Big Data Analysis using Hadoop

Archana Bopche
Assistant Professor
Department of Computer Engineering
Terna Engineering College, Navi Mumbai, India

ABSTRACT

In this paper a new set of approaches is present for analyzing dataset which is called as Big data. The big data is most wrongly interpreted term, when one talk about big data in terms of computers it is concept which explains about gathering, organizing, analyzing the data and from these steps to get information from this data. With this approach one is not able to do previously because there had challenges across one or more of the 3V's of bigdata.Firstly volume which means too big data second is verity which has too complex data and velocity which has fast data. With the help of Bigdata performance of organization is improved. As HADOOP is a prominent framework for big data implementation, so paper present overall architecture of Hadoop with details of its components.

General Terms

Big data, HDFS

Keywords

Big data.HDFS

1. INTRODUCTION

Companies from all over world have been using the data since a long time so that they can take better decision for betterment of companies' growth.

1.1 What is Big Data

Big data is a large amount of data which is difficult to store, process using the traditional database because data available has large amount of data, complex data and fast data. The big data is mostly unstructured data which is unlike of structured data which we access through RDBMS .This is one of the important reason why the concept of Big data was first introduced by Google,Facebook,Linkedin,ebay etc

1.2 Current Challenges in big data-

Sources of Bigdata

- **Science**-Data bases from astronomy,genomics,environmental data.
- **Humanities and Social Sciences**-Scanned books,historical documents,Social interaction data ,new technology like GPS.
- **Business and commerce**-Corporate sales, Stock market transaction,census,airline traffic,online shopping.
- **Entertainment**-Internet images,Hollywood movies,MP3 files
- **Medicines**-MRI and CT scans,patient records

Explosion in Quantity of Data-

- **Air Bus A380**
  1 billion line of code, 640TB per flight each engine generate 10 TB every 30 minutes
- **Twitter** Generates approximately 12 TB of data per day
- **New York Stock exchange** 1TB of data every day.
- **Storage capacity** has doubled roughly every 3years since the 1980s.
- Billions of Google queries per day.
- **Facebook** generates -500 TB/day
- **Mobile users in India**-9 Millions, highest growth rate in the world.

There are 4 categories of companies-

- **Class 1**-Who comes under the productization of big data Ex-Amazon,Facebook,Microsoft
- **Class 2**-Who has taken hadoop and productized it,Although hadoop is open source system but these companies has productized it.
  Ex-Coudera,MAPR
- **Class 3**-These companies have value added to hadoop.
- **Class 4**-These provides to services to class1 and class2 level companies.

1.3 Google's Technology to handle with the current problems

- **GPS**-Google file System,which is distributed file system.
- **Big Table** –It is a Kind of database ,but no sql ,no indexing no join can handle semi structured data.
- **Map Reduced Technology**-This is a data processing framework.
- **Lot of data analytical too**l-Other tools to make data analysis easy.

2. INTRODUCTION TO HADOOP

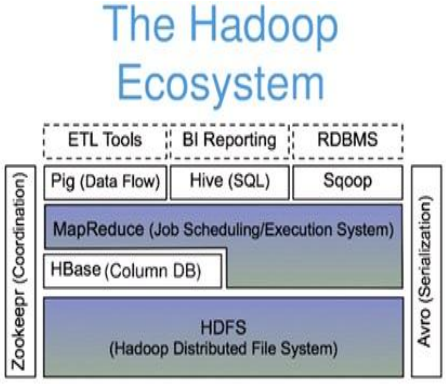

Fig-1

2.1 Introduction to Apache Hadoop

- It is the Open source software platform that lets one easily write and run applications that process huge amount of data. It includes
    - Hadoop Core(HDFS,Map Reduce)
    - Hadoop Ecosystem(Hbase,Pig,,Hive,Sqoop,Zookeeper)
    - Yahoo is the biggest contributor

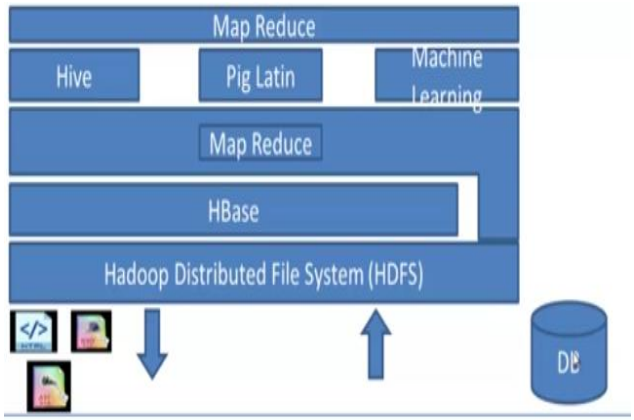

Fig 2

2.2 Hadoop Assumptions

It is written with large cluster of computer in mind and is built around the following assumptions:

- Hardware will fail
- Processing will be run in batches. Thus there is an emphasis on high throughput as opposed to low latency
- Applications that run on HDFS have large data sets. A typical file in HDFS is gigabyte to terabyte to peta byte in size.
- Moving computation is cheaper than Moving data
- Portability is important.

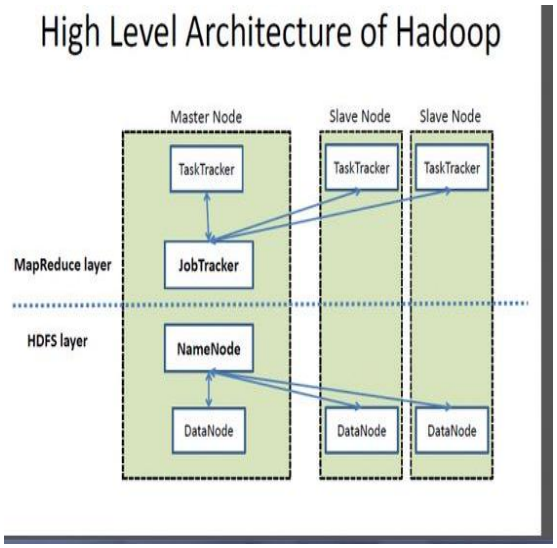

Fig 3

2.3 Hadoop Configuration files-

- Hadoop-env.sh
- Core-Site.xml
- Hdfs-site.xml
- Mapred-site.xml
- Hadoop-metrics-Properties
- Log4j.properties

2.4 Hadoop deployment mode-

- Standalone or local Mode-
    - No daemons running
    - Everything runs on a single JVM
    - Good for developing mode
- Pseudo-distributed mode
    - All daemons running on a single jvm
    - A cluster simulation on a single machine
- Fully distributed mode
    - Hadoop running on multiple machines on a cluster
    - Production environmet

2.5 Pseudo distributed mode-

- **Hdfs-site.xml**

```
<? Xml version=”1.0”?>
<!...hdfs.site.xml…>
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
<property>
<configuration>
```

- **Core-site-xml**

```
<? Xml version=”1.0”?>
<!—core site.xml…>
<propetry>
<name> fs default.name</name>
<value>hdfs ://localhost 8020</value>
</proetry>
</configuration>
```

3. HADOOP COMPONENTS COMPARISON

Table1: Hadoop Components Comparison

Name	Drizzle	HBase	Hive	MongoDB	Redis	Cassandra
Description	Here emphasis is on performance over compatibility which is MYSQL fork with a pluggable micro-kernel.	It is based on Apache Hadoop and based on big table which is wide column store	DataWarehouseSoftware for Querying and Managing Large Distributed Datasets, built on Hadoop	It is most popular document stores.	In-memory Database with configurable options which is performance vs. persistency	Wide-column store based on ideas of Big Table and Dynamo DB
Implementation language	C++	Java	Java	C++	C	Java
Database Model	RDBMS	Wide Column Store	Relational DBMS	Document Store	Key–Value Store	Wide Column Store
Consistency Concepts	-	Immediate Consistency	Eventual Consistency	Eventual Consistency ,Immediate Consistency	-	Eventual Consistency, Immediate Consistency
Concurrency	Yes	Yes	Yes	Yes	Yes	Yes
Replication Method	Master–Master Replication, Master–Slave Replication	Selected Replication factor	Selected Replication factor	Master–Slave Replication	Master–Slave Replication	Selected Replication factor
Durability	Yes	Yes	Yes	Yes	Yes	Yes

4. CONCLUSION

Hadoop had started in the year 2002 with the project name Apache Nutch.Hadoop developed by Dough Cutting.Hadoop was first inspired by papers published by Google outlining its approch to handling an avalanche of data.Hadoop ecosystem framework provides scalability to store large volume of data on commodity hardware. With hadoop no data is too big. Earlier the data which was considered use less, now organization can take help to take decision for future growth of organization. There is enough scope in big data field

because there are live project running on this for example Smart Refrizrator, Smart AC etc.There are large scope in machine learning based project is upcoming years where data plays vital role and in this regard hadoop may play major role.

5. REFERENCES

[1] https://gavinbadcock.files.wordpress.com/2013/02/hadoopschema1.jpgDing, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[2] JimmyLin"MapReduceIsGoodEnough?"The control project. IEEEComputer32(2013).

[3] Umasri.M.L, Shyamalagowri. D, Suresh Kumar. S "Mining Big Data: Current status and forecast to the future"Volume 4,Issue1,January2014ISSN:2277128X

[4] Albert Bifet "Mining Big Data In Real Time"Informatica 37(2013)15–20DEC2012

[5] Bernice Purcell "The emergence of "big data"technology and analytics"Journal of Technology Research 2013.

[6] Sameer Agarwal†,Barzan Mozafari X, Aurojit Panda†,Henry Milner †,Samuel Madden X, Ion Stoica "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data"Copyright©2013ìACM978-1-4503-19942/13/04

[7] Yingyi Bu_Bill Howe_Magdalena Balazinska_Michael D.Ernst "The HaLoop Approach to Large-Scale Iterative Data Analysis "VLDB 2010 paper "HaLoop: EfficientIterative Data Processing on Large Clusters.

[8] ShadiIbrahim⋆_HaiJin_LuLu"HandlingPartitioningSkewinMapReduceusingLEEN"

[9] https://www.pinterest.com/pin/454722893608174095/