Big Data and Its Implementation using HADOOP

Suraj Bala[1] Prof. Anil Mishra[2]
[1]M.Tech Scholar [2]Professor
[1,2]Department of Computer Science Engineering
[1,2]GITM, Gurugram, India

Abstract— In this paper, we are showing that how we can store and process high volume of data (terabytes or petabytes) on computer clusters built from commodity (cheap) hardware. For this, we are using "Apache Hadoop" a software framework that supports data-intensive distributed applications under a free license. In this paper a proposed technique for implementing the storage and processing of huge data sets we are installing a Hadoop Vanilla Cluster (Multi node cluster setup - Pseudo Distributed Mode) on Ubuntu Environment. Thereafter, showing how we can do data ingestion, processing and migration of data using different Hadoop Eco – Systems. The success rate of our proposed scheme depends upon the extraction and configuration techniques which have been applied in this work.

Key words: Big Data, Cluster, Data Processing, Hadoop, Extraction, HDFS

I. INTRODUCTION

There is a rampant increase in the amount of data being produced from varied sources. This can be attributed to the instrumentalisation of the current society and personnel's leading to storage and production of vast amounts of data. Since, the data being produced is huge with a lot of variety and the rate of production is also rapid. Thus, the traditional systems fail to manage this data and this is what led to the buzz word called Big Data. Big Data is a term which refers to the explosion of variety of data produced from disparate sources [1]. It is characterized by five features or attributes i.e. high volume, variety, veracity, visibility and velocity. Since, this kind of data is beyond the management scope of traditional systems therefore in order to mine such kind of data we need analytics' solutions that can help in gaining insights from both structured and unstructured data. At present scenario its instrumental to blend both big data and analytics into a single entity termed as big data Analytics. Analytics involves examination of data to derive meaningful insights such as hidden patterns and trends that can in turn benefit the organizations in making important business decisions and developing newer business models. The problem of data deluge imposes potential challenges involved in processing and extracting useful information from data. It also requires skills for management and analysis of huge data sets. Cloud computing serves as a quintessential solution for handling big data and hosting big data workloads. Cloud computing has revolutionized the way in which computing resources can be utilized by providing facilities such as pay per use, rapid elasticity and dynamic scalability. It provides the users with an illusion of infinite storage and compute capacity. The cloud resources can be used in private mode through private cloud or can be shared publicly using a public cloud such as Amazon EC2 and Microsoft Azure. Cloud therefore serves as a scalable technology with low upfront investment costs. Thus, the proposition value associated with using cloud as a platform for carrying out analytics is quite strong and therefore it is well suited for carrying out scalable data analytics. Hadoop is a technology that can be used for handling big data. It can play a significant role in opening gates to new insights out of data and can easily handle flood of huge unstructured data sets coming from sources such as sensors, mobile devices and social media. This paper presents about how hadoop can be used as technology on cloud for meeting the big data needs of users and discusses about the proposed hadoop based workflow for handling big data. We also present a case study of analysis carried out on movie data for mining many useful information from it which includes finding the number of movies released between a given period and the number of movies having a certain rating besides other informations.

II. RELATED WORK

In the research paper, the researchers have discussed the assisting developers of BDA Apps for cloud deployments. The research work outline challenges in analyzing big data for both *data at rest* and *data in motion*. They have described two kinds system for big data which is at rest namely NoSQL systems for interactive data serving environments and systems for large scale analytics based on Map Reduce paradigm, such as Hadoop, The have discussed that the NoSQL systems are designed to have a simpler key-value based data model having in-built *sharding*, so this work in a distributed cloud based environment. In contrast, to run long running decision support and analytical queries consuming and possible producing bulk data by use Hadoop based systems. For processing data in motion, they have present use-cases and illustrative algorithms of data stream management system. In a research paper explained that an thought-provoking thing of the cloud paradigm the cost of using 1000 machines for 1 hour, is the same as using 1 machine for 1000 hours, these implies that a Hadoop job's performance can potentially be improved, while incurring the same cost, this proved Hadoop is put organized to feat parallelism.

III. HADOOP: INTRODUCTION

Hadoop is an opens source software framework which is used for distributed storage and it can process the datasets of big data by using Map reduce programming model. The model consists of computer clusters which are being built from commodity hardware. In Hadoop the modules are designed with fundamental assumption in which the hardware fails with common occurrence and it is automatically handled by framework. The Hadoop distributed file system consists of a storage part in the core of apache Hadoop. Hadoop is spited up into large blocks and among clusters it is distributed. The code which is packaged

is transferred into nodes to process the data parallel. The nodes are manipulated to access the data which takes data locality as the advantage. The process becomes faster and efficient when it uses conventional supercomputer architecture where computation and data are distributed through high speed networking. Map reduce programming model is an implementation process which is used for large-scale data processing.

A. *Hadoop Architecture*

The package in Hadoop contains the java archive files and some of the scripts which help to start the Hadoop. The package provides file system and operating system level abstractions and a Mapreduce engine and a Hadoop distributed file system.

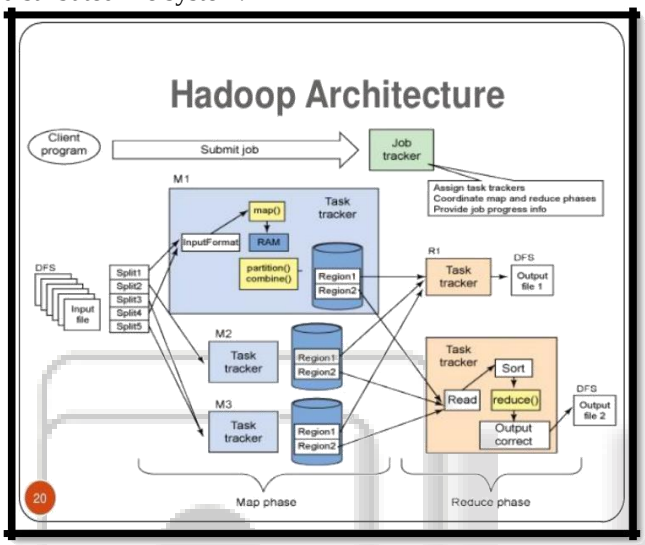

Fig. 1: Hadoop Architecture

To work with effective scheduling every Hadoop file system provides awareness on location and name of the rack where the worker node exist. Hadoop application uses this information which is provided by Hadoop file system to execute the code on the node where the data is and fails when it uses same rack or switch that reduces backbone traffic.

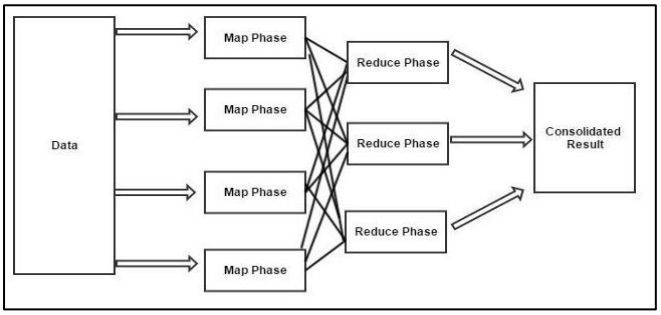

Figure 2: Hadoop Map Reduce model

This method is also used in Hadoop distributed file system to replicate the data across multiple racks. To reduce the impact of rack power outage or switch failure this approach is used. A single master and multiple worker nodes have been included in the small Hadoop cluster. In master node job tracker, task tracker, data and name node has been included. In data node and task tracker, it is possible to have only data and can compute inly worker nodes. These nodes are usually used in non-standard applications. The most requirements for Hadoop are java runtime environment. The term Secure Shell (SSH) can be set up between two nodes which will be used for standard start up and shutdown scripts.

B. *Reliability:*

In hadoop by default three copies of a dataset is made i.e. by default replication factor of a hadoop job is three. This feature offers its users with a reliable and robust framework because even if one of the machines holding that data set goes down the system will still be running as the data will be available at other replicated locations.

IV. HDFS ARCHITECTURE

HDFS has master/slave architecture. An HDFS cluster consists of a single Name Node, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of Data Nodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files.

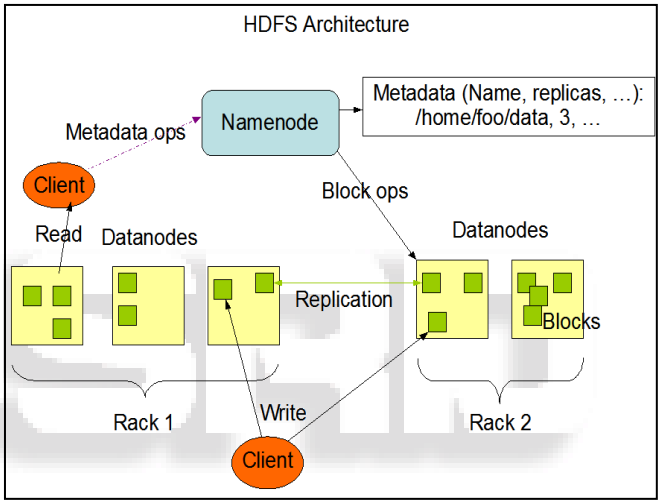

Fig. 3: HDFS Architecture

Internally, a file is split into one or more blocks and these blocks are stored in a set of Data Nodes. The Name Node executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to Data Nodes. The Data Nodes are responsible for serving read and write requests from the file system's clients. The Data Nodes also perform block creation, deletion, and replication upon instruction from the Name Node.

The Name Node and Data Node are pieces of software designed to run on commodity machines. These machines typically run a GNU/Linux operating system (OS). HDFS is built using the Java language; any machine that supports Java can run the Name Node or the Data Node software. Usage of the highly portable Java language means that HDFS can be deployed on a wide range of machines. A typical deployment has a dedicated machine that runs only the Name Node software. Each of the other machines in the cluster runs one instance of the Data Node software. The architecture does not preclude running multiple Data Nodes on the same machine but in a real deployment that is rarely the case.

The existence of a single Name Node in a cluster greatly simplifies the architecture of the system. The Name Node is the arbitrator and repository for all HDFS metadata.

The system is designed in such a way that user data never flows through the Name Node.

V. WORKING AND ORGANIZING MAP REDUCE

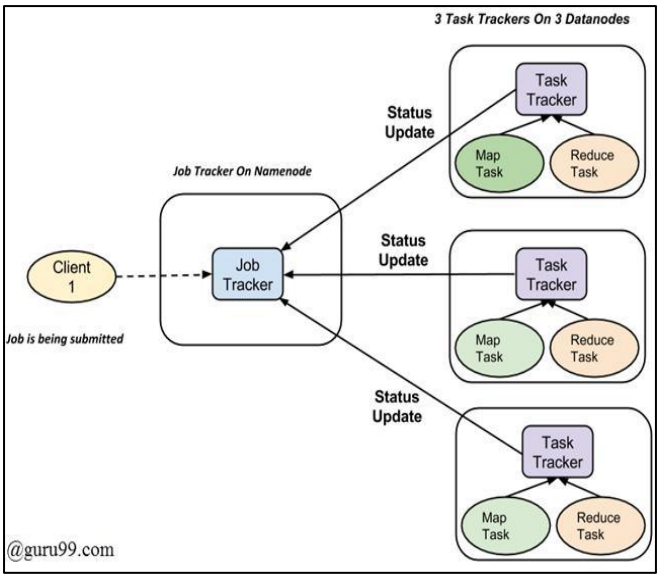

Fig. 3: Map Reduce

Map reduces works or organizes in two phases that is map and reduce. The map splits and processes the data. Reduce shuffles the data. There are two entities which complete the process namely, job tracker and multiple task trackers. Job trackers acts like the masterpiece which is responsible for the complete execution of all submitted jobs. Multiple task trackers act like the slaves. The multiple data nodes can run into a cluster when a job is divided into multiple tasks. To run on different nodes job tracker has to coordinate with the activity which is been scheduled by tasks. The job of a task tracker is to send the progress to job tracker. In the event of task failure, the job tracker can reschedule the jobs on different task tracker.

A. Hadoop MapReduce

Map reduce is an application which is used to write large number of data in parallel or on large clusters of commodity hardware in a reliable manner. It is based on a distributed computing which is a processing model and a programming model. Map and reduce are the two important tasks for the map reduce algorithm. A set of data will be converted into another set of data by using map and the every element will be broken down into tuples. The tuples can be key and value parts. The task of a reduce takes the output from a map as an input and combines the data's tuples as a small er set of tuples. The task of a reduce is performed when the map completes its job.

B. Advantages of Map Reduce

The map reduce is easy to scale the data processing over many multiple computing nodes. Mappers and reducers are the term which used for data processing primitives. In the concept of map reduce model, it will be a non-trivial process when the decomposition of data processing application takes place in Mappers and reducers. The configuration changes when we write an application in the map reduce form and scaling the application over hundreds or thousands of machines in a cluster. This is the process which attracts many users' to use this model.

VI. CONCLUSION

For analyzing the large amount of data there should be some technical advancement. To analyzing the huge data need of some scientific potential. In large variety of application domains the process is cost effective and faster method should be implemented.

REFERENCES

[1] S.Vikram Phaneendra, E.Madhusudhan Reddy,"Big Data-solutions for RDBMS problems- A survey", In 12thIEEE/ IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[2] Zujie Ren; Jian Wan; Weisong Shi; Xianghua Xu; Min Zhou, "Workload Analysis, Implications, and Optimization on a Production Hadoop Cluster: A Case Study on Taobao," in *Services Computing, IEEE Transactions on* , vol.7, no.2, pp.307-321, 2014.

[3] T. White, Hadoop-The Definitive Guide. Sebastopol, CA, USA: O'Reilly, 2009.

[4] Khan, M.; Yong Jin; Maozhen Li; Yang Xiang; Changjun Jiang, "Hadoop Performance Modeling for Job Estimation and Resource Provisioning," in *Parallel and Distributed Systems, IEEE Transactions on* , vol.27, no.2, pp.441-454, Feb. 1 2016

[5] Aveksa Inc. (2013). Ensuring "Big Data" Security with Identity and Access Management. Waltham, MA: Aveksa.

[6] Hewlett-Packard Development Company. (2012). Big Security for Big Data. L.P.: Hewlett-Packard Development Company.

[7] Kaisler, S., Armour, F., Espinosa, J. A., Money, W. (2013). Big Data: Issues and Challenges Moving Forward. International Confrence on System Sciences (pp. 995-1004). Hawaii: IEEE Computer Soceity.

[8] Katal, A., Wazid, M., Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. IEEE, 404-409.

[9] Marr, B. (2013, November 13). The Awesome Ways Big Data is used Today to Change Our World. Retrieved November 14, 2013.

[10] Amit Verma et al.,"Cross Layer Feedback Design: Optimization For Energy Efficient Mobile Devices Protocols Stacks Over Wireless Sensor Networks", National Conference on Emerging Trends in Communication held at SVIET-BANUR(District Patiala, PUNJAB) pp. 90, on 20th -21st February, 2009.

[11] Yi Yao; Jianzhe Tai; Bo Sheng; Ningfang Mi, "LsPS: A Job Size-Based Scheduler for Efficient Task Assignments in Hadoop," in *Cloud Computing, IEEE Transactions on* , vol.3, no.4, pp.411-424, Oct.-Dec. 1 2015.

[12] Shang, Weiyi, et al. "Assisting developers of big data analytics applications when deploying on hadoop clouds." Proceedings of the 2013 International Conference on Software Engineering. IEEE Press, 2013.

[13] Gupta, Rajeev, Himanshu Gupta, and Mukesh Mohania. "Cloud computing and big data analytics: what is new from databases perspective?." Big Data Analytics. Springer Berlin Heidelberg, 2012. 42-61.

[14] Kambatla, Karthik, Abhinav Pathak, and Himabindu Pucha. "Towards Optimizing Hadoop Provisioning in the Cloud." HotCloud 9 (2009): 12.
[15] Shakil, Kashish Ara, and Mansaf Alam. "Data Management in Cloud Based Environment using k-Median Clustering Technique." IJCA Proceedings on 4th International IT Summit Confluence 2013-The Next Generation Information Technology Summit Confluence 2013 (2014): 8-13.
[16] Kashish Ara Shakil Mansaf Alam, A Decision Matrix and Monitoring based framework for infrastructure performance enhancement in cloud based environment, Advances in Engineering and Technology Series, Vol-7, Page 147-153, 2013, Elsevier.
[17] Yi Yao; Jianzhe Tai; Bo Sheng; Ningfang Mi, "LsPS: A Job Size-Based Scheduler for Efficient Task Assignments in Hadoop," in *Cloud Computing, IEEE Transactions on* , vol.3, no.4, pp.411-424, Oct.-Dec. 1 2015.
[18] Shang, Weiyi, et al. "Assisting developers of big data analytics applications when deploying on hadoop clouds." Proceedings of the 2013 International Conference on Software Engineering. IEEE Press, 2013.
[19] Gupta, Rajeev, Himanshu Gupta, and Mukesh Mohania. "Cloud computing and big data analytics: what is new from databases perspective?." Big Data Analytics. Springer Berlin Heidelberg, 2012. 42-61.
[20] Kambatla, Karthik, Abhinav Pathak, and Himabindu Pucha. "Towards Optimizing Hadoop Provisioning in the Cloud." HotCloud 9 (2009): 12.
[21] Shakil, Kashish Ara, and Mansaf Alam. "Data Management in Cloud Based Environment using k-Median Clustering Technique." IJCA Proceedings on 4th International IT Summit Confluence 2013-The Next Generation Information Technology Summit Confluence 2013 (2014): 8-13.
[22] Kashish Ara Shakil Mansaf Alam, A Decision Matrix and Monitoring based framework for infrastructure performance enhancement in cloud based environment, Advances in Engineering and Technology Series, Vol-7, Page 147-153, 2013.

