ISSN: 2393-8528

IJICSE
Open Access

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 1 Issue 3; Page No.16-19

Review on hadoopm technology using cloud Computing

Mriglekha Chakraborty [1], Harshita Tiwari [2], Neetu sisodia [3]

[1]Research scholar Computer Science & Engineering Technology, Jyoti Vidyapeeth Women's, University, Jaipur, Rajasthan, India

Mriglekha9@gmail.com

[2]Research scholar Computer Science & Engineering Technology, Jyoti Vidyapeeth Women's, University, Jaipur, Rajasthan, India

harshitaswety@gmail.com

[3]Research scholar Computer Science & Engineering Technology, Jyoti Vidyapeeth Women's, University, Jaipur, Rajasthan, India

neetusisodia1992 @gmail.com

ARTICLE INFO

Received 12 Oct. 2014
Accepted 30 Nov. 2014

Corresponding Author:

Mriglekha Chakraborty

1 Research scholar Computer Science & Engineering Technology, Jyoti Vidyapeeth Women's, University, Jaipur, Rajasthan, India

Email: Mriglekha9@gmail.com

Key words: Hadoop,Big data, map reduce , Hdfs, yarn.

ABSTRACT

How astonishing now a days how the popular companies like Amazon,Gmail,Facebook,Enomaly,GoGrid,Microsoft,Netsuit,Rackspace are storing and handling such a large amount of data . Its hadoop in cloud computing that accumulates petabytes and terabyte or digital information ,cloud computing has evolved with great interest in the present used technologies with took their further subparts as hadoop and big data along with it .Now hadoop ,it is open source framework software for large scale analysis and storage of various clusters and data sets which is designed to establish on low cost hardware .There are various segments of hadoop as big data that is any collection of data sets that is so complex that it is hard to manage using the traditional processing method, hadoop distributed file system that has the capability of storing large amount of the data on single server with thousand machine or more ,eco system ,map reduce is a runtime infrastructure that decreases larger threads into smaller form as much as possible. So now we are taking cloud computing as a mass connection of the interconnected system in a private or public network.

INTRODUCTION

Man is still the most extraordinary computer .The exigency of hadoop in cloud computing emerges from where cloud computing is host based on internet service i.e. any service on demand or data stored on in internet anything out of it would not be saved in any electronic system but only on cloud. Now ,there are three services that subdivide the functionality of cloud computing : SaaS(Software as a service) renders the service to the client as their specific requirement but no need of spoil money on making license or any server so the cost is itself minimized and multiple number of end user can be provided service on single specimen that runs on cloud . ,Paas(Platform as a service)a user is provides an encapsulated environment to built high level application that runs on the providers infrastructure ,it also renders the application server and virtually made OS and network, configuration setting and many more facility to help along with the manageability and scalability requirement of application ,IaaS(Infrastructure as a service)is a provisional model that allows a very important service of storage ,database management and such demanded services are offered. The service provider is responsible for all storing, running and maintaining all the equipments as well as owns these equipments. Hadoop is an apache made computing cluster designed specially designed for storing ,maintaining data and Is used in cloud computing environment .Hadoop manages the unstructures data ,big data (collection of cluster of data sets).Then mapreduce processes these data sets on cloud .

II. HISTORY

Hadoop was created by Doug Cutting and Mike Cafarella in 2005.Cutting was working with Yahoo that time named it after his son's toy elephant It was evolved to support the Nutch search engine project

.Nutch is an open source web search engine based on java and Lucene. The goal 0f nutch is web scaled, crawler based search it had properties like distributed when necessary, sort /merge based processing .In the duration 2006-2008 hadoop project split out of nutch .Finally it made its appearance in 2008.

III. WHAT IS HADOOP?

- Hadoop is java based program that supports the processing of large data set i.e. big data in a computing environment to rapidly transfer data and to continue the performed operation and the main strength lies in its ability to scale across thousands of commodity server that don't share memory or disk space .For an instance hadoop can be considered as an ecosystem – it is a combination of many heterogeneous components all of which runs on a single platform .The major breakthrough advantage of hadoop is that organization and businesses can now find value in data that was considered useless recently .There are two different keys functional components within this ecosystem. They are**:**

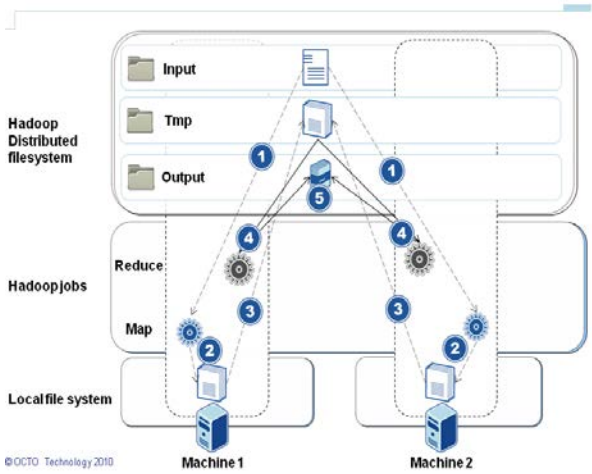

Figure1: Hadoop architecture

IV. BIG DATA

Big data is any encompassing term in which enlarged set of data that is hard and complex to deal with using relational database that was used traditionally. The objectives include analysis ,capture ,curation ,search ,sharing, storing visualization and privacy violation etc. Scientists were regularly encountering problem handling big data in many areas including meteorology, genomics, connectomics and biological and environmental research .Big data is when the data itself is a part of problem.The following characteristics are suitable for big data

V. CLOUD COMPUTING

VI. HADOOP USING CLOUD COMPUTING:

1. Hadoop render the extent data to reliable, efficient, scalable, economical using very simple java programming interface.Map reduce provide simple interface to complicated and distributed computing in build in cloud computing.
2. We can easily understand how to design java application on cloud using hadoop like mapreduce structure ,hdfs, hod, understating the hadoop api ,configure a hadoop cluster on linux, administer of hadoop environment ,design principal for cloud on hadoop.
3. When high traffic among the nodes on network and collision is create so resolve this problem we use hadoop to create ourselves isolate network like VLAN.so that we have to configure or install the hadoop cluster on cloud.
4. Hadoop distribution (cloudera CDH,IBM biglnsights,mapreduce,Hotonworks) can be lauched and run on public cloud like AWS,Rackspace,MS Azure,IBM smartCloud etc.,which offer infrastructure as a service(IAAS) in a public cloud.
5. The map reduce as a service perform If u want to quick run mapreduce job in your workload so the Amazon's EMR(elastic mapreduce)perform without having to install hadoop environment on cloud. Mainly ,the hadoop programming expertise within your organization…
6. To prevent and control over the hadoop infrastructure so create the private cloud, private cloud render platform as a service layer that offer pre build pattern for deploying hadoop cluster, the big reason to create private cloud i.e. deployment would be around data security , access control and visibility.
7. The requirement of data security SOX, PII, HIPPA for public cloud before moving data into hadoop cluster.

There are many project or new product to running hadoop in cloud computing, to mapreduce enhancement over it's base platform like qubole, mirantis , VMware and Rackspace all announced product or service offering with cloud.

8. Hadoop involve batch oriented system including new incoming data come on to server so it's analytic into the scheduling job hadoop cluster on physical machine is always used or not in consuming power, leased space etc.

VI. TERMINOLOGY USED:

A. PIG :Pig was initially developed by yahoo for the users of hadoop basically for decreasing the mapping process and reduce programs .They eat almost anything ,Hence this name is given . The pig is a combination of two components: first the language and secondly the environment where the piglatin is executed .There is a certain process in Piglatin –**Load** in which the data is used from the Hdfs **,secondly** the process continues in which further set **transformations** take place where a covered and are translated into a set of mapped and reducer tasks ,finally the data is **dumb** on the screen ,& **stored** in the file somewhere.

B. **ZOOKEEPER:**

It is an open source Apache project and provide centralizes architecture for maintaining configuring information. It is top level sub project of hadoop

because render high availability through redundant service. Zookeeper data structure supports hierarchical structure like file system. The basic working t\of zookeeper is providing group services and distributed synchronization .There are many contributed tools used in zookeeper .Some of them are:

- Zkconf –generates the configuration for zookeeper ensemble.
- Zktop- monitors zookeeper in real-time.
- Zkexamples- Phunt's redundant example of useful bits of zookeeper ephemera.
- SPM for Zookeeper -performance monitor and alerting of zookeeper.
- Zk-smokest- laterncytest a zookeeper using zkpython

C. HIVE **:**

Hive was created when there came a problem with the running of SQL queries ,Then came HQL(hive squery language).and it is limited in the commands it understands and are subdivided into hive services in mapreduce jobs and can be executed in hadoop clusters .Hive queries can run in many ways .Some of them are :

- Command line interface called hive shell from a Java Database Connectivity
- Open database connectivity(ODBC) for leveraging the HIVE drivers , that's called HIVE thrift Client .But Hive thrift client is much like a database that is generally installed by the clients computer in the middle tier of the three tier of the and communicate with the Hive applications on the server and then it can be used on any services which is written in the languages like C++ ,Java ,Python ,Ruby .The look of Hive is just similar to the traditinal SQL query .

D. HDFS(HADOOP DITRIBUTED FILE SYSTEM)

Hadoop distributed file system is content of apache hadoop project…hdfs grant fault torrent file system stand for rapidly transfer data between system nodes so that hadoop system will be continued if system fails Hadoop framework provides hadoop distributed file system for map reducing to analyze in large dataset...hdfs include master slave architecture that mean one system manages more system in which each cluster contain single name node that organize operation of file system and manage the data nodes in other words the hdfs stores data in dedicated server is called name nodes and application data store on other server is called data node ….hdfs procedure high performance fetching data into the hadoop cluster..it organize big data analytical application .hdfs deal 40 petabyte data of yahoo enterprise.

E. YARN

another resources locator assigns memory to CPU and helps in storage of the present running applications in hadoop clusters .The other generations which were discovered in past years could Yet not run other applications mapreduction applications .Yarn enhances this by enabling other frameworks like spark which helps open further possibilities

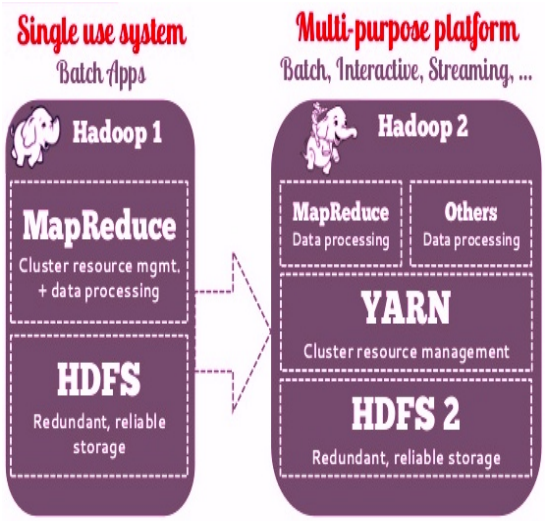

Figure 2: components of hadoop.

VII. CONCLUSION

The brief usage of the data and storing the data as per as requirement and accessing them according to the requirement and today's daily usage the evolution of hadoop and simultaneously the major use of big data also came into existence .Therefore after this various applications came into existence which enhanced the characteristics and properties of bigdata.

REFERENCES

1. Study of scientific data process on cloud using hadoop by Chen Zhang, Hans De Sterck ,Ashraf Aboulnaga, Haig Djambazian .
2. Scale-up vs Scale-out for Hadoop: Time to rethink by Raja Appuswamy, Christos Gkantsidis, Dushyanth Narayanan, Orion Hodson, and Antony Rowstron
3. Scaling big data with hadoop and solr by Hrishikesh Vijay Karambelkar.
4. hadoop operations and cluster management cookbook by Shumin guo.
5. Hadoop Real World Solution Cookbook by Jonathan R.Owens,Brain Femiano,Jon Lentz.

IX. FUTURE SCOPE:

As the factor is known that now Big data and hadoop are very popular technologies .Now a days hadoop is used in every technology like Mahout etc helping its data storage and other machine learning tasks .Going further ,In the coming time there is the fact that there is definitely much more scope of hadoop and its related technologies in future and its related career opportunity for the software developers and that would be there deed that how do they help them to get a bulk amount of cheque This can also rule as a major platform to start a fruitful career. When overtime data will increase the scope for hadoop will increase much more .Starting from GB ,next it TB then it will increase PB of data ,that is already been generated. Engineers

with knowledge of Java can undertake projects on hadoop .Giant merchants are investing on hadoop .In India generation of data is increasing day by day .So, its good till any next big thing replaces it .

X. EVALUATION OF HADOOP -

There are many companies which use hadoop's mapreduce in various streams and processes like some of them are

Yahoo!-

There are more than 100000 CPUs 40000 computers running hadoop . One of Hadoop's largest applications is Yahoo.It owns the biggest clusters i.e. 4500 nodes (2*4 cpu boxes w 4*1 TB dsik & 16 Gb RAM).yahoo web search runs on more that 10000 core linux clusters that is used majorly .

Twitter-

Apache hadoop is mainly used to store and process tweets ,log files and many types of generated data all across twitter in compressed LZO files .Both Scala and java

Are used to access hadoop 's Mapreduction APIs .Commiters are employeed on Apache Pig ,Apache Avro ,Apache hive and Apache Cassandra and contributed to open source also .

Adobe-

These are used in various ways to internal use and processing.About 30 nodes are running through HDFS ,the range starts from 5 to 14 nodes running HDFS and hadoop in cluster form ,and plans for 80 nodes clusters are coming .This production unit is runnig from oct 2008.

Facebook-

In the year 2013 june 13 it was announced that the data of facebook has grown till 100 PB.On November 8 2012 they said that the data is growing one PB per day in the warehouse .They were claimed to have the largest hadoop cluster in 2010.i.e 21 PB of total storage