International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 8, ISSN No. 2455-2143, Pages 221-223
Published Online December 2019 in IJEAST (http://www.ijeast.com)

ENHANCING PERFORMANCE AND EFFICIENCY FOR BIG DATA ANALYTICS APPLICATION IN HADOOP MAPREDUCE ENVIRONMENT

Purushotham Naidu V
M.Tech Scholar,
Department of Computer Science and Engineering
BMS Institute of Technology and Management
Doddaballapura Main Road, Avalahalli,
Yelahanka, Bengaluru-560064
Karnataka, India

Anjan K Koundinya
Associate Professor and PG Coordinator
Department of Computer Science and Engineering
BMS Institute of Technology and Management
Doddaballapura Main Road, Avalahalli, Yelahanka,
Bengaluru-560064
Karnataka, India

***Abstract* - Hadoop finds its area in Big Data Analytics for analysing huge amounts of data. Hadoop implements MapReduce to distribute data to different clusters. Data compression is adopted in order to reduce the memory space occupied by the data. The concept of MapReduce performance with Data Compression focuses on a number of compression codecs of Hadoop cluster such as snappy, gzip, Iz4, bzip2 and deflate. The Big data analytics in health care faces good benefits and also with all its associated components focuses with the proposal of a big data health care architecture. Big data analytics is an emerging field for extraction of closely connected information from very huge data-sets and focuses on the improvement of decision making with improved decision making. The educational system and academic trends of students needs to map up with the current trends in technological advancement which accumulates large amount of data which is unstructured and needs to be analysed. Data mining tools are required to obtain information with meaning by converting unstructured data to structured data. Data has become necessary part of every individual, industry, economy, business function and organization. As this data set increases, selecting the relevant information becomes a tedious task. The on-command and on-demand nature of digital universe gives creation of a data category called the Big Data because of its sheer volume, variety and velocity. It proposes computational and analytical challenges which includes measurement errors, scalability and storage bottleneck and noise accumulation**.

Index Terms – Big Data, Hadoop, MapReduce, HDFS, Data Mining.

I. INTRODUCTION

The field of Big Data has attained fast expansion by providing various tools to manage, accumulate and analyse data to make better decisions. Big data in health care focuses on huge benefits by improving the quality of detecting diseases at earlier stages and effective in the medical care. Cloud computing also plays an important role which provides an on demand close services for processing, analysing and storing huge amount of data. Big data includes many tools and platforms for analytics in health care such as Cassandra,

Hadoop Distributed File System (HDFS), MapReduce, Mahout, HBase, PIG, Hive and Zookeeper.

The study focuses on two scenarios, firstly data compression and map output are utilised. The execution time is better for input file with raw-text as snappy and deflate only. Next the results are compressed for map output which does not increase the performance compared to the uncompressed data. Secondly, compressed input file of bzip2 is utilised with the uncompressed MapReduce. The bzip2 files as input are compatible to word count job like raw text file. This can save storage space more than 70 percent of raw text file. Hadoop benchmarks which also has other applications for large output file of reduce phase are also focused.

The Big data in health care can be used to understand Structured, Semi Structured or Unstructured data and handle in an efficient way. Enhanced measures will be considered to cure the diseases. The field of Cloud computing is also being used to process huge (big) data in health care. Cryptographic techniques can be adopted and implemented to achieve a good frame work for sharing of sensitive information with patients on cloud with improved security.

II. METHODOLOGY

The compression codecs of Hadoop cluster as mentioned in the abstract are configured in an XML configuration file as codec properties. The research includes the utilisation of data compressed with word count MapReduce as follows:

A. *Scenario 1 (Map output compression)*: Raw-file, gzip and bzip2 of 4.8 GB was used as an input file. The MapReduce process utilises the compression codec's such

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 8, ISSN No. 2455-2143, Pages 221-223
Published Online December 2019 in IJEAST (http://www.ijeast.com)

as snappy, gzip, Iz4, bzip2 and deflate to obtain a compressed map output.

B. Scenario 2 (Input file compression): A variety of input size such as 4.8 GB, 7.2 GB, 9.6 GB and 14.4 GB was used. Uncompressed map and reduce process was used in MapReduce process.

In all the scenarios only 1 master node with 3 replications and 4 data nodes are used. A study based on data compression with MapReduce in Hadoop cluster was presented.

The Big data health care architecture is used to process and analyse large scale health care data on cloud computing environment with Hadoop clusters. The traditional software systems are replaced by the health care applications by the adoption of cloud services for processing. The records of health accumulated from various sources as datasets will be input to the NOSQL Database such as MongoDB and performance test is conducted. A NOSQL database such as MongoDB is used to enhance MapReduce with the help of an underlying Hadoop framework.

An enhanced version of data processing on Hadoop Clusters in cloud environment is the main focus. The process includes scalable applications since it runs on large servers and can accumulate many users. The generated data is sent to other servers and the cloud where the data is integrated together using Hadoop MapReduce and which can be used to obtain and manage the results at a faster rate. The model contains various other components such as analytics of health care, batch scheduling, standard schema validator, processing, semantic practitioner, Mongo data reader, query formulator, processing layer, Big data container.

III. IMPLEMENTATION

Big data is an emerging field of data mining. Data mining plays an important role in the field of education where progress of society is a key aspect. The implementation of Apriori algorithm provides high performance with the technique of MAP reduce in Hadoop framework. The pattern of student's choice for industrial training course combinations is predicted after processing through MAP Reduce Hadoop Data Mining Technique.

The input dataset is collected from students and MAP Reduce technique is applied which is stored in HDFS. The input is provided to the mapper which maps data to the output which in prior is split into various clusters. The combiner obtains the output from the mapper for combining the output together and sending it to the reducer. Hadoop organises the work by diving the tasks into Map and Reduce tasks. Hadoop Distributed File System (HDFS) includes the following components such as Mapper, Reducer, Name Node, Data Node, Resource Manager and Job Tracker.

Jobs can be made to run over HDFS with the help of MapReduce. Hadoop clusters are utilised for applications that can be scalable to run as a cluster over multiple machines with the help of MapReduce. The reducer's function is to display the output in the form of <key, value> pair after aggregating the tuples derived from the mapper once the input is fed to the mapper.

The input data as large volume containing course combinations is passed through the mapper function using HDFS which runs on a single node cluster parallel using HDFS. Meaningful data is obtained from the Reducer function by converting data into meaningful data by individual tuples. This strengthens the decision making within students by using the reducer which classifies the data opted by a greater number of students and enhances the decision making of institutions as well as students more demanding courses for industrial trainings.

IV. CHALLENGES

Data is the aggregation of characteristics and components that are related to each other in some sense and varies in some other. Large volumes of raw data are created every day. Data is not only produced in large volumes but also in variety, veracity and velocity.

The following are the issues and challenges concerned with Big Data: Complexity, Incompleteness, Timeliness, Security, Scale, Privacy and Heterogeneity. The Big Data mining techniques are as follows: Hadoop, MapReduce.

The challenges concerned in deployment of Hadoop are as follows: Troubleshooting is difficult, investing in more hardware than required/used, mixed workloads, Complex system with low level API's, Specialised Skills.

The solution for the above is as follows: Instead of maintaining a single cluster, clusters can be grouped according to the task they perform, providing a high degree of separation. Extensive optimization of Hadoop by existing tools and tuning methods. Sizing the cluster by taking every day's load pattern of the previous few months or weeks, and then scaling it as when the need be, could be more cost effective. It is much more challenging to deploy Hadoop over system running Linux, instead of Windows or OS X.

V. RESULTS AND DISCUSSIONS

Results have been observed that the execution time between bzip2 and raw text is the same. But the advantage is that bzip2 reduces the memory space which faces a performance improvement and the Hadoop compression can be used to compute the word count MapReduce execution time with a bzip2 as an input file in the Hadoop cluster.

Hadoop Distributed File System (HDFS) stores huge volumes of data. MapReduce is utilised to extract Knowledge to help students in decision making and select courses relevant to their industrial training. Preferable courses are derived based on course combination for training the students. HDFS is used for running the tasks over MapReduce and the aggregated results are obtained.

VI. CONCLUSION

A study based on data compression with MapReduce in Hadoop cluster was presented. The focus is also on Hadoop benchmarks which also has other applications for large output file of reduce phase.

The proposal of Big data analytics in health care improvises the results by eliminating the traditional practices of using EHR, EMR, etc., by doctors and health care professionals for decades whose performance is still unpredictable.

VII. REFERENCES

[1] Maninder Jeet Kaur and Ved P Mishra (2018), "Analysis of Big Data Cloud Computing Environment on Healthcare Organizations by implementing Hadoop Clusters", *The Fifth Hct Information Technology Trends (Itt)*, Dubai, Uae, Nov. 28 - 29.

[2] Rahul Khullar, Tushar Sharma, Tanupriya Choudhury and Rajat Mittal (2018), "Addressing Challenges of Hadoop for BIG Data Analysis", *International Conference on Communication, Computing and Internet of Things (IC3IoT).*

[3] Kritwara Rattanaopas and Sureerat Kaewkeeree (2017), "Improving Hadoop MapReduce Performance with Data Compression: A Study using Wordcount Job", *14th International Conference on Electrical Engineering / Electronics, Computer, Telecommunications and Information Technology (ECTI-CON).*

[4] Pratiyush Guleria and Manu Sood (2017), "Big Data Analytics: Predicting Academic Course Preference Using Hadoop Inspired MapReduce", *Fourth International Conference on Image Information Processing (ICIIP).*

[5] Ahmed Qasim Mohammed and Rajesh Bharati (2017), "An Efficient Technique to Improve Resources Utilization for Hadoop MapReduce in Heterogeneous system", *International Conference on Intelligent Communication and Computational Techniques (ICCT) Manipal University Jaipur, Dec 22-23.*

[6] Motahar Reza, Badrinath Tripathy, Harsh Ranjan and G.Anant Kumar (2017), "Study and Analysis of Hadoop Cluster Optimization based on configuration properties", *International Conference on Innovations in Power and Advanced Computing Technologies [i-PACT].*

[7] PrathyushaRani Merla and Yiheng Liang (2017), "Data Analysis using Hadoop MapReduce Environment", *IEEE International Conference on Big Data (BIGDATA).*

[8] Mugerwa Dick, Jeong Geun Ji and Youngmi Kwon (2017), "Practical Difficulties and Anomalies in Running Hadoop", *International Conference on Computational Science and Computational Intelligence.*

[9] Nivedita V and Geetha J (2017), "Optimization of Hadoop Small File Storage using Priority Model", 2nd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, India.

[10] Sushila Maheshkar, Bhavishya Mathur, Raj Roushan and Ajay Kumar Mallick (2017), "Automatic Hadoop cluster deployment and Management tool", *Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN).*

[11] Ahmed Qasim Mohammed and Rajesh Bharati (2017), "An Efficient Technique to Improve Resources Utilization for Hadoop MapReduce in Heterogeneous system", *International Conference on Intelligent Communication and Computational Techniques (ICCT) Manipal University Jaipur, Dec 22-23.*

[12] Pavan Kumar Pagadala, M. Vikram, Rajesh Eswarawaka and P. Srinivasa Reddy (2017), "Join Operations to Enhance Performance in Hadoop MapReduce Environment", *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications.*

[13] Guilherme W. Cassales, Andrea Schwertner Charao, Manuele Kirsch-Pinheiro, Carine Souveyet and Luiz-Angelo Steffenel (2016), "Improving the Performance of Apache Hadoop on Pervasive Environments through Context-Aware Scheduling", *Journal of Ambient Intelligence and Humanized Computing*, March.

[14] Priyam Jain, Satyaranjan Patra and Pankaj Richhariya (2016), "Enhance Performance of Mapreduce Job on Hadoop Framework using Setup and Cleanup", *International Journal of Computer Applications (pp. 0975 – 8887)* Volume 155 – No 8, December.