A Review on Cloud based Hadoop Environment

Ms. Vandana Vijay
Lectuter
S. S. Jain Subodh P.G. College, Jaipur
vandanavijay161978@gmail.com

ABSTRACT - Cloud Computing is emerging as a most recent computational paradigm. It is an internet based technology which enables any type of business and organizations to use highly sophisticated computer applications. Cloud computing assures to change the way people use computers for accessing, modifying or saving the personal and business information. Cloud computing has several benefits for end users and businesses (elasticity, self-service provisioning, pay-per-use). Despite of its advantages, it has many disadvantages outline low scalability factor, no support for stream data processing). Hadoop is an open-source software framework that stores the data and run applications on different machines. It can handle thousands of terabytes of data. Hadoop has a file system known as HDFS which is distributed in nature. It allows faster transfer of data amongst nodes. If a node gets fails then also it allows systems to continue operating which reduces the risks of a system failure. Hadoop framework needs to be implemented in cloud computing to overcome its drawbacks. This paper is structured as follows: Section 1 Introduction explain the use of Hadoop in Cloud environment. Section 2 describes Hadoop architecture (HDFS and MapReduce). Section 3 defines Cloud Computing and its service models (IaaS, PaaS, SaaS). Section 4 provides a literature review on research papers related to Cloud computing with Hadoop. Finally paper is concluded in Section 5.

KEYWORDS - Cloud computing, Hadoop, Hadoop Distributed File System, Hadoop Ecosystem, MapReduce.

1. INTRODUCTION

In current scenario of Internet age, almost all the companies have migrated their data as well as applications to the cloud due to the popularity of the Internet. Management of big distributed data like cloud is a big challenge. For processing such gigantic amount of data, the traditional methods of database management are not appropriate since these approaches fail to handle huge size of data. Hence, in order to handle such large volume of heterogeneous data, companies are now coming up with different alternatives. One of the widely accepted solutions is Hadoop. It is the open source implementation of MapReduce. MapReduce works on Hadoop Distributed File System (HDFS).

2. HADOOP

a) Hadoop

Hadoop is used by many companies and social sites (LinkedIn, Yahoo, IBM, and Twitter, Facebook). It is an open source framework. It is written in Java language. It has two important parts namely, HDFS and MapReduce. HDFS supports storage of data in distributed manner while MapReduce support processing of data in an efficient manner. Hadoop architecture components are shown in figure 1. It follows Master and Slave node concept. Masters node contain Name node, Secondary node and Job Tracker while slave node contain Data nodes and Task Tracker. Job Tracker imitate the new tasks and track them accordingly. Task Tracker manages data processing, collects result and reports the status to Job Tracker.

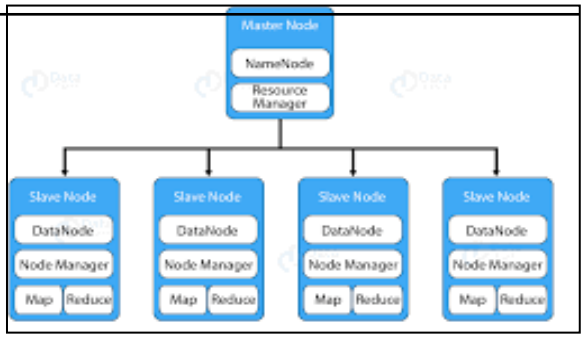

Fig 1: Hadoop Master/Slave architecture

b) Hadoop Distributed File System (HDFS)

It have a NameNode and multiple DataNodes. NameNode manages system metadata and act as a master server. It also maintains file system namespace and maps data from database to DataNode. The HDFS architecture is shown in figure 2.

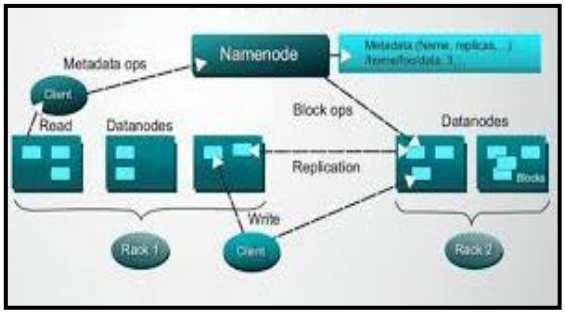

Fig 2: HDFS architecture

DataNode stores data and manages storage of physical nodes. It deals with read/write requests from users or clients. Through a registration process each DataNode in the cluster, makes itself available to the NameNode during startup. Every DataNode also informs NameNode about the blocks that it has possessed by sending a block report. Reports are sent periodically or after a fixed interval of time or whenever a change takes place in the block. Every DataNode sends heartbeat messages to the NameNode. These messages confirmed that the data node is still in operational state and that the data it is holding is available and safe. In case of DataNode failure error mechanisms comes in action so that the failure can be overcome and the block can be made available to other DataNode in the cluster. The typical size of the block is 128MB, but it can be change according to client. HDFS is fault tolerant in nature. Each block is stored in more than one data node in order to provide accessibility during system failure. A replication mechanism is used to implement the fault-tolerance feature.

c) MapReduce

It is a programming framework. It allows parallel and distributed processing on large data on multiple nodes. It consists of two important parts: A Job Tracker which is also known as master node. And multiple Task Trackers also called slave nodes. The Job Tracker accepts job requests, splits data input, defines tasks for the job, assign tasks across slaves nodes, monitors progress and handle failures. The Task Tracker executes tasks as ordered by the master node. The task can be executed either using a map function or reduce function. The MapReduce architecture is shown in figure 3.

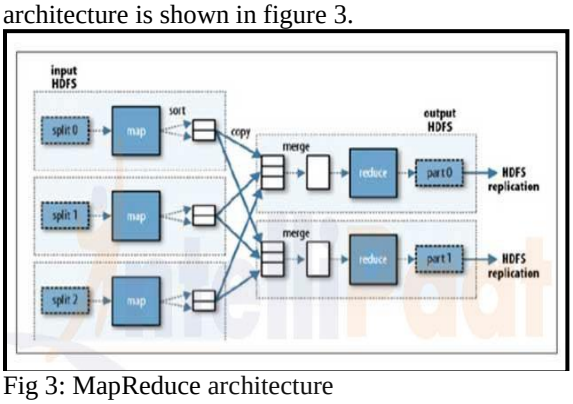

Fig 3: MapReduce architecture

The Map function receive data in Key, Value form . It returns a list of pairs in a different domain as shown in equation below:

Map Function (k1, v1) → list (k2, v2).

The Reduce function produces a collection of values in the same domain as shown in equation below:

Reduce Function (k2, list (v2)) → list (v3).

3. CLOUD COMPUTING

Cloud computing is one of the emerging technologies in the current time scenario. It has generated lot of interest and competition in the industry. Cloud is a network of servers that pools different resources. Cloud computing growth is reducing the cost of computation, application hosting, content storage and delivery in today's time. Cloud services can be grouped into categories depending upon either the type of service being provided or on the basis of location as shown in figure 4. According to Almorsy et al. (2016), the three basic service models are 1. Infrastructure-as-a-service (IaaS): In this service, the cloud providers deliver computation resources, network and storage as an internet-based service. 2. Platform-as-a-service (PaaS): In this service, the cloud providers deliver platforms, tools and other business services. 3. Software-as-a-service (SaaS): The cloud providers deliver applications hosted on the cloud infrastructure as internet-based service for end users, without requiring installing the applications on the customers' local machines. As per Zhang, (2010) applications can be deploy on Public, Private or Hybrid clouds. Public clouds offer their resources to the general public. Private clouds are used by a single organization. A hybrid cloud is a combination of both, public and private that tries to overcome the limitations of both.

Why Hadoop in the Cloud?

Hadoop clusters are run in the cloud due to following reasons listed in the table 1.

1	Lack of space	If the customer require Hadoop clusters, but don't have space for servers.
2	Flexibility	Everything is controlled through cloud provider APIs and web consoles.
3	Speed of change	It is faster to launch new cloud instances or allocate new database servers than to purchase, unpack, rack, and configure physical computers.
4	Lower risk	In the cloud, customers can easily get how many resources they use, so there is little risk of under commitment or over commitment. If some resource malfunctions, thatresources get discarded and new one are allocated.
5	Worldwide availability	Cloud providers have data centers all around the world.

Table 1: Hadoop Cluster in the Cloud

4. LITERATURE REVIEW

Malhotra et al. (2018) proposed GENMR model. It converts RDBMS queries to Map Reduce codes. It can effectively process data at Cloud repositories to overcome the limitations of existing RDBMS system. **Tripathi et al. (2018)** developed a cloud enabled hadoop framework. It combines cloud technology with the conventional hadoop framework to support the spatial big data solutions. **Bashir (2017)** provides review and analysis of MapReduce in cloud computing. Theyconcludes that cloud MapReduce has high scalability and also simplifies the large-scale data computation. **Alam&Shakil (2016)** proposed Hadoop based workflow for handling big data. Huge amount of data as well as big data can be managing in very easy ways in less amount of time. In the research work it was found that the average processing time is very less while processing in the cloud environment. **Ikhlaq&Keswani (2016)** Review Big Data methods and approaches of cloud computing implementation addressed. **Patil et al. (2016)**proposed Secured Hadoop as a service. It process big data on cloud with security and also provide hassle free platform which keeps away users from Hadoop configurations and gives out Hadoop service as web application service. Ansari et al. (2015) proposed Data Cleaning mechanism in Hadoop, Push Model and caching. The Data cleaning clears the already present memory content. It increases execution process. The Push Model enables the job tracker to push the heart beat to the task tracker in order to work directly. **Dash & Panda (2014)** propose a platform which integrates the Cloud, Big Data, NoSQL, Hadoop and analytic tools to efficiently capture, store and analyse complex datasets. **Voruganti (2014)** implements MapReduce through two components, JobTracker and TaskTrackers. **Gupta &Saxena (2014)** proposed big data implementation using Hadoop. It is the most required technology for Cloud Computing. They provide set up Hadoop cluster backed by HDFS running on ubuntu operating system. **Lu et al. (2012)** describes three important parts of Hadoop, HDFS, MapReduce andHBase. Hadoop shows good performance in dealing with large data sets concurrently, but there are still some limitations like failure of NameNode, HDFS small files, Job Tracker overload.

5. CONCLUSION

In this survey paper, the role of Hadoop in context of Cloud computations is investigated. The key issues, including Hadoop architecture, HDFS, MapReduce and Cloud Computing environment is described. Hadoop is widely used for big data processing in cloud platforms. It hides the complexity of parallel execution across hundreds of servers in a cloud environment. It can process terabytes of data. This paper concludes that, on cloud platform, Hadoop has been proven to be a useful tool for distributing the processing over as many processors as possible. It ensures a powerful, robust and fault tolerant system that can be used to deploy huge data set processing. Hadoop is the first choice for cloud computations. But further research must be done in near future to increase its efficiency, so that maximum utilization can be made from it.

6. REFERENCES

[1] Alam, M., & Shakil, K. A. (2016). "Big Data Analytics in Cloud environment using Hadoop", arXiv preprint arXiv:1610.04572.

[2] Gupta, N., & Saxena, K. (2014). "Cloud computing techniques for big data and hadoopimplementation", International journal of Engineering Research & Technology, *3*(4), 722-726.

[3] Ikhlaq, S., & Keswani, B. (2016),"Computation of Big Data in Hadoop and Cloud Environment", IOSR Journal of Engineering, *6*(1), 31-39.

[4] Patil, A. U., Patil, R. U., Pande, A. P., & Patil, B. S. "Secured Hadoop as A Service Based on Infrastructure Cloud Computing Environment"

[5] Tripathi, A. K., Agrawal, S., & Gupta, R. D. (2018). a Comparative Analysis of Conventional Hadoop with Proposed Cloud Enabled Hadoop Framework for Spatial Big Data Processing. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, *45*, 425-430.

[6] Bashir, B. (2017), "An Approach of MapReduce Programming Model For Cloud Computing", International Journal of Advanced Research in Computer Science, *8*(2)

[7] Ansari, S. M., Chepuri, S., & Wadhai, V. (2015),"Efficient Map Reduce Model with Hadoop Framework for Data Processing"

[8] Lu, H., Hai-Shan, C., & Ting-Ting, H. (2012, October),"Research on Hadoop cloud computing model and its applications", In *2012* third international conference on networking and distributed computing (pp. 59-63). IEEE.

[9] Malhotra, S., Doja, M. N., Alam, B., & Alam, M. (2018), "Generalized Query Processing Mechanism in Cloud Database Management System", In *Big* Data Analytics (pp. 641-648). Springer, Singapore.

[10] Kumari, S., 2014, "A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications4, 2250–3153

[11] Kaur, G., Kaur, M., 2015, "Review Paper On Big Data Using Hadoop", International Journal of Computer Engineering & Technology (IJCET) 6, 65–71.

[12] Peter Mell and Timothy Grance, "The NIST Definition of CloudComputing",(draft),http://www.nist.gov/custom

cf/get_pdf.cfm?pub_id=909616, accessed November 12, 2015

[13] Zhang, Q., Cheng, L., & Boutaba, R. (2010),"Cloud computing: state-of-the-art and research challenges", Journal of internet services and applications, *1*(1), 7-18.

[14] Almorsy, M., Grundy, J., & Müller, I. (2016),"An analysis of the cloud computing security problem", arXiv preprint arXiv:1609.01107.

[15] http://www.ijarse.com /images / fullpdf/ 1412339121 _346 _ IJARSE.pdf

[16] Dash, M., & Panda,R.N.(2014),"The Big Datain the Cloud and HADOOP Technology", International Journal of Advance Research In Science And Engineering. Vol. No.3, Special Issue (01), ISSN-2319-8354.

[17] Voruganti, S.(2014), "MapReduce a Programming Model for Cloud Computing Based on Hadoop Ecosystem"*I*nternational Journal of Computer Science and information Technologies, Vol.5(3),3794-3799.