Int.J.HIT.TRANSC:ECCN.Vol.3: Issue 2A(1) (2017) Page 62-66 ISSN: 0973-6875

Available Online at www.hithaldia.in/locate/ECCN

ORIGINAL CONTRIBUTION

Survey Related to HADOOP and Application Oriented Approach

Moumita Mantri*, Pranab Goswami and Manasija Bhattacharya

Depertment of Information Technology, Haldia Institute of Technology, Haldia, India

(Received Date: 20th August, 2017; Acceptance Date: 30th September, 2017)

ABSTRACT

In the present scenario storing massive amount of data has been a big problem for normal database system. The search engines, social media website etc. receives a large amount of data every second from corner of the world. So large amounts of storage space are required to store the data sent by the client, which is not possible by the normal database system. Clustering is the solution to this problem of big data. In order to create a cluster some modern technique such as distributed storage and distributed computing are needed. And to implement distributed storage and distributed computing technologies like Hadoop is used. Here the concept of cloud computing is used to provide a unique service named "Hadoop as a service". This is a unique service apart from SaaS, PaaS, Staas, and IaaS.

Key words: Big data; Cluster; Hadoop; Cloud computing;

1. INTRODUCTION

Hadoop [12] is an open source software framework for storing data and running application on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. Hadoop is a product of Apache which is built on JAVA. It works on HDFS [3] (Hadoop Distrubuted File System) and Mapper Reducer [2]. Here we are working on Hadoop over cloud, so we need to know what cloud is. The cloud is just a metaphor for the Internet and other networks. The word "cloud" often refers to the internet and more precisely to some datacenter full of servers that is connected to the Internet. However, the term "cloud computing" [4] refers to the software and services that have enabled the internet cloud to become so prominent in everyday life. A cloud can be a wide area network (WAN) like the public Internet or a private, national or global network. The term can also refer to a local area network (LAN) within an organization. Here cloud computing is used for proper utilization of hardware. Many companies provide SaaS (Software as a Service) but none of the companies provides cluster as a service. Even AWS (Amazon Web Services) which is a giant in the technology market at cloud computing can provide Hadoop as a Service. According to statistics about 80% of internet traffic gets transferred from AWS.

2. NEED OF HADOOP

In the world of growing internet technology massive amount of data gets posted on the internet every second. As the data that are posted on the internet stays on a server machine and everyone can access them any time when needed. But with the increase of data space of the server machines are reducing at a large rate. In order to overcome this problem of storage of Big Data cluster is needed.

So a question arises that what big data actually is? Big data [1] is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structure of our database architectures.

Every day 2.5 quintillion bytes of data are created-so that 90% of the data in the world today has been created in the last two years alone.

According to IBM, 80% of data captured today is unstructured, from sensors used to gather climate information, posts to social media sites,

*Corresponding Author: moumita.mantri@gmail.com

digital pictures and videos, purchase transaction records, and cell phone GPS signals, to name a few. All of this unstructured data is Big Data.

Unstructured data comes from information that is not organized or easily interpreted by traditional database or data models, and typically, it's text-heavy. Metadata, Twitter tweets, and other social media posts are good example of unstructured data.

2.1 Statistics on Big Data

On the most visited websites over the internet we collected some stats which show how the amount of data increases over these websites in the recent years:-

Google

In 2012, Google received over 2 million search queries per minute. Fast forward to 2014 ant that number has more than doubled. Today, Google receives over 4 million search queries per minute from the 2.4 billion strong global internet population.

Face book

The face book warehouse stores upwards of 300 PB of Hive data, with an incoming rate of about 600 TB. In the last year, the Face book warehouse has seen a 3x growth in the amount of data stored in it.

Mobile data traffic

Global mobile data traffic grew 74% in 2015.Global mobile data traffic reached 3.7 Exabyte's per month at the end of 2015, up from 2.1 Exabyte's per month at the end of 2014.

Twitter

Every second, on average, around 6,000 tweets on twitter which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year. The chart below shows the number of tweets per day throughout Twitter's history.

3. WHY DO INTEGRATE HADOOP WITH CLOUD?

There is a lot of hardware available in our locality such as in labs in the college or our personal laptops or PCs but we cannot make proper use of these hardware. If we use these hardware's to create cluster then we could use the hardware's in a proper manner.

In the public cloud we are sharing infrastructure with other people on the internet. Enterprises could also use and pay for the Hadoop Cluster [8].

4. IMPLEMENTATION OF HADOOP OVER CLOUD IN LINUX

Hadoop is based on the concepts of distributed storage and distributed computing [5]. In distributive storage multiple numbers of hard drives are merged together in order to make cluster. In distributive computing both the hard drive and RAM of the system are merged. Hadoop works on scale out process [9]i.e. all the systems are connected in a horizontal manner. This mechanism avoids single point failure.

4.1 How to setup multiple nodes Hadoop Cluster in context of Distributed Storage

Hadoop works on master slave model where the master is known as the **Name Node** and the slave is known as a **Data Node**.

STEP 1: Get the **Hadoop RPM** from the apache website

http://www.apache.org/dyn/clouser.cgi/hadoop/common/

And run the command--**#yum install hadoop **-- on the Red Hat Linux Terminal.

STEP 2: Get the Java RPM from Oracle Website

http://www.oracle.com/technetwork/java/javase/downloadl/jdk7-downloads-1880260.html

And run the command --**#yum install jdk**-- on the Red Hat Linux Terminal

STEP 3: Run the following commands on the Linux terminal to setup **HDFS** name node and data node

#vim/etc/hadoop/hdfs-site.xml & #vim/etc/hadoop/core-site.xml

The above command open the hdfs-site.xml page and core-site.xml in which we write the following code to configure name node and data node

hdfs-site.xml

Entry for name node

<configuration>
<property>
<name>dfs.name.dir</name>
</property>
</configuration>

core-site.xml

Entry for name & data node

<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://IP Of NameNode:9001</value>
</property>
</configuration>

Entry for data node

<configuration>
<property>
<name>dfs.data.dir</name>
</property>
</configuration>

STEP 4: Now format the name node

#hadoop namenode-format

STEP 5: Command to start name node

#hadoop-daemon.sh startnamenode

STEP 6: Command to start data node

#hadoop-daemon.sh startdatanode

As discussed earlier hadoop also works on map reduce. Distributed computing is introduced to configure the concept .

4.2 How to setup multiple Hadoop Cluster in context of Distributive Computing

Hadoop works on master slave models where the master is known as the Job Tracker and the slave is known as a Task Tracker [7].

STEP 1 and STEP 2 are same as describe in 3.1

STEP3: Run the following commands on the Linux terminal to setup Job Tracker and Task Tracker.

The above command opens the mapped-site.xml page

Mapped-site.xml

Entry for Job Tracker & Task Tracker

<configuration>

<property>

<name>mapped.job.tracker</name>

<value>IP Of Job Tracker:9002</value>

</property>

</configuration>

STEP4: Command to start the Job Tracker

#hadoop-daemon.sh start jobtracker

STEP5: Command to start the Task Tracker

#hadoop-daemon.sh start tasktracker

4.3 How to transfer Hadoop Cluster over cloud

As many of the Cloud companies virtually transfers IaaS (Infrastructure as a Service) in which they transfer full flash operating system to their clients by using protocol such as SSH(Secured Sell), RDP(Remote Desktop Protocol). For example AWS (Amazon Web Services) transfer Linux OS through SSH protocol & Windows OS through RDP. But none of these companies use a protocol called VNC(Virtual Network Computing)[10][11] in which a graphical desktop sharing system that uses the Remote Frame Buffer protocol(RFB) to remotely control another computer. So here VNC & SSH is used to transfer full flash Linux OS in which Hadoop cluster is pre-installed.

We can also see Hadoop using web portal. In the browser address bar type the IP address of name node. There are two management ports 50070 & 50030 where we see how HDFS and Mapped Reduce Works.

Clients can see how many node & task trackers are in the cluster. They can even put their files and how the cluster works.

5. HOW ACTUALLY CLUSTER WORKS

Many people do not know that where the data is stored i.e. in Name Node, job tracker. Data node. Task tracker [6]. But actually when clients puts file on cluster file is not stored in the name node and job tracker. Client put the file on job tracker, job tracker transfer the information of the file to name node and name node said job tracker to perform the work. Then job tracker said task tracker to perform the work instead of doing itself. After performing the task, Task Tracker saves the data to data node. Data node only stored Meta data. By default 3 replication is made on cluster, but we can change it. And by default block size is 64 Mb, but we can also change it.

6. CONCLUSION

Hadoop cab be created by using local resources cluster which is being much efficient. Like collage's has lots of systems which can use to create a cluster. And make easy to use Hadoop cluster to everyone by using cloud computing concept. One good thing about this paper is according to requirement modification on cluster size is possible. This technology help to reduce the infrastructure cost. With the help of Hadoop Cluster we can run the large program. Fast processing is also done by using Hadoop.

References

[1] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A Review Paper on Big Data and Hadoop",in International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-3153.

[2] P. Sudha, Dr. R. Gunavathi2 "A Survey Paper on Map Reduce in Big Data" in International Journal of Science and Research (IJSR), Volume 5 Issue 9, September 2016, ISSN (Online): 2319-7064.

[3] Sneha D.Borkar , Prof.Chaitali S.Surtakar "A REVIEW PAPER ON THE HADOOP DISTRIBUTED FILE SYSTEM" in International Journal of Research In Science & Engineering, Volume: 1 Special Issue: 1, e-ISSN: 2394-8299, p-ISSN: 2394-8280.

[4] Santosh Kumar and R. H. Goudar "Cloud Computing – Research Issues, Challenges, Architecture, Platforms and Applications: A Survey",in International Journal of Future Computer and Communication, Vol. 1, No. 4, December 2012

[5] Dr.A.P.Mittal, Dr.Vanita Jain and Tanuj Ahuja "Google File System and Hadoop Distributed File System- An Analogy" in International Journal of Innovations & Advancement in Computer Science, Volume 4, Special Issue March 2015, ISSN 2347 – 8616

[6] Madhavi Vaidya, Madhavi Vaidya "Critical Study of Hadoop Implementation and Performance Issues" Research In IT: Exploring the Horizon in patkar-varde college on 30th and 31st August 2013 Patkar College

[7. Ruchi Mittal , Ruhi Bagga in International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

[8] Garvit Bansal, Anshul Gupta, Utkarsh Pyne , Manish Singhal, Subhasis Banerjee "A Framework for Performance Analysis and Tuning in Hadoop Based Clusters" in 15th International Conference on Distributed Computing and Networking(ICDCN)

[9] Zhuozhao Li and Haiying Shen "Performance Measurement on Scale-up and Scale-out Hadoop with Remote and Local File Systems " in 2016 IEEE 9th International Conference on Cloud Computing (CLOUD), Electronic ISSN: 2159-6190

10] Md. Sanaullah Baig1, Rajasekar M.2 and Balaji P. "Virtual Network Computing Based Remote Desktop Access" in International Journal of Computer Science and Telecommunications, Volume 3, Issue 5, May 2012

[11] Priyadarshani Raskar , Sejal Patel , Pragati Badhe, Prof. Archana Lomte "Virtual Network Computing- A Technique to Control Android Phones Remotely" in International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 2 February, 2014 Page No. 3991-3995

[12] James Horey, Edmon Begoli, Raghul Gunasekaran, Seung-Hwan Lim, and James Nutaro "Big Data Platforms as a Service: Challenges and Approach" in Hotcloud12-final61