Volume 5, Issue 9, September 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Study of Hadoop-Related Tools and Techniques

Deepika P*, Anantha Raman G R
Department of CSE, ACE,
India

Abstract: ***Today's world cannot be even imagined without storage. The amount of data which is being generated by us is growing day by day. Traditional approaches fails to manage such large volume of data. Big Data is a collection of large volume of structured, unstructured and semi-structured data that are clustered together. Various Big Data tools and techniques are already in use for managing that data efficiently and effectively. Among the widely available tools and techniques, Hadoop plays a major role in the IT market. It is a framework for managing the massive amount of heterogeneous data. Many leading giants such as IBM, Microsoft, Yahoo, Amazon are working with this technology. Almost 100% of the other big giants would move on to Hadoop in the upcoming years. This paper is a study of various Hadoop-related tools and techniques.***

Keywords: Big Data, Hadoop, cluster, heterogeneous, HDFS, MapReduce

I. INTRODUCTION

In the earlier days, traditional approaches were used to organize and store the data. An organization will have a separate computer to store the data. This approach is well suited when the amount of data is less. But the data which is being generated in the recent years is really huge (i.e. petabytes of data). It is a tedious task to manage this huge volume of data with the traditional approaches. And so, the concept of Big Data came into picture.

Big Data concept is used in many real world applications. Many marriage matching sites are now using Big Data tools to find out the best match, it is also greatly helpful in understanding the needs of the customers and targeting the customers, it is already been used in sick baby unit to predict the infection of a sick baby 24 hours before any physical symptom occurs, it is also used in operating the Google's self driving car. Big Data is an emerging and interesting technology where everybody can contribute to its development.

Big Data is nothing but a collection of large amount of data that are clustered together. Big Data includes structured, unstructured and semi-structured data. Big Data includes three main characteristics. Velocity denotes how fast the data is getting transferred to and fro, Volume defines the quantity of data it has to manage and Variety means to manage different varieties of data. The organizations such as Amazon, Google, and Facebook are using Big Data to manage their transactions and also to target their customers.

Managing this massive amount of data becomes simple with the Big Data tools and techniques. There are several different types of tools and techniques available for managing the Big Data. Few of them are Hadoop, Cassandra, Storm, MongoDB, Riak, Hive, Pig. Most of these technologies and tools are based on Java only with few exceptions. Knowing Core Java is an added advantage to work with this environment. Big Data is about managing and organizing the data. So the knowledge of Data Warehousing is also highly recommended. Most of the tools and technologies of Big Data are Open Source which makes it easier for the developer to code and test.

Hadoop is the leading tool for managing the Big Data. Apache Hadoop is the open source project of the Apache software foundation. It runs the applications under the Map-Reduce algorithm where a huge task is broken in small pieces and distributed to process in parallel on different nodes. A recent prediction says that almost 100% of the leading giants would adopt Hadoop in the upcoming years. It is also predicted that Hadoop will yield an annual growth rate of 58% by 2022.

A study on various Hadoop-related tools and techniques will give an insight of how this massive amount of data is getting managed, stored and organized.

II. RELATED WORK

A. What is Hadoop?

Hadoop is the software framework which was developed by Apache Software Foundation. Hadoop framework is written in Java with the intent to handle large clusters of data. Hadoop can manage a huge volume (i.e. petabytes) of data. Hadoop runs the task under the MapReduce algorithm and Hadoop Distributed Fie System (HDFS). MapReduce is nothing but a model for processing a task by splitting a huge task into several others sub-task and is distributed to different nodes to process in parallel. HDFS, as the name suggests, it is the distributed file system for storing the data in the clusters. Thus by using MapReduce and HDFS, Hadoop manages the massive amount of data. Other than HDFS and

MapReduce, there are several other Hadoop related tools and technologies. Notable ones are Ambari, Avro, Cascading, Chukwa, Fiume, HBase, Hive, Hivemail, Mahout, Oozie, Pig, Sqoop, Spark, Tez and Zookeeper. Fig. 1. depicts the Hadoop architecture.

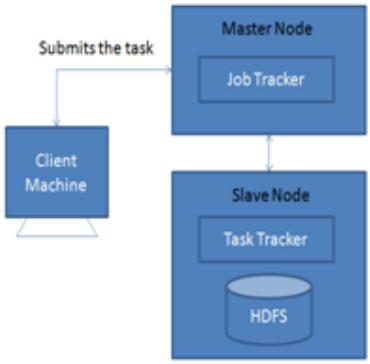

Fig. 1. Hadoop Architecture

B. Why Hadoop?

Working with Hadoop is quiet simple with the knowledge of Core Java and few related concepts of Data Warehousing. Hadoop library has been designed in such a way that it automatically identifies and handles failure which makes it more efficient in the way it does not need to depend on any hardware platform to detect failures. Any server can be removed from the cluster or added to the cluster dynamically and Hadoop continues with its operation. Hadoop is designed in such a way that it can work with any platform.

C. Who are the users of Hadoop?

Amazon uses Hadoop to build their product search indices and also to process their millions of sessions. Adobe uses it internal data storage and processing. Cloudspace is using Hadoop for their client projects. Hadoop is been used by eBay for their search optimization and research. Facebook uses Hadoop for machine learning and to store their copies of internal log. IIT, Hyderabad uses Hadoop for Information Retrieval and Extraction research projects. Last.fm uses it for charts calculation and dataset merging. Twitter is also using it to manage the data that is been generated daily in their website. Apart from these big players, IBM, Rackspace, The New York Times, Linkedln, University of Freiburg, University of Glasgow and lot more are using Hadoop.

D. Who are Hadoop Vendors?

Everybody around the world is gaining knowledge about Big Data. Many vendors today support Hadoop to a greater extent. Notable ones are AWS, Cloudera, Microsoft, MapR, Oracle and IBM.

III. TOOLS AND TECHNIQUES RELATED TO HADOOP

A. Ambari

Ambari is the project developed by the Apache Software foundation to support Hadoop by making its management simpler by maintaining the Hadoop Clusters. It provides an environment which is easy to maintain using its RESTful APIs. Using Ambari, the system adminisrators can easily manage, provision and monitor a Hadoop cluster. Various Operating Systems which supports Ambari are OS X, Windows and Linux. Fig. 2. depicts the Ambari architecture

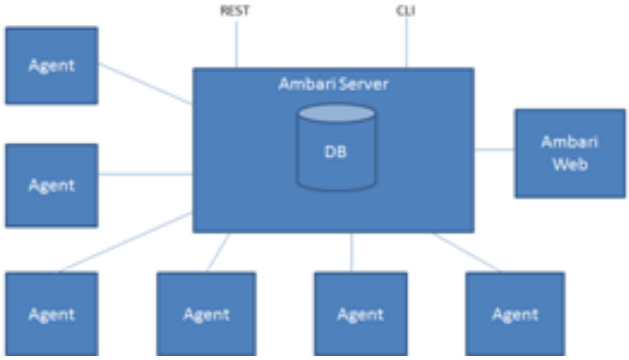

Fig. 2. Ambari Architecture

B. HBase

Apache HBase is an open source, distributed database which was built on top of HDFS. It aims at storing millions and billions of data with support to fault-tolerance. HBase is similar to that of the Google's Bigtable. Although it supports the storage of large databases, it is not a direct replacement of the SQL database. It's performance is increasing in the recent days and now it supports the messaging platform of Facebook. HBase is based on Java and is OS independent. The architecture of HBase is depicted in Fig. 3.

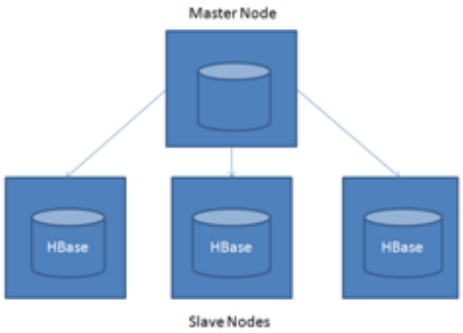

Fig. 3. HBase Architecture

C. Tajo

Tajo is the datawarehouse which is distributed for managing the Big Data. It was developed by Apache. Initially, it uses HDFS as the storage layer and the storage gets completed with its own query engine. This query engine will allow direct control of execution and also data flow. This is because it has various evaluation strategies, SQL standards and also optimization methods. The supported OS of Tajo are Linux and Mac. Fig. 4. depicts the Tajo architecture.

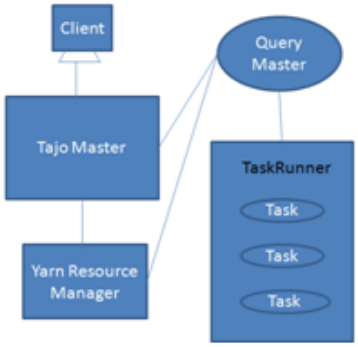

Fig. 4. Tajo Architecture

D. MapReduce

MapReduce is a framework for handling the huge volume of datasets in a cluster. MapReduce was initially found by Google. It consists of a Map() phase and a Reduce() phase. The Map() phase will perform the sorting and filtering operations and the Reduce() phase will then perform a summary operation on the sorted data. It is said to be the heart of Hadoop. It is OS independent. Fig. 5. is the architecture of MapReduce.

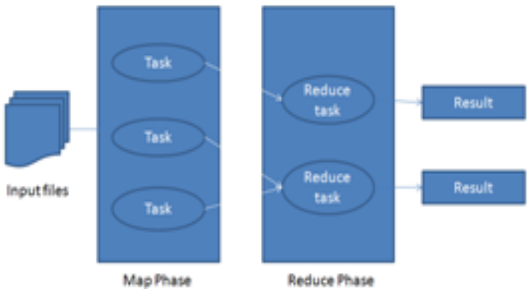

Fig. 5. MapReduce Architecture

E. HDFS

HDFS (Hadoop Distributed File System) is the distributed file system which is based on Java for the storage of large datasets. HDFS was initially developed by Apache and now it is the sub-project of Apache Hadoop. HDFS provides high fault-tolerance when compared with the other distributed file systems. It also provides high throughput, scalability and can be deployed on hardware of low-cost. Windows, Linux and OS X are the Operating system which supports HDFS. HDFS architecture is shown in Fig. 6.

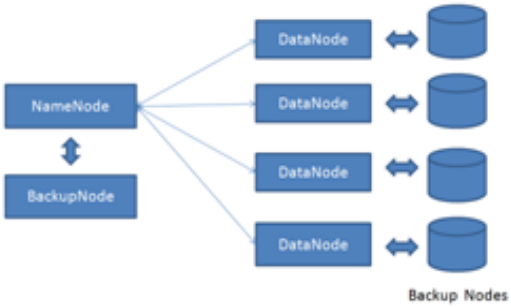

Fig. 6. HDFS Architecture

F. Hive

Hive is the open source project which was developed by Facebook. Same like Tajo, Hive is also the data warehouse for managing large datasets which is based on Hadoop. It uses HiveQL which is similar to SQL language, for managing the datasets. Being inconvenient to use HiveQL, it also allows the designers to use MapReduce concept. It is OS independent. The architecture of Hive is depicted in Fig. 7.

Fig. 7. Hive Architecture

IV. COMPARISON OF THE TOOLS AND TECHNIQUES

A. Ambari and Mapreduce

TABLE I COMPARING AMBARI AND MAPREDUCE

	Ambari	MapReduce
Use	For managing Hadoop Clusters	For managing Hadoop Clusters
Mechanism	Uses RESTful APIs	Uses MapReduce algorithm
Security	Through Kerberos	Through HDFS
Advantage	Simplicity, centralized management	Scalability, fault-tolerant
Performance	Uses Dashboard for monitoring performance	Performance increases by renting more nodes
OS	OS X, Windows, Linux	OS Independent

B. HBase and HDFS

TABLE II COMPARING HBASE AND HDFS

	HBase	HDFS
Use	Storage of large datasets	Storage of large datasets
Fault-tolerant	High	High when compared with HBase
Security	Thrift Gateway	Authorization mechanisms
Advantage	Automatic failover support	Low cost, high bandwidth
Disadvantage	Not suitable for small datasets	Not suitable for small datasets
OS	Independent	Windows, Linux, OS X

C. Tajo and Hive

TABLE IIII COMPARING TAJO AND HIVE

	Tajo	Hive
Use	Data warehouse for managing big data	Data warehouse for managing big data
Storage	Uses HDFS and its own query engine	Uses HiveQL
Speed	Greater speed	Lesser speed

Advantage	Greater connectivity to Java program	Good fit for organizations
Usage	Not widely used	Widely used
OS	Linux and MAC	Independent

V. RESULTS AND CONCLUSION

The data which is being generated everyday will definitely increase in the upcoming years but it has no hope of getting decreased. So the future market will have high impact on storage. When it comes to the word "storage", then obviously it is going to be Big Data. It is the leading technology were most of the organizations are working with, and many other organizations will adopt to it in few years. The future is going to be nothing without Big Data as it is the main technology used for storage. Knowing Big Data has another main important factor. It is the one which is going to be the "future of business". Learning the importance of Big Data is highly recommended for the survival in the IT market and also for employment. Big Data cannot be understood without Hadoop. It is the basis of storage where huge volume of clusters is managed.

Hadoop is not a single term. It has many inter-related technologies, terms and tools. This paper is just an introduction and comparative study of Hadoop-Related technologies and tools such as Ambari, HBase, Hive, MapReduce, HDFS and Tajo. This will motivate the beginners to learn more about this upcoming technology. The employers of an organization will have an added advantage by knowing it, as this is going to be the future of the IT market and it may also create employment for the fresher. Apart from all these, knowing a new technology is always a boon and not a bane in this competitive world.

REFERENCES

[1] http://hadoop.apache.org/
[2] Puneet Singh Duggal and Sanchita Paul, "Big Data Analysis: Challenges and Solutions," Nov. 2013
[3] http://datamation.com/
[4] http://3g.sina.com.cn/
[5] http://cubrid.org/
[6] Sangeeta Bansal and Dr.Ajay Rana, "Transitioning from Relational Databases to Big Data," in vol.4, Issue 1, Jan 2014
[7] Stephen Kaisler, Frank Armour, J. Alberto Espinosa and William Money, "Big Data: Issues and Challenges Moving Forward," 2012
[8] http://research.google.com/archive/mapreduce.html/
[9] Arkady Zaslavsky, Charith Perera and Dimitros Georgakopoulus, "Sensing as a service and Big Data,"
[10] Hao, Chen and Ying Qiao. " Research of Cloud Computing based on the Hadoop platform.", Chengdu, China: 2011, pp.181-184, 21-23 Oct 2011
[11] http://marktab.net/
[12] Brad Brown, Michael Chui, and James Manyika, "Are you ready for the era of 'big data'?, Oct 2011
[13] http://craighenderson.co.uk/
[14] www.ibm.com/software/data/infosphere/hadoop/
[15] https://ambari.apache.org/
[16] https://hive.apache.org/
[17] https://hbase.apache.org/
[18] http://hortonworks.com/
[19] www.pcworld.com/
[20] http://sethsiddharth.com/
[21] Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri, "Security issues associated with Big Data in Cloud Computing," vol.6, No.3, May 2014
[22] A. Katal, M. Wazid, R.H. Goudar, "Big Data: Issues, Challenges, tools and Good practices," Aug. 2013

