International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Text Mining with Apache Hadoop over different Hadoop Clusters Architectures

E. Laxmi Lydia, Gorapalli Chandra Sekhar,Madhu BabuChevuru, Dasari Ramya, K. Vijaya Kumar

Abstract: ***Big data is very much practical for real time applicational systems. One of the mostly used real time application worldwide are on unstructured documents. Large number of documents are managed and maintained through popular leadingBig Data platform is Hadoop. It maintains all the information at Hadoop Distributed File System in Blocks. Irrespective of datasize, BigData has opened its path to store and analyze the data which has consumed time. To overcome this, Hadoophas designed cluster process for large volumes of unstructured data computations. Three different cluster architectures like Standalone, Single node cluster and multi node clusters are considered. In this paper, Big Data allows Hadoop platform to boost the processing speed overlarge datasets through cluster architectures, which are studied and analyzed through text documents from newsgroup20 dataset.It identifies the challenges on text mining and its applications using ApacheHadoop.***

Keywords:Big Data, Hadoop Cluster, Standalone mode, Pseudo cluster mode, Fully Distributed mode,

I. INTRODUCTION

Data processing for single centralized resource is a dedicated network carried out by independent components of Hadoop cluster. These systems incluster are named as “Shared Nothing” systems as they share nothing maong the nodes except network connection.

This will minimize the latency of processing in Hadoop cluster whenever queries are posed in large database of data. Hadoop clusters maintain virtual machines over cloud to overcome energy consumption and enhance the productivity. Hadoop empowers the technology of distributing process over high degree fault tolerance. It is applicable tool used especially for Big Data.

Revised Manuscript Received on 30 July 2019.

***** Correspondence Author

E. Laxmi Lydia*, Associate Professor, Department of Computer Science Engineering, Vignan's Institute of Information Technology,Andhra Pradesh, India.

Gorapalli Chandra Sekhar,PG Schloar, Department of Computer ScienceEngineering,Vignan's Institute of InformationTechnology, Andhra Pradesh,India.

Madhu Babu Chevuru, Asst. Professor,Department of Computer Science Engineering,VFSTR deemed to be University.

Dasari Ramya,Junior Research Fellow,Departmentof Computer Science Engineering,Vignan's Institute of Information Technology, Andhra Pradesh, India.

K. Vijaya Kumar,Associate Professor, Department of Computer Science Engineering, Vignan's Institute of Information Technology,Andhra Pradesh, India.

Key constraints required for Hadoop cluster

- Hadoop clusters are focused design processing clusters pointed at analyzing and storing for unstructured data. Hadoop clusters are data processing computational clusters that distribute the work over multiple cluster nodes in parallel.

- Hadoop operates for partitioning the data into chunks and distributed every chunk to an individual cluster node for analysis.
- Data nodes in Hadoop cluster may not be uniform for handling processes in cluster node.

Following are some of the most favouring circumstances in Big Data through Hadoop clusters:

- A Hadoop cluster absolutely performsparallel processing to help with the analysis,but clusters are lackingdue to increase of data theprocessing power. These clusters are scaled to maintain analysis by making modifications to the additional appended cluster nodes that has application logic.
- Hadoop clusters are inexpensive for commodity hardware, so they can be used in any organization but server hardware is expensive.
- Cluster nodes wont create any problem if single data node fails, it has an replication files with other nodes (replication factor). It becomes an challenging issue if the NameNode fails in a cluster.

II. LITERATURE SURVEY

Ankita Saldi et al[1], focused on statistical data generated in industries. When the generated data is in different formats, environment becomes more challenging to perform functioning. They have choosenHadoop environment to overcome the issues on large data records by appling parallel implementation of datanotes in single node. The authors concentrated on sequential execution of datanodes. Every datanode running at Hadoop cluster using SMs (Streaming Multiprocessors) has obtained profitable desired result for raw industrial data that is collected in a year.

E. Laxmi Lydia et al [2], worked on Flume, MapReduce, Pig and Hive components on Hadoop for traditional database management system considering twitter data. They have examined Hadoop framework by utilizing apache components on enormous information. They identified the performance of Hadoop components throughcomparision. All these have properly organized in Hadoop distributed file system (HDFS) and use such data in easy manner. In 2016, E. Laxmi Lydia et al [4], processed their work on big data with technological improvement by calling a new system data acquisition. This accomplishes three stage proceedings using Hadoop, peocedures and technologies, handling through cloud computing.

Ehab Mohamed et al [3],suggested that Hadoop can manage and work on large datasets at fault-tolerant platform. Their research work has studied on data transmission through blocks in Hadoop with optimal efficacy.Also suggested the improvement of cloud environments for job scheduling on multiple constraints mostly for parallel distributed systems and quality requirements..Sujit Roy et al.[7], also done their research study on data blocks for assigned jobs in Hadoop. Nishant Rajput et al.[14] described the study on HDFS and wraping up the processing data as well as running Hadoop components on clusters. This has attracted clients to work on cluster over servers.

Avanish Sing at al.[8], did their comparative study on virtual machine working in Hadoop environment. They have explained the perception of containerization which describes the cluster importance in Hadoop on a container. The virtual machine uses container services to analyze the random data on distributed file system cluster. It will sort the data through provided services and the result is finally analysed by virtual machine. This has improved the efficiency of the clusters by using containers. This helps user to have optimized data at data centers by services over virtual machine.

Yojna Arora et al[5], explored the solutions to the existing difficulties in handling big data. Major solution is through distributed computing using Hadoop software. They described the processing frameworks in Hadoop like general-purpose processing, Abstraction, SQL, Graph processing, machine learning, real-time/ streaming frameworks with Hadoop architectures having Hadoop core, HDFS (Data Storage), Hbase (Automatic Control), YARN (task and resource) and PIG(data access). K. Tamilselvi et al.[9] explained about similarly but they introduced a term parallelism by facing different challenges while handling large amount of generated data.

C.S.Arage et al.[10] suggested digitizing data in healthcare organizations for treatmenting patients based on the solutions. These are provided by the measures of diagnoses for predicting hospital readmission risks and identifying the cost saving opportunites for the people. Hadoopclusters processes datasets to perform distributed computations and MapReduce programs for Electronic Health Records database.Hadoop process the data with the integrateion of map-reduce as it works with the dynamic read-write layer.

Jayalakshmi DS et al.[6], applied big data processing through data processing engine in cloud computing. Her virtual machines are connected in cloud data centers. Thi will lead in high productivity and reduce resource an energy consumption. They have implemented the challenge faced by the virtual Hadoop clusters based on response tim and energy consumption using various virtualizatio technologies.

Ruchi Mittal and Ruhi Bagga [11], implemente performance analysis on multiple machines under Pseudo-Distributed mode using Hadoop. Worked on parallel distributed programming approach to deal with accelerated applications. The results haveanalyzed the time by performing similar operations on multiple machines.

Fernando G. Tinetti et al.[12], explained assess cluster configuration byverifying parallelization efficiency on independent environments. The constructed softwares are performed on distributed systems to enhance the Hadoop architectures. Yannan Ma et al.[13], suggested new algorithms in distributed systems providing security and allowing high speed on reading files i.e, multi-node reading.

Yicheng Huang et al.[15], interpreted Hadoop clusters through tools based on the requirements that decreases the complexity of large systems. They proposed that Hadoop clusters are powerful and implemented unified models to it. This has reached the real-world benefits on various kinds of services.

III. METHODOLOGY

A. Hadoop Cluster

Hadoop distributed cluster architectures varies in different aspects to process the data from data centers by hardware and software network communicateions. Data centers maintains racks that has similar collection of nodes functioning for different user jobs. A rack may contain 20 to 30 nodes connected to only single server using Ethernet. Clusters in Hadoop uses switch at cluster level. Whenever large clusters are formed then more number of Hadoop architectures are combined. Hadoop cluster involves three major components. They are as follows:

- *Client Nodes* – Hadoop installs client node to load data by setting all the required configurations to Hadoop cluster.

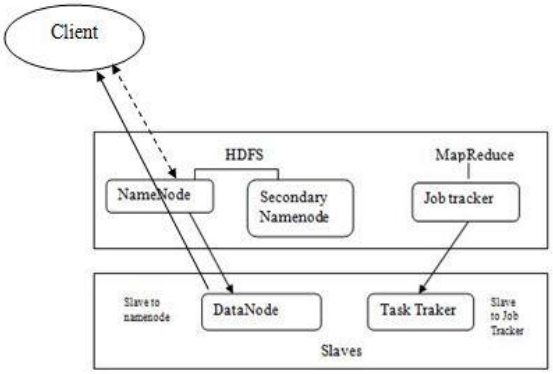

Fig1: connectivity of client node with cluster nodes.

Client node submits information to Jobs by loading data into the cluster and process data.The final output is obtained from the client node after the job is completed as shown in Fig1.

- *Master Node* – This node helps Hadoop cluster to store the data by using HDFS. It allows to perform parallel computations for the stored data using MapReduce.Master nodedescribes the mastor node having its major components HDFS and Mapreduce, internally maintains data of NameNode, Secondary NameNode and JobTracker shown in Fig2.Here every single node has its own function to be performed.
  - → *NameNode* maintains a record of incoming data to DataNode, it keeps track of the access time of the files called metadata.
  - → *Secondary NameNode*is the backup node for Namenode. It will be helpful in critical situations likenamenode failures, backups.
  - → *JobTracker* uses MapReduce to check the processing of data parallely.

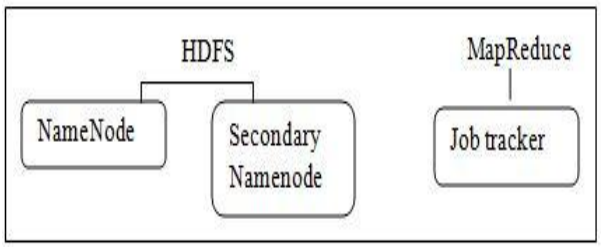

Fig2: Specified node to describe Master node

- *Slave Node*- Slave node is also known as worker node. This node is responsible for functioning jobs assigned by the user to store as well as perform computations. A single master node can have multiple slave nodes.

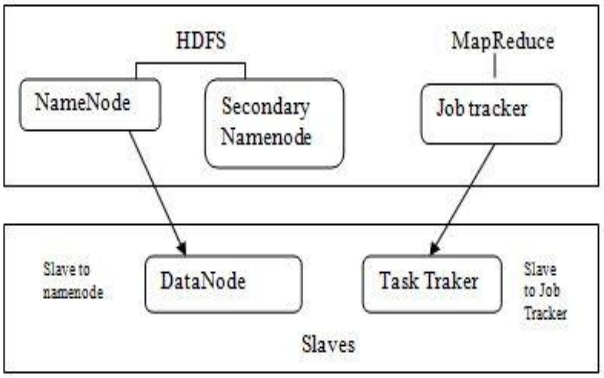

Fig3: Specified nodes to function Slave node

Each single slave node runs on TaskTracker and a DataNode of Master node in the cluster for communication. DataNode serviceis inferior to the NameNode and TaskTracker service isinferior to the JobTracker as shown in Fig3.

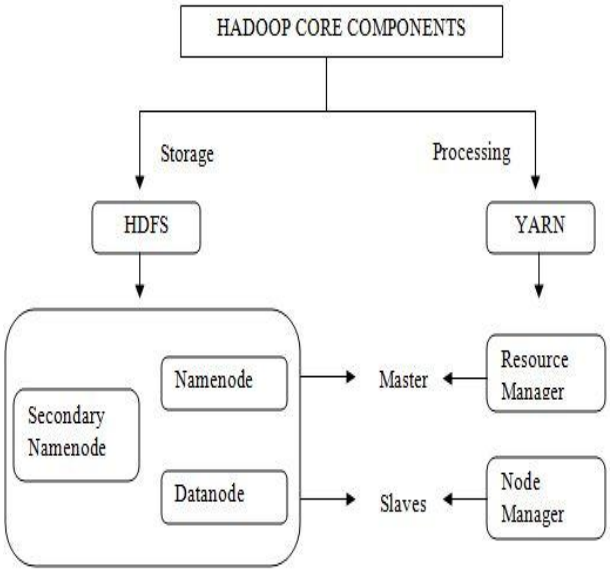

Fig4: Core components in Hadoop

Figure4 describes the essential core components for Hadoop cluster for storage and preprocessing.

B. Configuration requirements for Hadoop Cluster

Hadoop cluster is build with connection of minimum three computer systems, which are also known as nodes. These systems are connected each other to work in parallel by monitoring from single system known as cluster. Clusters using Hadoop software is known as Hadoop cluster. It is designed especially for analyzing and processing vast unsupervised data in distributed environment. Hadoop clusters are sequentially organized using racks. Different configurations of Hadoop clusters are in three different modes.

a. Standalone cluster mode:
This acts as the default mode for Hadoop clusters. Here HDFS is not applied in this mode. Every input and output are handled by the local file system. This cluster is put in use to handle debugging. Three xml files like mapreduce (mapred-site.xml), hdfs (hdfs-site.xml) andcore (core-site.xml) are used to configure the standalone Hadoop cluster setting.

b. Single node Hadoop cluster- Pseudo Distributed mode:
This cluster allows all the daemons to run at single node itself. Every node is specified by its work, one node for master node, one node for data node, job tracker, task tracker. Pseudo distributed mode has most advantageous factor called replication factor, which is helpful for backups. It store the data at various places basd on the replication factor. It uses three configuration files to setup HadoopPesudo distributed Hadoop cluster.

c. Multiple node Hadoop cluster- Fully distributed mode:
In this cluster mode, data is distributed over many different nodes.

Table 1: List of all Hadoop configuration files

File Names	Description for configuration filenames
Hadoop-env.sh	Hadoop runs on environment variables.
core-site.xml	Input/Output settings are configured to HDFS and MapReduce.
hdfs-site.xml	Daemons configuration settings to (namenode,secondary namenode, datanode).
mapred-site.xml	MapReduce settings are configured.
yarn-site.xml	Resource and Node manager settings are configured.
masters	Maintains number of systems that run secondary NameNode.
slaves	Maintains number of systems that run a datanode and a Node manager.

C. Procedure for Standalone node Hadoop cluster

Step1: Start the cluster by considering a single system
*Step2:*Download and install Java and set path
Step3: Download and install Hadoop and set path
*Step4:*Execute simple programs over MapReduce

D. Procedure for Pseudo node Hadoop cluster

Step1: Start the cluster by considering a single system
*Step2:*Download and install Java and set path
Step3: Configure SSH connectivity

Step4: Download and install Hadoop and set path
Step5: Open Hadoop configuration files and edit mapred-site.xml, hdfs-site.xml and core-site.xml files.
Step6: Check Hadoop directory and format Namenode
Step7: After formatting starts all the daemons. We can check using jps command for running nodes like Namenode, Resource Manager, DataNode, Secondary Namenodes, JobTracker, TaskTracker.
Step8: Browse the status and execution of nodes in the web.

E. Procedure for Multi node Hadoop cluster

Step1: Start the cluster by considering any number of systems. Consider one system as master (NameNode) and others as slaves (Datanodes)
*Step2:*Check their IP address and verify its connectivity through ping
Step3: Configure SSH connectivity in all nodes
*Step2:*Download and install Java and set path
Step4: Download and install Hadoop and set path
Step5: Open Hadoop configuration files and edit mapred-site.xml, hdfs-site.xml and core-site.xml, yarn-site.xml files for all nodes.
Step6: Create Namenode only in the Master node.
Step7: Create DateNodein all Slave nodes.
Step8: format Namenode in Master node
Step9: After formatting starts all the daemons. We can check using jps command for running nodes like Namenode, Resource Manager,Secondary Namenodes in Master node and jps command for running nodes like DataNode and Node Manager
Step8: Run programs and browse the status as well as execution of nodes in the web.

IV. RESULT ANALYSIS

For the result analysis, text documents from newsgroup20 are taken and compared with respect to the time and size of the documents. Documents with size 20 KB, 50 MB, 250MB, 750 MB, 1GB are considered and executed over different Hadoop cluster modes shown in Table2.
Fig5 describes the comparative analysis of Hadoop clusters. The graph analysis was based on Table2 results. We observed that, when fewer documents were given to execute i.e, with less size of data in KB, Hadoop standalone cluster mode and fully distributed mode are generating results in short time. When more number of documents were given to execute i.e, with more size of data in MB and GB, Hadoop Fully distributed cluster moed is generating results in short time.For large data we can observe that fully distributed mode are more advantages that Standalone and Pseudo distributed mode.

Table2: Data Size and time taken to execute in different modes of Hadoop clusters.

Data Size	Time Taken(in seconds)		
	Standalone	Pseudo	Fully Distributed
20 KB	4.6 sec	4.6 sec	4.4 sec
50 MB	119.66 sec	115.6 sec	98.7 sec
250 MB	260.7 sec	243.3 sec	119.1 sec
750 MB	706.9 sec	650.45 sec	520.32 sec
1 GB	945.86 sec	843.31 sec	674.6 sec

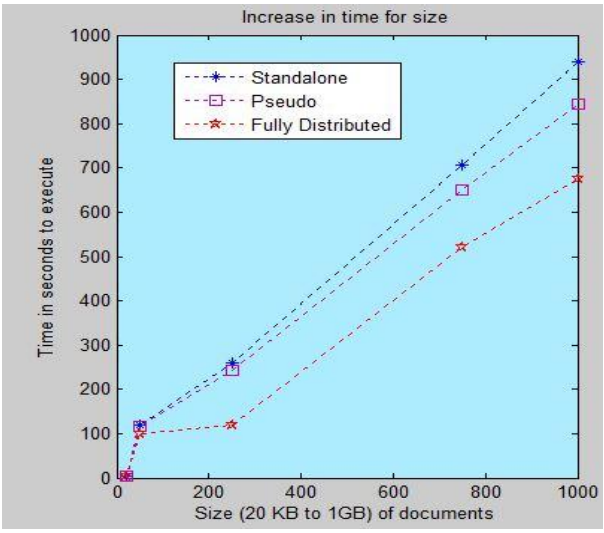

Fig5:Comparative analysis on Hadoop cluster modes.

V. CONCLUSION

Hadoop enables its framework over different distributing processing. Its scalability for large datasets, cost-effectiveness for storing data, flexibility for unstructured data, speed have grabbed large organizations to adopt Hadoop platform. This paper describesdifferent modes of Hadoop clusters clearly andHadoop clusters using Java codes for large text documents from newsgroup20 datasets on various distributed systems.Study and Performance of different cluster modes identified that large datasets are very much advantageous on fully distributed mode and datasets with less data can be well performed over Standalone and Fully- Distributed. This contributesrunning of low cost commodity computers.

REFERENCES

1. Ankita Saldi, Abhinav Goel, Dipesh Yadav, Ankur Saldhi, Dhruv Saksena, S. Indu, “Big Data Analysis using Hadoop Cluster”, 2014 IEEE International Conference on Computational Intelligence and Computing Research, 978-1-4799-3975-6/14.
2. E. Laxmi Lydia, M. Ben Swaroop, “ Big Data Analysis using Hadoop Components like Flume, MapReduce, Pig and Hive”, IJCSET, November 2015, vol5, Issue 11, 390-394.
3. Ehab Mohamed, Zheng Hong, “Hadoop- MapReduce Job Scheduling Algorithms Survey”, 2016 International Conference on Cloud Computing and Big Data, 978-1-5090-3555-7/16. DOI 10.1109/CCBD.2016.54
4. Dr. E. Laxmi Lydia, M. Vijay Laxmi, Dr.M.Ben Swarup, “A Literature Inspection on Big Data Analytics”, International Journal of Innovative Research in Engineering and Management(IJIREM), ISSN:2350-0557, Volume-3, Issue-5, September-2016.

5. YojnaArora, Dr. Dinesh Goyal, "Hadoop: A framework for Big Data processing & storage", International Journal of Application or Innovation in Engineering of Management(IJAIEM), ISSN 2319-4847, Volume 6, Issue 7, July 2017.
6. Jayalakshmi D S, Syeda Rabiya Alam, R.Srinivasam, "Approaches to Deployment of Hadoop on cloud platforms: Analysis and Research Issues", IEEE International Conference on recent trends in Electronics Information & Communication Technology(RTEICT), May 2017, India.
7. Sujit Roy, Subrata Kumar Das, IndraniMandal, "Hadoop Periodic Jobs using data blocks to achieve efficiency", International Journal of scientific research in computer science engineering and Information technology (IJSRCSEIT), ISSN:2456-3307, volume 3, Issue 3, 2018
8. Avanish Singh, P.Gouthaman, ShivankitBagla, and AbhishekDey, "Comparative study of Hadoop over containers and Hadoop over Virtual Machine", International Journal of Applied Engineering Research, ISSN 0973-4562, volume 6, Issue 6 (2018), pp. 4373-4378
9. K. Tamilselvi, V.Sumithra, K. Dhanapriyadharsini, "Big data analytic using Hadoop Technology", International Research Journal of engineering technology(IRJET), e- ISSN: 2395-0056, Volume 05, Issue 01, Jan-2018
10. Mr.C.S.Arage, M.P. Gaikwad, Rohit Tadasare, Ronak Bhutra, "Analyse Big Data Electronic Health Records Database using Hadoop Cluster", International Research Journal of Engineering and Technology(IRJET),e- ISSN: 2395-0056, Volume 05, Issue 03, Mar-2018
11. Ruchi Mittal and RuhiBagga,"Performance Analysis of Hadoop with Pseudo-Distributed Mode on Different machines", International Journal of Computer Sciences and Engineering (JCSE), vol.-3(6),PP(113-117) June 2015, E-ISSN: 2347-2693.
12. Fernando G. Tinetti, Ignacio real, Rodrigo Jaramillo, and Damian Barry, "Hadoop Scalability and Performance testing in Heterogeneous Clusters", Internation Conf. Par. And Dist. Proc. Tech. and Appl. PDPTA'15.
13. Yannan Ma, Yu Zhou, Yao Yu, Chenglei Peng, Ziqiang Wang, Sidan Du, " A Novel approach for improving security and storage efficiency on HDFS", Elsevier ScienceDirect International Conference on Ambient Systems, Networks and Technologies(ANT 2015), 631-635.
14. Nishant Rajput, Nikhil Ganage, Jeet Bhavesh Thakur, "Review paper on Hadoop and MapReduce ", International Journal of research in engineering and technology, vol.-6(9), sep-2017, eISSN:2319-1163.
15. Yicheng Huang, XingtuLan, Xing Chen, WenzhongGuo, " Towards Model-Based Approach to Hadoop Deployment and Configuration", 2015 12th Web Information System and Application Conference (WISA). DOI:10.1109/wisa.2015.65.

FUNDING ACKNOWLEDGMENT

This work is financially supported by the Department of Science and Technology(DST), Science and Engineering Research Board(SERB) under the scheme of ECR. we thank DST-SERB for the financial support to carry the research work.