4
BIG DATA: TECHNOLOGICAL ADVANCEMENT IN THE FIELD OF DATA ORGANIZATION

Tabish Mufti
Department of Computer Science and System Studies Mewar University, Gangrar, Chittorgarh
Deepak Kumar
Amity Institute of Information Technology, Noida

Abstract
In this paper we will be discussing about Big data and what it is all about. Growing demand for online services is rapidly increasing. Now days large amount of data has been generated from various sources and this data need to be store and organised for proper usable. In Big Data the size of Data or File is at least minimum 1 Terabyte. So the Amount of time requires opening such files to open it is 10 minute or 5 minutes respectively and storage of such file take lot of amount of time. These are the type of problems which arises here while opening such a huge amount of data. The configuration of a machine also plays a vital role in processing such data. So just think how many challenges you are going to face while opening such files leave a part that you want to do analysis on it. In such situations the concepts which comes to a picture is Hadoop.

Keywords: Hadoop, Dig data, Terabytes, Storage, Access, Data Size, Analysis, Configuration, Apache Hadoop, Map Reduce, HDFS

Introduction

In Today's IT world, every day large amount of data is being generated from various sources. It is said that 90% of world data is being generated in last two years and this accerated trend will be continue. All this data is coming from smart phones, social networking, trading Platforms, machines and remain other sources and the questions is that weather we are going to use it or not. In early 2000 the companies like Google was having large quantity of data was simply too large to pump through single database. Such companies paid large amount to their respective database vendors. This problem was been faced by large IT companies. Company like Google first started storing all this data in Distributed File system and started store this data on distributed file system .The problem which Google starting facing was that data updating was being done manually.

One of the largest technological challenges in software systems research today is to provide mechanisms for storage, manipulation, and information retrieval on large amounts of data. Web services and social media produce together an impressive amount of data, reaching the scale of petabytes daily (Facebook, 2012).

These data may contain valuable information, which sometimes is not properly explored by existing systems. Most of this data is stored in a non-structured manner, using different languages and formats, which, in many cases, are incompatible (Bakshi, 2012; Stonebraker et al., 2010).

Take, for instance, Facebook, which initially used relational database management systems (DBMS) to store its data. Due to the increasingly large volume of information generated on a daily basis (from a 15TB dataset in 2007 to a 700TB dataset in 2010) (Thusoo et al., 2010), the use of such infrastructure became impracticable. Especially because, most of its data is unstructured, consisting of logs, posts, photos, and pictures. One of the Face book's largest cluster holds more than 100 PB of data, processing more than 60,000 queries a day (Facebook, 2012). Having achieved in September 2012 more than 1 billion active users, Face book may be considered one of the largest and most valuable social networks.

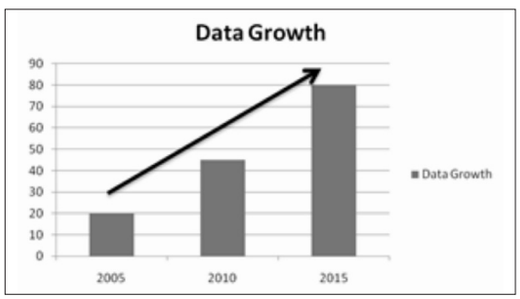
Companies holding large amounts of user data started to be evaluated not just by their applications but also by their datasets, specially the information that can be retrieved from them. Big companies like Google, Facebook and Yahoo have an aggregate value not only for their provided services but also for the huge amount of information kept. This information can be used for numerous future applications, which may allow, for example, personalized relationships with users.

Data storing was done manually in Distributed file system so Google started looking for mechanism through which data can be updated automatically that's why Google written a white paper known as Google Distributed File system on the other hand a person named as Dough Cutting who is came from search background was also working on the same frame work. He invented losing search engine which is used in java environment. Dough Cutting read the white paper and started implementation of that in 2002 and 2003 .On the other hand yahoo was keeping eye on this situation. Yahoo hired Dough Cutting and told him to fully implementation .He took two years to do complete implementation. Yahoo

implemented Hadoop as a Framework in its environment. Later Yahoo Gave Framework to apache and it became full fledge

Big Data

Big Data is a collection of huge amount of data that is large in nature and complex in processing using on hand database management tools or traditional data processing application like dbms, rdbms or Sql file. Now the bar graph show below shows the percentage growth in terms of data generation.



As we are talking about big data there are various types of data

- Structured Data
- Unstructured Data
- Semi-structured Data

Structured Data

Structured data is information, usually text files, displayed in titled columns and rows which can easily be ordered and processed by data mining tools. This could be visualized as a perfectly organized filing cabinet where everything is identified, labelled and easy to access. Most organizations are likely to be familiar with this form of data and already using it effectively

Unstructured Data

Unstructured data is raw and unorganized and organizations store it all. Ideally, all of this information would be converted into structured data however; this would be costly and time consuming. Also, not all types of unstructured data can easily be converted into a structured model. For example, an email holds information such as the time sent, subject, and sender, but the content of the message is not so easily broken down and categorized. This can introduce some compatibility issues with the structure of a relational database system

Semi-structured Data

Semi-structured data is data that has not been organized into a specialized repository, such as a database, but that nevertheless has associated information, such as metadata, that makes it more amenable to processing than data

II. Hadoop Map Reduce Methodology

Hadoop helps in saving large files with the help of distributed file system and helps to do Analytics on the top of it with the help of Google map reduce algorithm implementation

Map Reduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner. Now moving few years back we came across that somewhere around 70's and 80's we all were using transaction systems where we were using traditional databases we had fix , schemas fix with tables fix rows at relational database level. People know that what a kind of data store is and what kind of application they have to build on top of it. This went on 2000 and 2002. Application like facebook and tweeter came and started generating unstructured data. In structured we know what the table will look like, how rows will look like and so on. In unstructured data the data is categories in various formats consisting of audios, images, textual. So the data was unstructured and size was very large and it was growing.

IBM also developed his own definition of BIG data. Define the concept into FOUR V's different parameters.

- Volume
- Varity
- Velocity
- veracity

Volume: define the scale of data .IBM says size of data is increasing and it will be double with respect to time

Variety: Means various forms of data. Which include audios, videos, post, and images?

Velocity: The speed at which data is generating is huge and need to do speedy analytics

Veracity: how you know that data is genuine and results are right which are generated from the current data

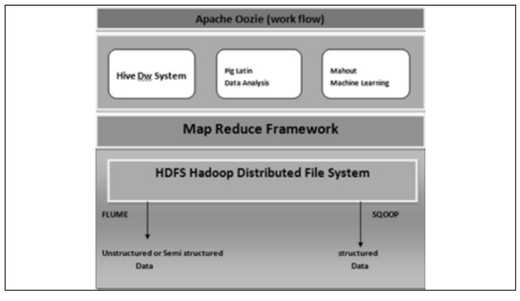
Limitation of Exiting Solutions

- Fixed of Scheme
- Cost
- Saving huge files and accessing them
- Perform Analytics

III.Hadoop Ecosystem

A Hadoop ecosystem not only consist of Distributed file system and Map reducing Algorithm but also helps in saving files. Distributed file system helps to save such big file and map reducing algorithm helps to do analytics on top of it. All that done in very speedy way. It is cost effective and come free of cost

Hadoop can be defined as Open Source Management system



Hadoop Architecture

Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware. There are mainly five building blocks inside this runtime environment (from bottom to top): the cluster is the set of host machines (nodes).

Components of Hadoop Architecture

Apache Oozie

Apache Oozie is the tool in which all sort of programs can be pipelined in a desired order to work in Hadoop's distributed environment. Oozie also provides a mechanism to run the job at a given schedule.

Hive DW System

Hive DW components is somewhere equal to data warehouse but a difference with it uses portion of data whereas data whereas Hive Dw uses entire set of data

Pig Latin Data Analysis

Apache Pig is an abstraction over Map Reduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows. Pig is generally used with Hadoop. we can perform all the data manipulation operations in Hadoop using Pig.

Mahout Machine Learning

It is a Artificial intelligent kind of search engine.

Apache Mahout is an open source project that is primarily used in producing scalable machine learning algorithms. It works in the following manner suppose you are search online to buy new gadget say laptop you browse various web sites to see the product now without purchasing you close website and next day to browse again for any other work so while working online you will see the previous day data which you browsed might be suggesting you. Now Mahout comes to a picture it influence you to buy that product which you have not done in previous day.

Now these three chucks of block built a layer of abstraction on top of Mapreduce and HDFS

HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

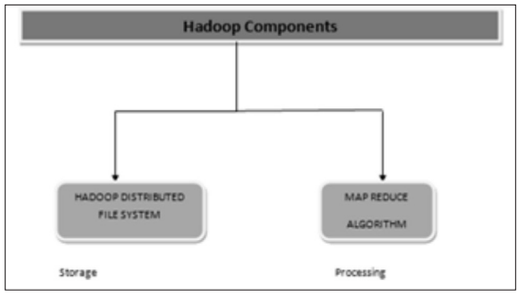
Features of HDFS

- It is suitable for the distributed storage and processing.
- Hadoop provides a command interface to interact with HDFS.
- The built-in servers of name node and datanode help users to easily check the status of cluster.
- Streaming access to file system data.
- HDFS provides file permissions and authentication.

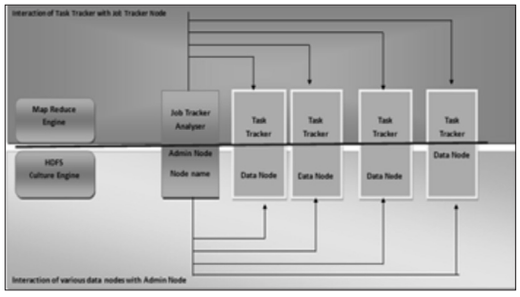
Flume: It is a framework which provides to store unstructured data which is continuously generated from Applications LinkedIn, facebook in form of videos, audios, posts etc. So Flume pulls data from that system and stores it in Hadoop.

SQOOP: is a Framework. When you are doing big data implementation by using hadoop .Most of the time your transactional data is in old system and you want to pull that data from old system to hadoop distributed file system. Sqoop helps to do that. It proves very simple interphase it pulls data from RDBMS and stores it in hadoop implementation.

These two frameworks not only help to pull the data from systems but also push the data back to the systems.



HDFS helps to store whereas Mapreduce helps to performs analytics on data. Below Diagrams shows how various components interact logically with one another.



HDFS with Map Reduce

In a typical collection of machine there will be one heavy duty server called as Admin Node and list of commodity machines spread across clusters. Admin node is divided into two parts

Job tracker and Name node.

Name node directs correlates HDFS part of it. Name node on admin node manages everything from sorting managing and accessing these huge Big Data files. Now these files are directly correlated with data nodes with respect to Admin node each node. So the other part of Admin node technology consists of Job tracker. Whose job is to manage map reduce algorithms implementations, they both runs at the admin node levels. A typical Admin machine is a heavy duty machine with 64 GB of RAM machine.

Now moving towards commodity machines these are those machines which are used in save data on distributed systems. These types of machines are divided into two type

Data Node and Task tracker. Data node is responsible for distributed file system aspects whereas task tracker is use for map reducing algorithm implementation.

Hadoop solves limitation of existing solutions

Now lets analysis how Hadoop solve the limitation of the existing solutions like RDBMS based data ware housing solutions

- It helps to save large files : large files whose size is in Zeta byte can be saved using HDFS
- Maps reduce Framework helps to do Analytics on the entire set of data. It's not like data warehouse where a part of data is being use. Map Reduce provide analytics on whole data.
- Hadoop takes less time to do Analytics while in data warehousing it takes lot of time.
- Scale out Architecture: since it is a Scale out architecture you can change the amount of time take to do processing. Suppose in a cluster you have got one name node and 30 commodity machines and your 1 zeta byte taking 2 hours to perform analytics on top of it. If you increase number of commodity machines says 50 then the amount of time would be just half to perform analytics. So you keep on scaling out or increasing commodity machines the entire computational time will keep on decreasing that's how Hadoop solves this problem.

Big data is a huge complex set of data where the challenge is not only to save the data but also to read, browse and cache and do meaningful analytics on top

The existing solution like data warehousing built on top of RDBMS have challenge with respect to storing and accessing Big data.

Hadoop is an open frame work meant for analytics using HDFS as the storage mechanism and Map reduce as an analytic platform

Hadoop provide solution for storing data on HDFS and doing analytics using Map reduce on the entire set of data.

Conclusion

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. Hadoop is designed to run on cheap commodity hardware, It automatically handles data replication and node failure, It does the hard work – you can focus on processing data, Cost Saving and efficient and reliable data processing. It has many similarities with existing distributed file systems.

However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data.

HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is part of the Apache Hadoop Core project

References

1. *D.J. Abadi, Data management in the cloud: Limitations and opportunities, IEEE Data Engineering Bulletin 32 (1) (2009) 3–12*
2. *Davies, J., & Graff, M. (2005). Performance in e- Learning: Online participation and student grades. British Journal of Educational Technology, 36(4), 657–663*
3. *T. Menzies, T. Zimmermann, Software analytics: So what? IEEE Software 30 (4) (2013) 31–37*
4. *The Age of Big Data. Steve Lohr. New York Times, Feb 11,2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html*