International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Distributed Data Storage and Retrieval on Cloud by using Hadoop

N. Brahmanaidu[1], Shaik Riaz[2]

[1] M. Tech (CNS) Student, K L University, Guntur (A.P), India

[2]Assistant Professor, Department of CSE, K L University, Guntur (A.P), India

Abstract: ***Huge amount of data cannot be processed and store on local systems of the data is in tera byte. It is almost impossible to process and analysis the results from the data. So in this paper we are proposing a distributed storage mechanism on cloud which works based on hadoop mechanism.***

Keywords: Distributed cloud, Cloud computing, Hadoop, Hdfs, Map reduce mechanism

1. Introduction

Hadoop is changing the perception of handling Big Data especially the unstructured data. Let's know how Apache Hadoop software library, which is a framework, plays a vital role in handling Big Data. Apache Hadoop enables surplus data to be streamlined for any distributed processing system across clusters of computers using simple programming models. It truly is made to scale up from single servers to a large number of machines, each and every offering local computation, and storage space. Instead of depending on hardware to provide high-availability, the library itself is built to detect and handle breakdowns at the application layer, so providing an extremely available service along with a cluster of computers, as both versions might be vulnerable to failures. Huge amount of data cannot be processed and store on local systems of the data is in tera byte. It is almost impossible to process and analysis the results from the data. So in this paper we are proposing a distributed storage mechanism on cloud which works based on hadoop mechanism. In this paper we assume a storage mechanism on cloud the sender's subscriber to a cloud daas. The cloud data base is built up based on hadoop technologies. In this technology data is stored in distributed manner called hadoop distributed file system (HDFS). Data is replicated in different data nodes which can be access by name node using logs that are present in name node. Map reduce model is used to process data on cloud various types of analytic can be performed by using map reduce codes. Which can be return in java, hive or pig Latin.

2. Architecture and Working

2.1 Hadoop Distributed File System (HDFS)

HDFS is designed to run on commodity hardware. It stores large files typically in the range of gigabytes to terabytes across different machines. HDFS provides data awareness between task tracker and job tracker. The job tracker schedules map or reduce jobs to task trackers with awareness in the data location. This simplifies the process of data management. The two main parts of Hadoop are data processing framework and HDFS. HDFS is a rack aware file system to handle data effectively. HDFS implements a single-writer, multiple-reader model and supports operations to read, write, and delete files, and operations to create and delete directories networks.

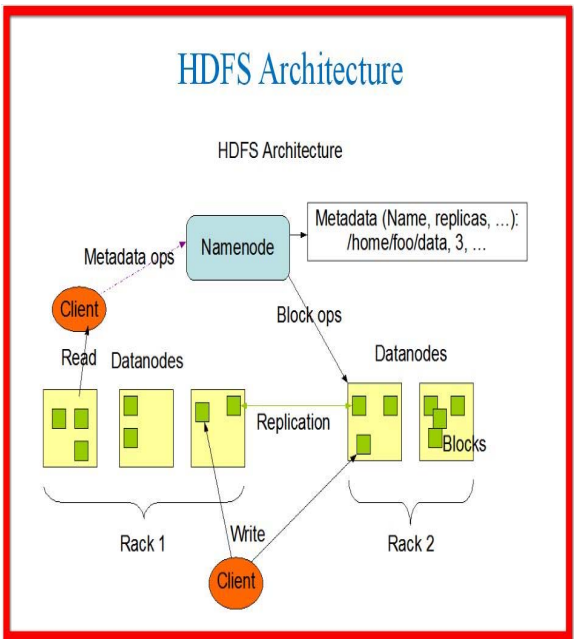

Figure 1: An overview of HDfs

3. Goals

- **Hardware Failure:** A core architectural goal of HDFS is detection of faults and quick, automatic recovery from them.
- **Need Streaming Data Access:** To run the application HDFS is designed more for batch processing rather than interactive use by users to streaming their data sets.
- **Portability Issues:** HDFS has been designed to be easily portable from one platform to another Across Heterogeneous Hardware and Software Platforms.

4. Data Processing Framework & Map Reduce

The data processing framework is the tool used to process the data and it is a Java based system known as Map Reduce.

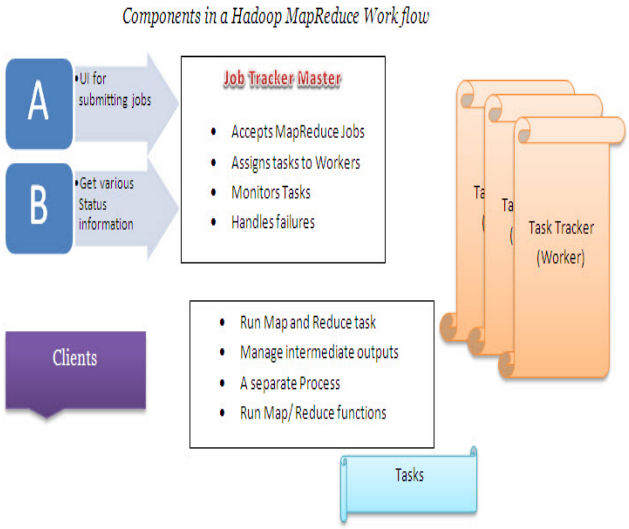

Figure 2: Map Reducing

5. Conclusion

We assume a s storage mechanism on cloud to the sender subscriber to a cloud Daas Hadoop enables surplus data to be streamlined for any distributed processing system across clusters of computers using simple programming models. It truly is made to scale up from single servers to a large number of machines, each and every offering local computation, and storage space. Instead of depending on hardware to provide high-availability, the library itself is built to detect and handle breakdowns at the application layer, so providing an extremely available service along with a cluster of computers, as both versions might be vulnerable to failures.

6. Future Work

We can upgrade this system to spark which is 100 times faster than hadoop**.**This module described the Map Reduce execution platform at the heart of the Hadoop system. By using Map Reduce, a high degree of parallelism can be achieved by applications. The Map Reduce framework provides a high degree of fault tolerance for applications running on it by limiting the communication which can occur between nodes, and requiring applications to be written in a "dataflow-centric" manner.

References

[1] http://hadoop.apache.org/hdfs/http://hadoop.apache.org/common/docs/current/hdfs_design.html
[2] Catanzaro, B., N. Sundaram, and K. Keutzer, \A MapReduce framework for pro- gramming graphics processors," in Workshop on Software Tools for MultiCore Sys- tems, 2008.
[3] Dean J. and S. Ghemawat, \MapReduce: Simpli ed Data Processing on Large Clusters," OSDI 2004.
[4] DeWitt, D. J., E. Robinson, S. Shankar, E. Paulson, J. Naughton, A. Krioukov, and
[5] J.Royalty, \Clustera: An Integrated Computation and Data Management System," VLDB 2008.
[6] Pike, R., S. Dorward, R. Griesemer, and S. Quinlan, \Interpreting the Data: Par- allel Analysis with Sawzall," Scienti c Programming 13(4), 2005.
[7] Scalable Scienti c Computing Group, University of Waterloo: http://www.math.uwaterloo.ca/groups/SSC/software/cloud, Retrieved date: Sep. 27, 2009.
[8] Soror, A., U. F. Minhas, A. Aboulnaga, K. Salem, P. Kokosielis, and S. Kamath, \Automatic Virtual Machine Con guration for Database Workloads," SIGMOD 2008.
[9] Yang, H. C., A. Dasdan, R.-L. Hsiao and D. S. Parker. \Map-reduce-merge: sim- pli ed relational data processing on large clusters," SIGMOD 2007.
[10] Zhang, C., H. De Sterck. \CloudWF: A Computational Work ow System for Clouds Based on Hadoop," The First International Conference on Cloud Computing, Beijing, China, 2009.