ISSN No. 0976-5697

Volume 6, No. 6, July-August 2015

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Large Scale Data Processing using HBASE Cloud Architecture

Ishwinder Kaur Sandhu
Department of Computer Science and Technology
ITM University
Gurgaon, Haryana

Neha Singh
Department of Computer Science and Technology
ITM University
Gurgaon, Haryana

Abstract: - In the present era, large amount of data is generated due to various Media. The rate of data generation is very high. This data is known to be "Big data". Big data mostly includes images and videos. Organizing Big data has turned difficult due to varying types of data and its increasing volume. This paper targets to provide image and video processing using cloud computing and analysis of big data using map reduce algorithm. To fulfil the needs a big data processing engine known as Hbase is used. Therefore, it is of a great provocation to handle large amount of data and also to improve its effectiveness using HDFS(Hadoop Distributed File System) and Hbase to store data.

Keywords: Cloud Computing, Map Reduce, Hbase, HDFS.

I. INTRODUCTION

As we are entering the new era of big data generation, where every second a large amount of data are generated. Instances such as digital processing, Internet, mobile devices and different types of sensors usually leads to the generation of Big data. Images and videos are the main source of these new generated data. An extensible computing power and practicing of data mining and also the capability for recognizing patterns is required for analysis of big data. It is magnified in processing images since the video and image processing algorithms turned much complex, demanding much computation potential.

In Image processing input is an image, for example a photograph or video frame whereas the output of image processing can be an image or a set of attributes similar to the image. Most of the image-processing techniques involve treating the image as 2-D signal and applying standard signal-processing techniques to it. The amalgamation of image processing with the cloud provide us with high-efficiency and performance image processing research background. This model is further substituted within a cloud computing framework.

Cloud not only provides the researchers with a large amount of storage and computation, but also it provide an ubiquitous environment to exchange knowledge and various algorithms.

II. RELATED WORK

HBase platform is used to process images parallely. For implementing algorithms in image processing our solution put forwards a Pass including the use of different languages. This proves one major differentiation among our work and others. HIPI (Hadoop Image Processing Interface)[3] is among the similar work.

Indifference to our work, to overcome the limitations in managing multiple image files in Hbase, Hadoop Image Processing Interface creates an interface for merging these files into a single large file. In this interface HipiImageBundle is used as an input. Maintenance needed to handle image storage requires no additional programming overhead. This is due to the reason that image storage is kept transparent to the users.

There exists a Remote sensing image analyzing system which uses the concept of Map reduce algorithm and further provides a programming module for image processing with its framework. This concept also uses the image as an input in Hbase. This concept is a fully Java-based framework. Distributed processing of video databases are performed parallely using the concept Parallel image database processing using map reduce in a cloud environment. Multiple series of video frames are stored using video. There exists a programming language named Ruby, used as a Mapper and thus runs on Hbase platform with streaming mode[2]. Therefore, the platform came out to be more flexible as it supports multiple languages. The workability of implementing Map Reduce model is considered by Large-scale Image Processing using Map Reduce[4].

III. PROPOSED WORK

HBase Cloud Architecture- Several clusters are built together in the HBase cloud computing infrastructure. Apache cloud stack[6] which is an open source software for creating, managing and deploying infrastructure as a service (iaas) to farm the VM's (virtual machine) and to provide platform as a service (paas).Storing and processing of big data takes place in HBase distributed.

There are three major components in the infrastructure: (1) To build a service portal for cloud computing large number of VM's are used. (2) High level computation and processing of big data is attained by implementing high performance Hbase clusters. (3) Access and storage of data is supported by data storage in cloud.

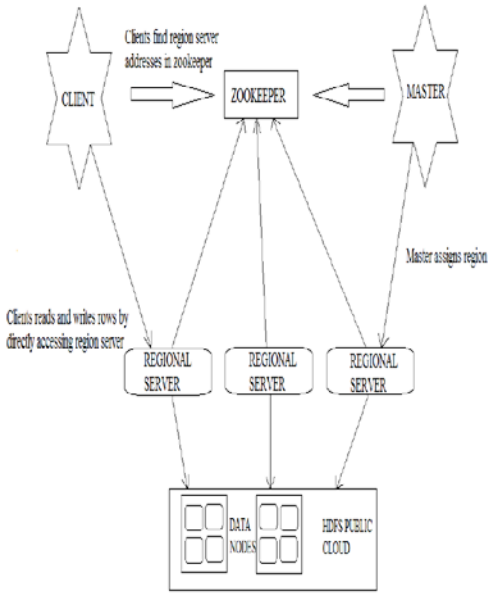

Fig 1. HBase Cloud Architecture

HBase Architecture- In HBase, tables are split into regions and are served by the region servers. These regions are vertically divided into column families named as "Stores". Stores are saved as files in HDFS[1][5]. Architecture of HBase is shown in FIG 2. HBase consists of three main components: the client library, a master server, and region servers. Region servers can be added or removed as per requirement. The master server: (1) Allocates regions to the region servers with the help of Apache Keeper (2) Manages load balancing across the region servers of the appropriate region.

It put off the busy servers and shifts the regions to less occupied servers, (3) Preserves the current state of the cluster by negotiating the load balancing, (4) In charged for schema changes and other metadata operations.(5) Regions: Regions are nothing but tables that are split up and spread across the region servers.

The region servers have regions that - (1) Communicate with the client and handles data-related operations, (2) Handles read and write requests for all the regions, (3) Decide the size of the region by following the region size thresholds. Region keeper: (1) Region keeper is an open-source project that provides services like maintaining configuration information, naming, providing distributed synchronization, etc., (2) Region keeper has ephemeral nodes representing different region servers. Master servers use these nodes to locate available servers, (3) Inclusion to availability, the nodes are also used to track server failures, (4) Clients communicate with region servers via keeper, (5) In pseudo and standalone modes, HBase itself will take care of keeper.

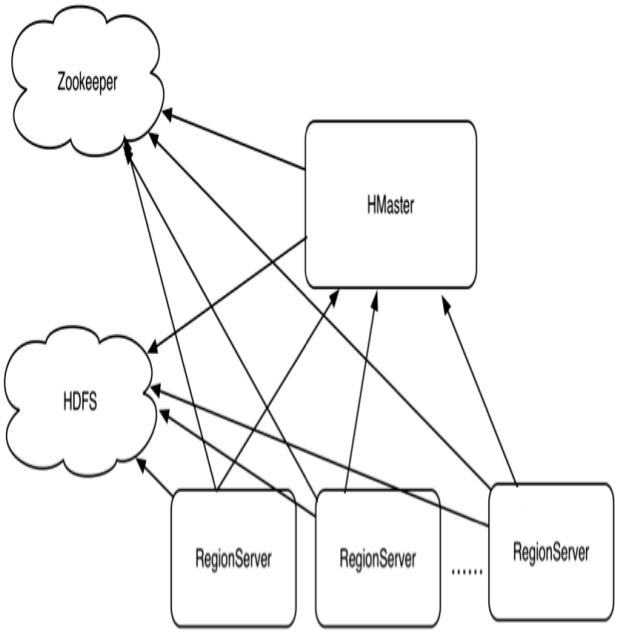

Fig 2. HBase Architecture

Zookeeper:
Zookeeper gets the details of region server. Information such as how many regions is present and which particular region server are accessing data nodes. Maintenance of Metadata is avoided by decreasing the overhead on top of Hadoop.

IV. CONCLUSION

Therefore the Hbase cloud architecture provides an open source real time processing of Big data using HDFS resulting in high performance and scalability. The most captivating feature is the ability to write user code that can generate files in Hbase's own format that can then be passed to the region servers, dodging the write path with minimal effect on abeyancy.

V. ACKNOWLEDGEMENT

The author wish to thank Asst.Prof. Neha Singh, ITM University, Gurgaon, India for providing requisite facilities.

VI. REFERENCES

[1] Barrachina Duque Arantxa, O'Driscoll Aisling, "A big data methodology for categorizing technical support requests using Hadoop and Mahout", Journal of Big data 2014.

[2] Harter Tyler, Borthakur Dhruba, Dong Siying , Aiyer Amitanand, Tang Liyin, Arpaci-Dusseau C. Andrea , Arpaci-Dusseau H.Remzi, "Analysis of HDFS Under HBase: A Facebook Messages Case Study", University of Wisconsin, Madison.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up Robust Features. In Computer Vision- ECCV 2006, pages 404-4016, Springer 2006.

[4] K. K MunetoYamamoto, "Parallel image database processing with Map Reduce and performance Evaluation in Pseudo Distributed Mode", International Journal of Electronic Commerce Studies, vol. 3, no. 2, 2012.

[5] "Hadoop Introduction", www.tutorialspoint.com/hadoop/

[6] "Apache cloud stack website", http://cloudstack.apache.org