadoop makes it feasible due to its distributed File System. It also assists people to carry on operation even though there is a failure of node. But, it is observed that a single point failure does not influence the failure of calamitous framework. Hadoop is an open-source Map Reduce execution that is intended for enormous Clusters. It comprises a single master node namely, the JobTracker and many slave nodes namely, the TaskTrackers as shown in Figure 2. As characterized in the article, JobTracker is responsible for parallelizing the activity execution across hubs and accordingly guaranteeing adaptation to internal failure [4][5].

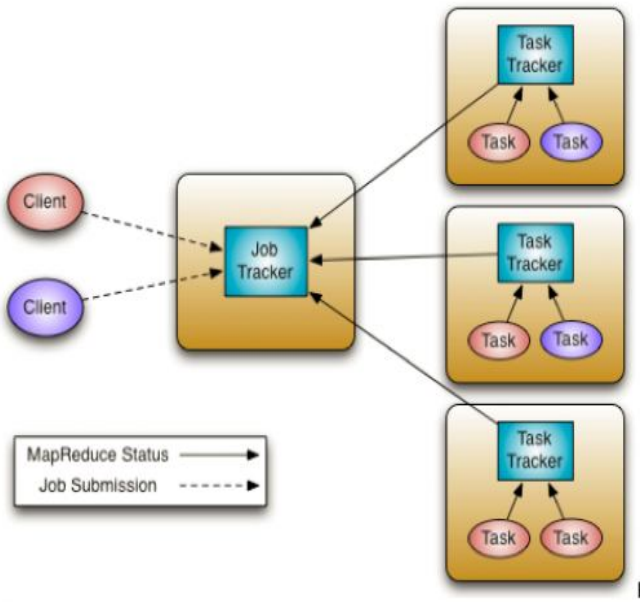

Figure 2: Procedure of Map reduce

4. RESULTS AND ANALYSIS

In this section deal with results and analysis of proposed system. The following Steps involved generating the results. Initially VMware player installation and running vmware player. In nest step Cloudera work station with creating a new java project in eclipse[6][7]. After creating a project in external location to setting jar file libraries and Setting jar files from client. After that creating a new class with Setting packages for that class and writing source code. And then creating jar file along with exporting the jar file. Open the input file in terminal and executing the Map Reduce file operations, output generated after Mapreduce is executed[8][9][10]. That content entering into HBase shell and Listing tables in HBase, finally showing table contents in HBase which are shown from Figure 3 to Figure 14.

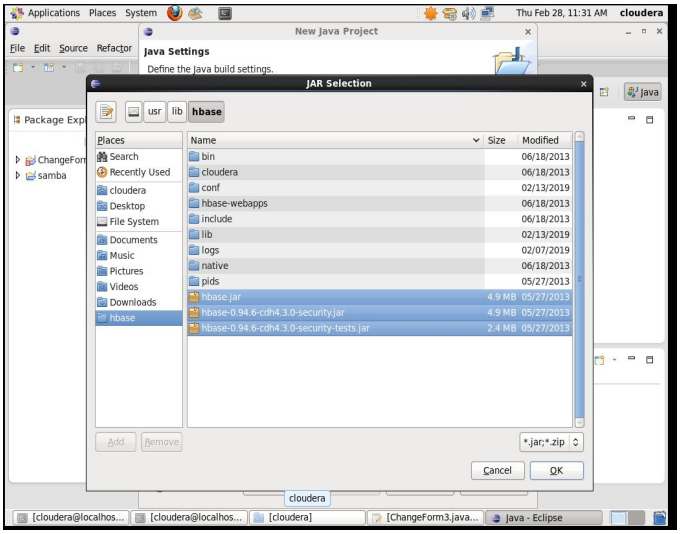

Figure 5: Setting jar files from hbase

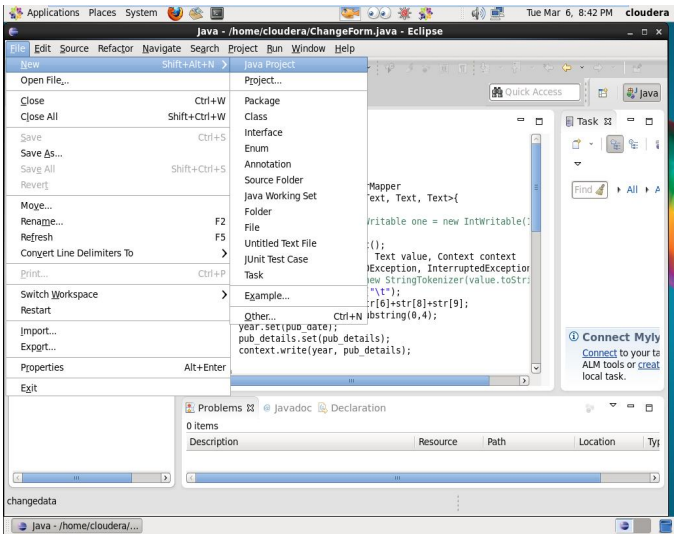

Figure 3: Creating a new java project in eclipse

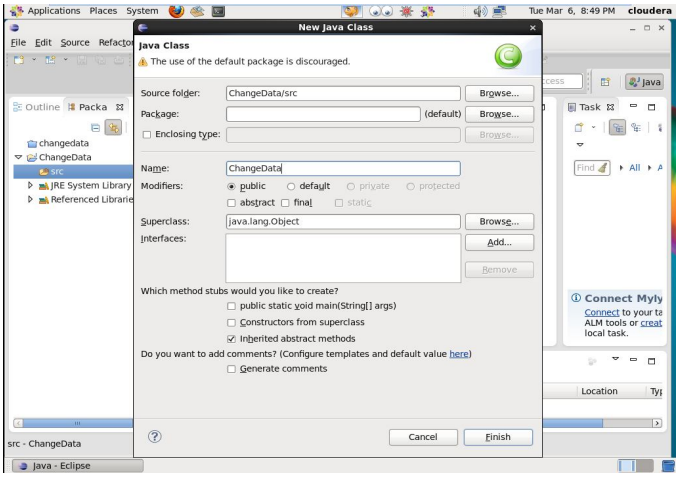

Figure 6: Creating a new class

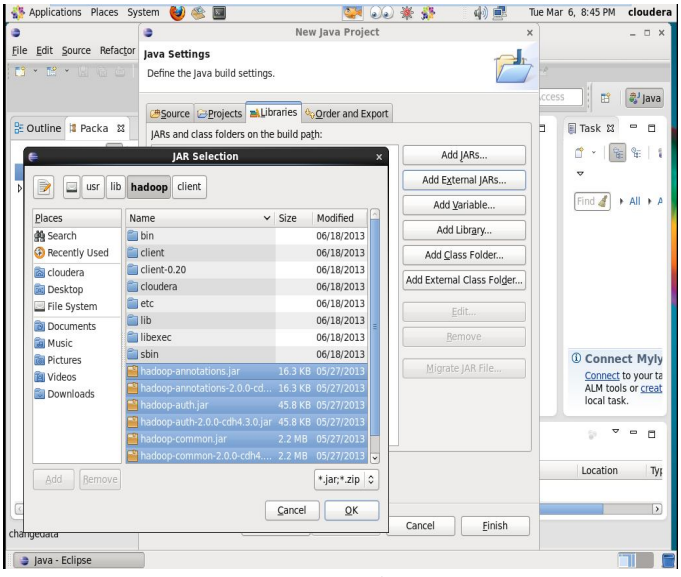

Figure 4: Setting jar file libraries

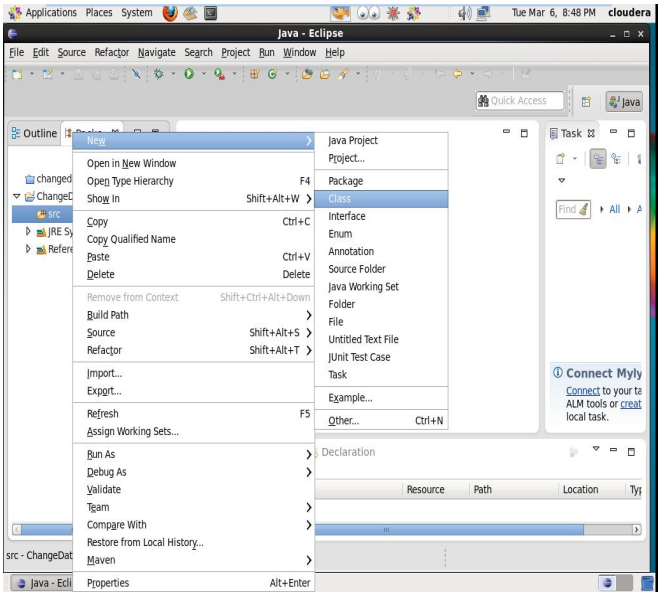

Figure 7: writing source code

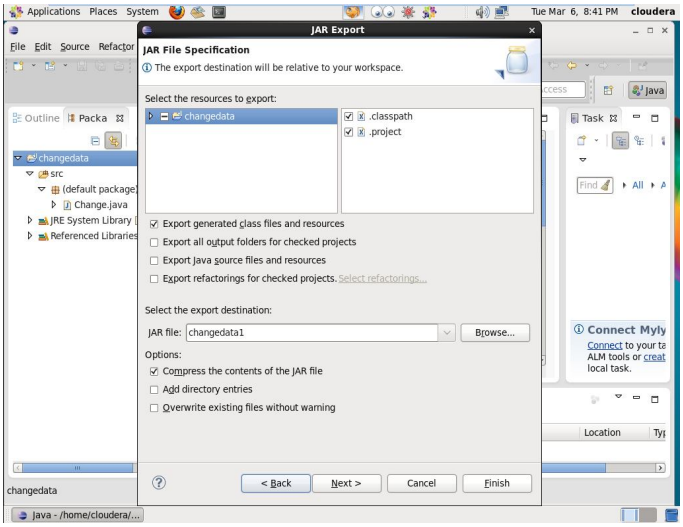

Figure 8: Exporting the jar file.

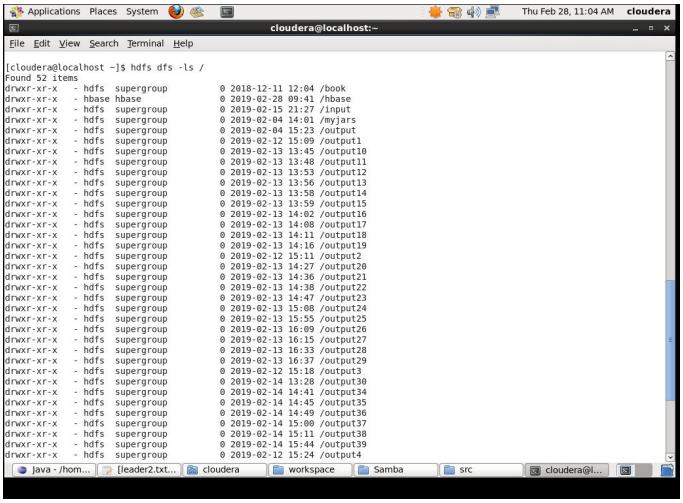

Figure 9: Opening the input file in terminal.

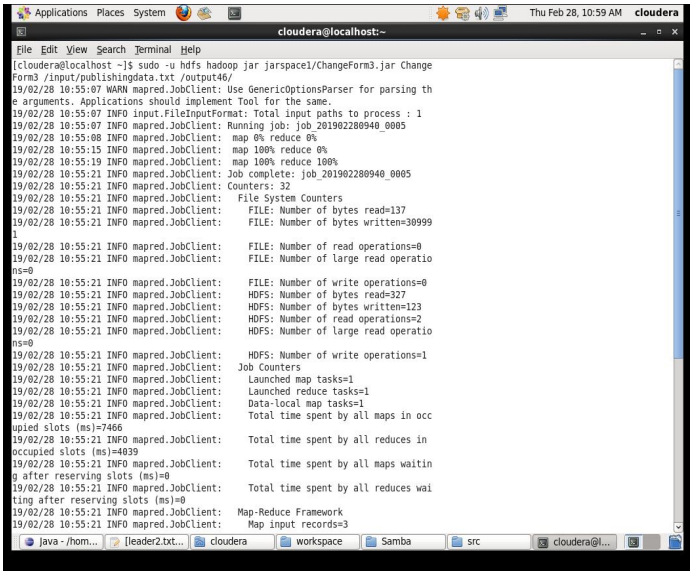

Figure 10: Executing the Map Reduce file

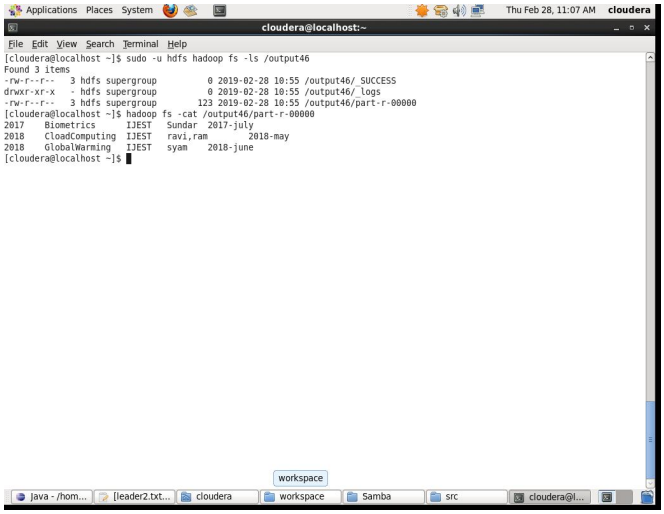

Figure 11: Output after executing Map Reduce file

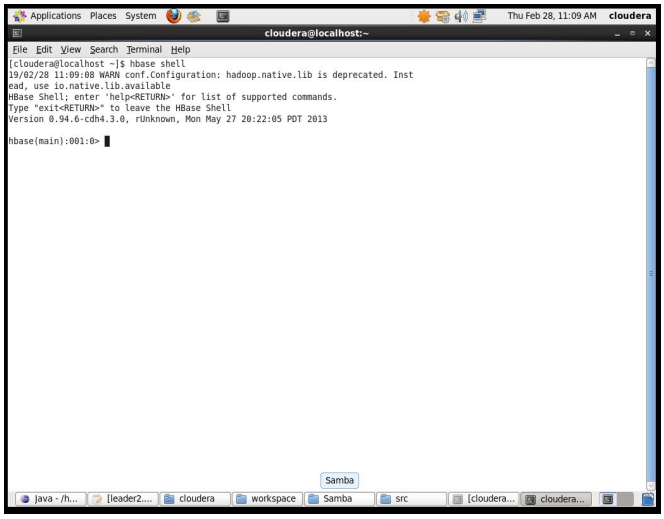

Figure 12: Entering into HBase shell

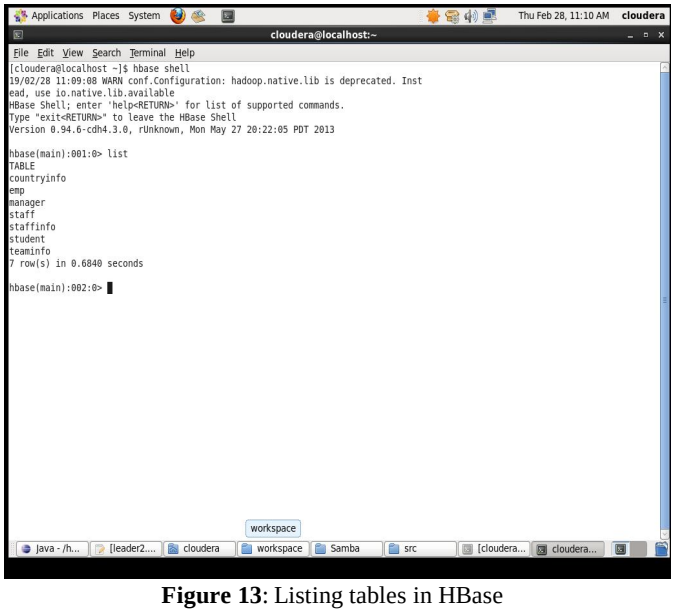

Figure 13: Listing tables in HBase

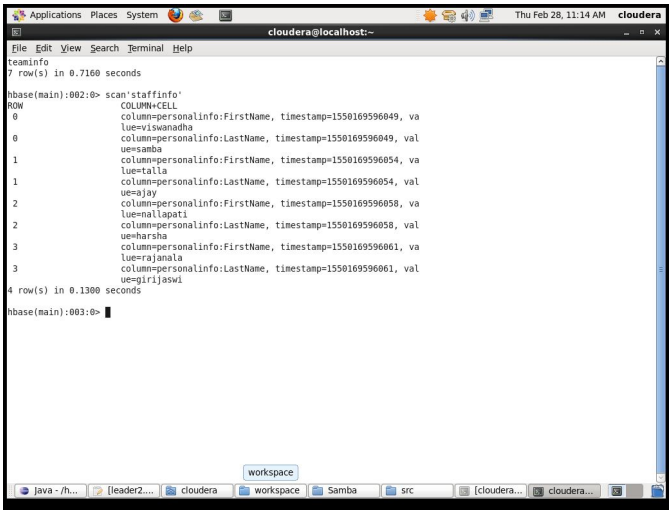

Figure 14: Showing table content in HBase

5. CONCLUSION

Hadoop is one of the most excellent platforms used for store and retrieve the information in huge volume at superior rate. The Hive tool can be used to progression and store up the correct information in a huge database, compared to other data mining and cloud procedures. It cans progression information that has a looser structure than would be possible in a relational database. This enables us to query and examine customary structured data. After this Unstructured data will be format into structured data. Hadoop has freshly adopted YARN which opens up the potential to go beyond map/reduce without altering everything. Note that some of the novel options do have migration path and also we still retain the access to all that big data we have in Hadoop as the extended use again of a number of the ecosystem.

.

REFERENCES

1. Afrati, F.N. & Ullman, J.D. **Optimizing Multiway Joins in a Map-Reduce Environment**. IEEE Transactions on Knowledge and Data Engineering, Vol.23, Issue.9, pp.1282-1298, Feb.2011.
2. Arba Asha Altaye, Dr. J. Sebastian Nixon **A Comparative Study on Big Data Applications in Higher Education,** International Journal of Emerging Trends in Engineering Research,Volume 7. No. 12, pp.739-745,December 2019.
3. Gu et al. **SHadoop: Improving MapReduce Performance by Optimizing Job Execution Mechanism in Hadoop Clusters**. Journal of Parallel and Distributed Computing, Volume74 Issue.3, pp.2166-2179, 2014.
4. Zaharia M. et al. **Improving Map reduce performance in heterogeneous environments.** Proceedings of the 8th USENIX conference on Operating systems design and implementation (OSDI), San Diego, California, USA. pp.29-42, 2008.
5. K.Jose Triny et al. **A Bigdata processing with Hadoop Map Reduce in Cloud Systems** ,International Journal of Emerging Trends in Engineering Research, Volume 8. No. 3, pp.752-758, March 2020.
6. Installation of cloudera: https://www.youtube.com/watch?v=oNQ812My5Hs
7. www.go.gliffy.com
8. Execution of a mapreduce program: https://www.acaglid.com
9. Hadoop mapreduce tutorial: https://www.youtube.com/watch?v=SqvAaB3vK8U
10. https://www.tutorialspoint.com/hbase/hbase_architecture.html