Class Note:

Subject: Big Data Technologies

Faculty: V. Divya

Topic: Introduction to Hadoop. Introduction.

Unit No.: 2
Lecture No.: 1
Link to Session
Planner (SP) S.No. ..... of SP
Date Conducted:

Introduction:

- Hadoop is an open source software framework that is used for storing and processing language large amounts of data in a distributed computing environment.
- It is designed to handle big data and is based on MapReduce programming model, which allows for the parallel processing of large datasets.
- Its framework is based on Java programming with some native code in C and shell scripts.
- It has two components:
  - → HDFS (Hadoop Distributed File system)
  - → YARN (Yet Another Resource Negotiator.

History of Hadoop:

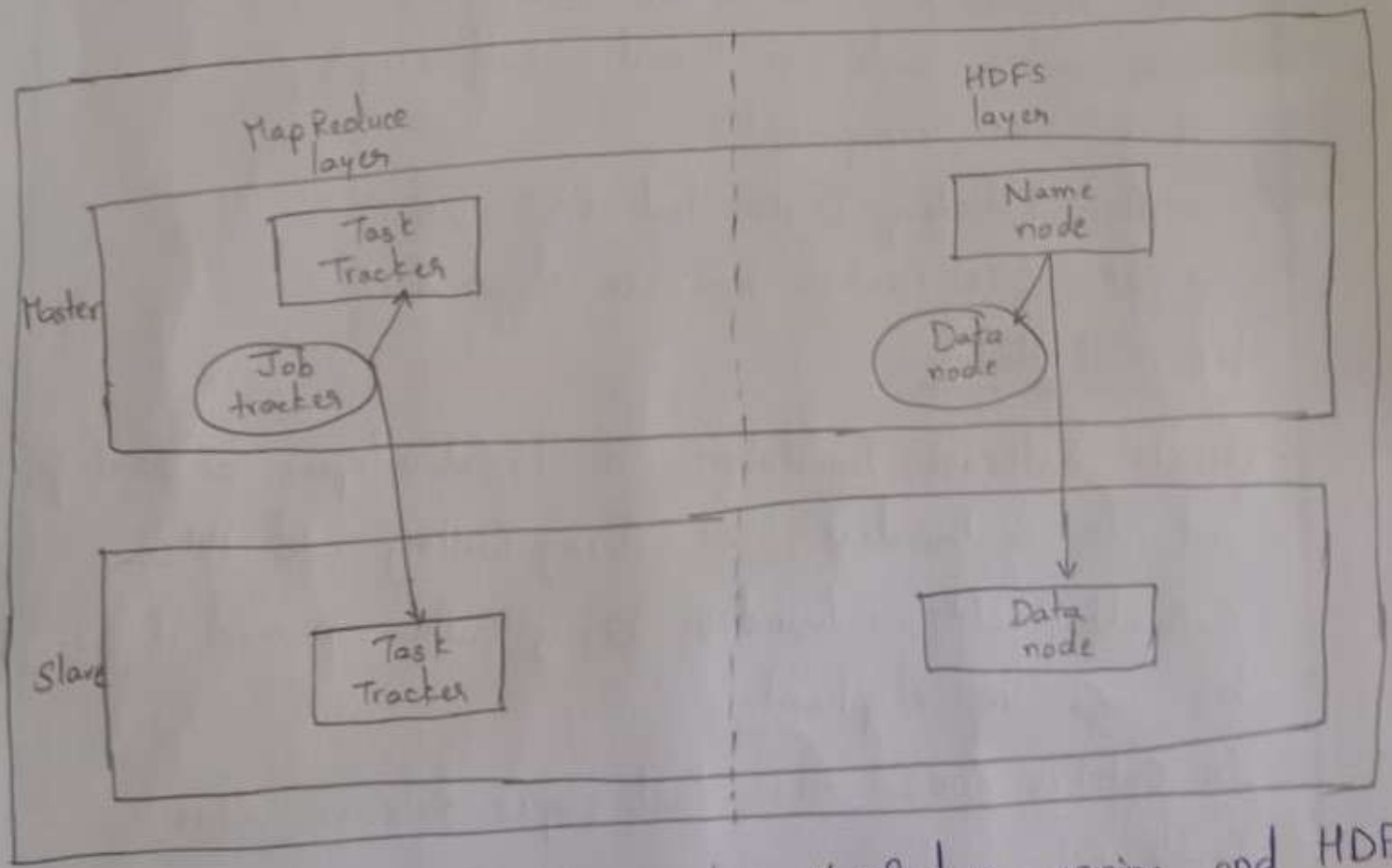
- Apache Software foundation is the developers of Hadoop and it's co-founders are Doug Cutting and Mike Cafarella. It's cofounder Doug Cutting named it on his son's toy elephant.
- In October 2003, the first paper release was Google File system.

- In January 2006, MapReduce development started on Apache Nutch which consisted of around 6000 lines coding for it and around 5000 lines coding for HDFS.
- In April 2006 Hadoop 0.1.0 was released.

Features of Hadoop

- It is fault tolerance
- It is highly available
- Its programming is easy
- It have huge flexible storage
- It is low cost.

Hadoop Architecture



- It is a package of file system, MapReduce engine and HDFS
- MapReduce engine can be MapReduce/MR1 or YARN/MR2.

Class Note:

Subject: BDT

Faculty: V. Divya

Topic: Hadoop Architecture

Unit No.: 2
Lecture No.: 2
Link to Session
Planner (SP) s.No. ..... of SP
Date Conducted:
Page No. 2

- A Hadoop cluster consists of a single master and multiple slave nodes.
- The master node includes Job tracker, Task tracker, Name Node and data node whereas the slave node include Data node and Task Tracker.

Advantages of Hadoop:

- Fast: In HDFS the data distributed over the cluster and are mapped which helps in faster retrieval. Even the tools to process the data are often on the same servers, thus reducing the processing time. It is able to process terabytes of data in minutes and Peta bytes in hours.
- Scalable: Hadoop cluster can be extended by just adding nodes in the cluster.
- Cost effective: Hadoop is open source and uses commodity hardware to store data so it really cost effective as compared to traditional relational database management system.
- Resilient to failure: HDFS has the property with which

it can replicate data over the network, so if one node is down or some other network failure happens, then Hadoop takes the other copy of data and use it. Normally, data are replicated thrice but the replication factor is configurable.

Year	Event
2003	Google released the paper, Google file System(GFS)
2004	Google released a white paper on MapReduce
2006	- Hadoop introduced - Hadoop 0.1.0 released - Yahoo deploys 300 machines and within this year reaches 600 machines.
2007	- Yahoo run 2 clusters of 1000 machines - Hadoop includes HBase
2008	- YARN JIRA opened - Hadoop becomes the fastest system to sort 1 terabyte of data on a 900 node cluster within 209 sec. - Yahoo clusters loaded with 10 terabytes per day - Cloudera was founded as a Hadoop distributor.
2009	- Yahoo runs 17 clusters of 24000 machines - Hadoop becomes capable enough to sort a petabyte. - MapReduce and HDFS become separate subproject. - Hadoop added the support for Kerberos.
2010	- Hadoop operates 4,000 nodes with 40 petabytes - Apache Hive and Pig released.
2011	- Apache Zookeeper released. - Yahoo has 42,000 Hadoop nodes and hundreds of petabytes of storage.

Class Note:

Subject : BDT

Faculty : V. Divya

Topic : Hadoop and its ecosystem.

Unit No. : 2
Lecture No. : 3
Link to Session Planner (SP) S.No. of SP
Date Conducted:
Page No. 3

2012 - Apache Hadoop 1.0 version released.
2013 - Apache Hadoop 2.2 version released
2014 - Apache Hadoop 2.6 version released
2015 - Apache Hadoop 2.7 version released
2017 - Apache Hadoop 3.0 version released
2018 - Apache Hadoop 3.1 version released.

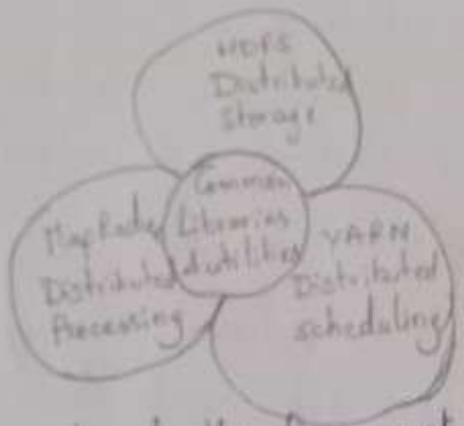
Hadoop and its ecosystem:

- Apache Hadoop is a computing environment in which input data stores, processes and stores the results.
- The environment consists of clusters which distribute at the cloud or set of servers. Each cluster consists of a string of data files constituting data blocks.
- The Hadoop system cluster shifts files in data blocks and the complete system consists of a scalable distributed set of clusters.
- Infrastructure consists of cloud for clusters where a cluster consists of sets of computers or PCs.
- Hadoop platform provides a low cost Big data platform, ~~takes~~ - just - ~~five~~ nnnodes. which is open source

and uses cloud services.

- Hadoop enables distributed processing of large datasets across clusters of computers using a programming model called MapReduce.
- The system characteristics are scalable, self manageable, self healing and distributed file system.

- Hadoop core components:



Hadoop core components of the framework are:

1. Hadoop common: The common module contains the libraries and utilities that are required by the other modules of Hadoop.
2. Hadoop Distributed File System (HDFS): A Java based distributed file system which can store all kinds of data on disks at the clusters.
3. MapReduce v1: Software programming model in Hadoop 1 using Mapper and Reducer. The v1 processes large sets of data in parallel and in batches.
4. YARN: Software for managing resources for computing. The user application tasks or sub-tasks run in parallel at the Hadoop, uses scheduling and handles the requests for the resources in distributed running of the tasks.

Class Note:

Subject : BDT

Faculty : V. Divya

Topic: Hadoop Distributed File system

Unit No. : 2
Lecture No.: 4
Link to Session
Planner (SP) s.No. ..... of SP
Date Conducted:
Page No. 4

5. MapReduce V2 - Hadoop 2 YARN based system for parallel processing of large datasets and distributed processing of the application tasks.

Features of Hadoop:

1. Fault efficient scalable, flexible and modular design which uses simple and modular programming model.
2. Robust design of HDFS
3. Store and process Big data.
4. Distributed clusters computing model with data locality.
5. Hardware - fault tolerant
6. Open Source framework
7. Java and Linux based.

Hadoop Ecosystem components:-

- The four layers in the figure are as follows:
  (i) Distributed storage layer
  (ii) Resource manger layer for job or application subtasks scheduling and execution
  (iii) Processing framework layer, consisting of Mapper and Reducer for the MapReduce process flow.

iv, AP is at application support layer

- Hadoop Streaming:
- HDFS with MapReduce and YARN based system enables parallel processing of large datasets.
- Spark provides in-memory processing of data, thus improving the processing speed.
- Spark and Flink technologies enable in-stream processing.

Hadoop Pipes:

- These are the C++ pipes which interface with MapReduce. Java native interfaces are not used in pipes.
- Apache Hadoop provides an adapter layer, which processes in pipes.
- A pipe means data streaming into the system at Mapper input and aggregated results flowing out at outputs.
- The adapter layer enables running of application tasks in C++ coded MapReduce programs.
- Applications which require faster numerical computations can achieve higher throughput using C++ when used through the pipes, as compared to Java.
- Pipes do not use the standard I/O when communicating with Mapper and Reducer codes.
- Cloudera distribution including Hadoop (CDH) version CDH 5.0.2 runs the pipes.
- Distribution means software downloadable from the website distributing the codes.

Class Note:

Subject : BDT

Faculty : Y. Divya

Topic: Hadoop Distributed File System

Unit No. : 2
Lecture No.: 5
Link to Session
Planner (SP) s No. ..... of SP
Date Conducted:
Page No. 5

Hadoop Distributed File system:

- Big data analytics applications are software applications that leverage large scale data.
- HDFS is a core component of Hadoop. It is designed to run on a cluster of computers and servers at cloud based utility services.
- HDFS stores big data which may range from GBs ($1GB = 2^{30}B$) to PBs ($1PB = 1015B$), nearly the $2^{50}B$).
- HDFS stores the data in a distributed manner in order to compute fast. The distributed data store in HDFS stores data in any format regardless of schema.
- HDFS provides high throughput access to data centric applications that require large scale data processing workloads.

→ HDFS Data Storage:

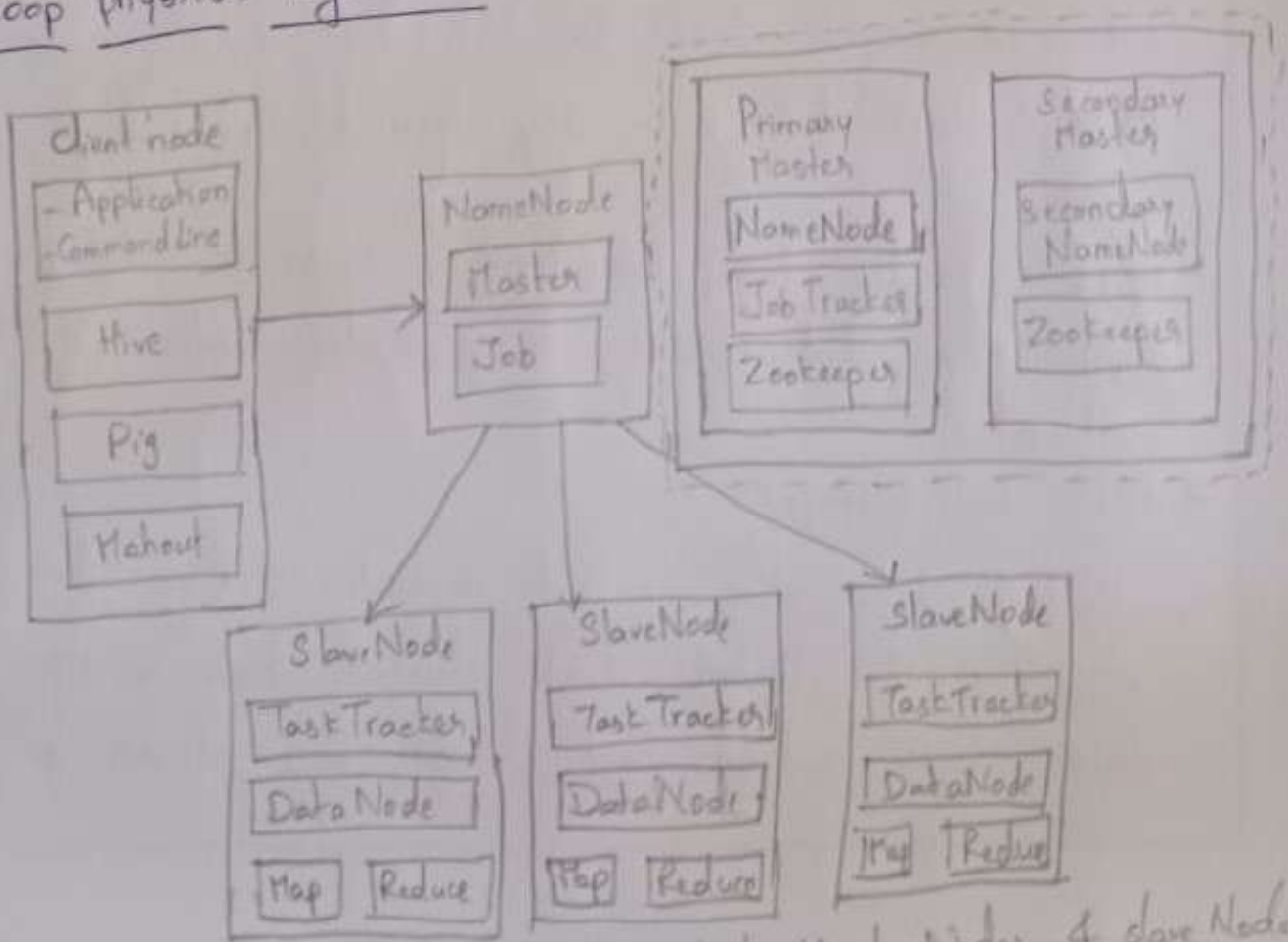
- Hadoop data store concept implies storing the data at a no of clusters. Each cluster has a no. of data stores, called racks. Each rack stores a no.

of DataNodes. Each DataNode has a large no. of data blocks.
- The racks distribute across a cluster. The nodes have processing and storage capabilities. The nodes have the data in data blocks to run the application tasks.
- A file, containing the data divides into data blocks. and its default size is 64MBs.

Hadoop HDFS features are as follows:

i. Create, append, delete, rename and attribute modification functions.
ii. Content of individual file cannot be modified or replaced but appended with new data at the end of file.
iii. Write once but read many times during usages and processing
iv. Average file size can be more than 500MB.

Hadoop physical organization



The client, master NameNode, MasterNodes & slave Nodes

Class Note:

Subject : BDT

Faculty : V. Divya

Topic : HDFS & Hapfedue framework & Programming model

Unit No. : 2
Lecture No. : 6
Link to Session
Planner (SP) s.No. ..... of SP
Date Conducted:
Page No. 6

HDFS Commands:-

- The above figure showed Hadoop common module, which contains the libraries and utilities.
- They are common to other modules of Hadoop. HDFS shell is not compliant with the POSIX. Thus, the shell cannot interact similar to Unix or Linux.
- Commands for interacting with the files in HDFS require /bin/hdfs dfs <args>, where args stands for the command arguments.
- Full set of Hadoop shell commands can be found at Apache Software websit.

-mkdir - Assume stu_filedir is a directory of student files. Command for creating the directory is
$Hadoop hdfs -mkdir /user/stu_filesdir

-put - $Hadoop hdfs -put copies file into directory

-ls - Assume all files to be listed. Then $hdfs dfs -ls command does provide the listing.

-cp - $Hadoop hdfs -cp copies file into directory

MapReduce framework and programming model:

- MapReduce is a programming model for distributed computing.
- Mapper means software for doing the assigned task after organizing the data blocks imported using the keys. A key specified in a command line of Mapper.
- Reducer means software for reducing the mapped data by using the aggregation, query or user-specified function. & provides a concise cohesive response for the application.
- Aggregation function means the function that groups the values of multiple rows together to result a single value of more significant meaning or measurement.
- Querying function means a function that finds the desired values.

Features of MapReduce framework are as follows:

i. Provides automatic parallelization and distribution of computation based on several processors.
ii. Processes data stored on distributed clusters of DataNodes and racks.
iii. Allows processing large amount of data in parallel.
iv. Provides scalability for usages of large number of servers
v. Provides MapReduce batch-oriented programming model in Hadoop version
vi. Provides additional processing modes in Hadoop 2 YARN based system and enables required parallel processing.

Class Note:

Subject : BDT

Faculty : V. Divya

Topic : Hadoop Yarn

Unit No. : 2
Lecture No. : 7
Link to Session Planner (SP) S.No ..... of SP
Date Conducted:
Page No. 7

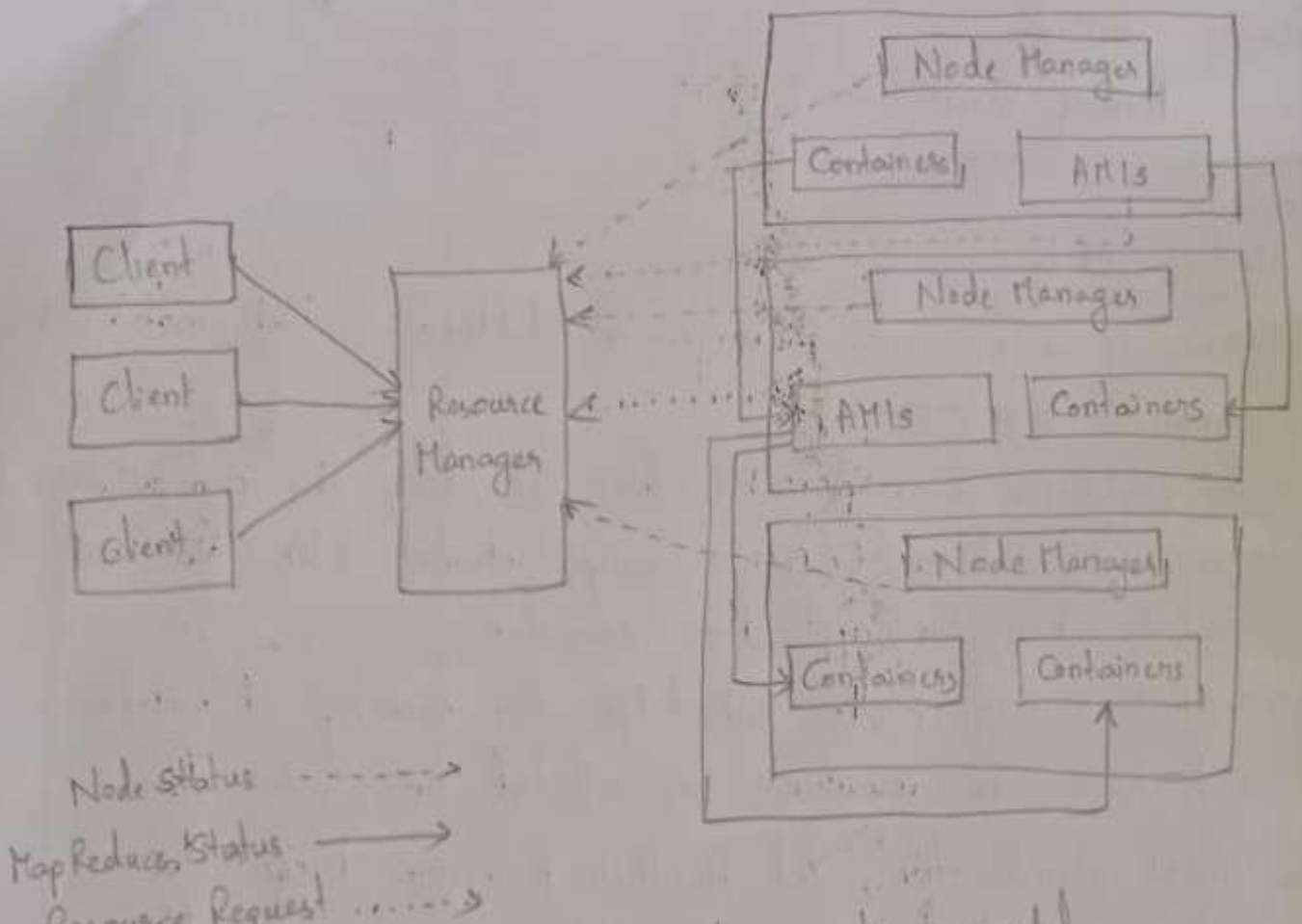
Hadoop YARN:

- YARN is a resource management platform. which manages computer resources.
- The platform is responsible for providing the computational resources, such as CPUs, memory, network I/O which are needed when an application executes.
- YARN manages the schedules for running of subtasks and uses the resources in alloted time intervals.
- YARN stands for Yet Another Resource Negotiator. and separates the resource management and processing components.
- YARN enables running of multi-threaded applications and manages & allocates the resources for the application subtasks and submits the resources for them at Hadoop system.

→ Hadoop 2 Execution model:

- The below figure shows the YARN components - Client, Resource Manager (RM), Node Manager (NM), Application Master (AM) and Containers.

- illustrates YARN components namely, Client, Resource Manager(RM), Node Manager(RM), Application Master (AM) and Containers.



YARN based execution model

- A MasterNode has two components: i, Job History Server and ii, Resource Manager (RM).
- A Client Node submits the request of an application to the RM.
- Multiple NMs are at a cluster.
- The AMI performs role of an Application Manager that estimates the resources requirement for running an application program or subtask.
- NM is a slave of the infrastructure.
- Each NM assigns a container for each AMI.
- RM allots the resources to AM and thus to AppIMs for using assigned containers on the same or other NM for running the application subtasks.

Class Note:

Subject: BDT

Faculty: V. Divya

Topic: Hadoop Ecosystem Tools, HDFS Design Features

Unit No.: 2
Lecture No.: 8
Link to Session
Planner (SP) s.No. of SP
Date Conducted:
Page No. 8

Hadoop Ecosystem Tools

- A simple framework of Hadoop enabled development of a number of opensource projects has quickly emerged
- They solve very specific problems related to distributed storage and processing model.

Ecosystem tool	Functionalities
ZooKeeper - Coordination service	- Provisions high performance coordination service for distributed running of applications and tasks.
Avro - Data serialization & transfer utility	- Provisions data serialization during data transfer blw application and processing layers
Oozie	- Provides a way to package and bundles multiple coordinator and workflow jobs and manage the lifecycle of these jobs
Flume - Large data transfer utility	- Provisions for reliable data transfer and provides for recovery in case of failure.
Ambari - A web based tool	- Provisions, monitors and manages and viewing of functioning of cluster, MapReduce, Hive and Pig APIs
Chukwa - A data collection system	- Provisions and manages data collection system for large and distributed systems

HBase - A structured data store using database
- Provisions a scalable and structured database for large tables.

Cassandra - A database
- Provisions scalable and fault-tolerant database for multiple masters.

Hive - A data warehouse system
- Provisions data aggregation, data summarization, data warehouse infrastructure, adhoc querying and SQL like scripting language for query processing using HiveQL

Pig - A high level dataflow language
- Provisions dataflow functionality and the execution framework for parallel computations.

Mahout - A machine learning slw
- Provisions scalable mlc learning and library functions for data mining and analytics.

i, Hadoop Ecosystem:
Consider ZooKeeper, Oozie, Sqoop and Flume.

HDFS Design features:
- Write once/read many design is intended to facilitate streaming reads.
- File may be appended, but random seeks are not permitted. There is no corelations of data.
- Conversed data storage and processing happen on the same server nodes.
- Moving computation is cheaper than moving data.
- Reliable file system maintains multiple copies of data across the cluster. Consequently, failure of a single node will not bring down the file system.
- A specialized file system is used, which is not designed for general use.

Class Note:

Subject : BDT

Faculty : V. Divya

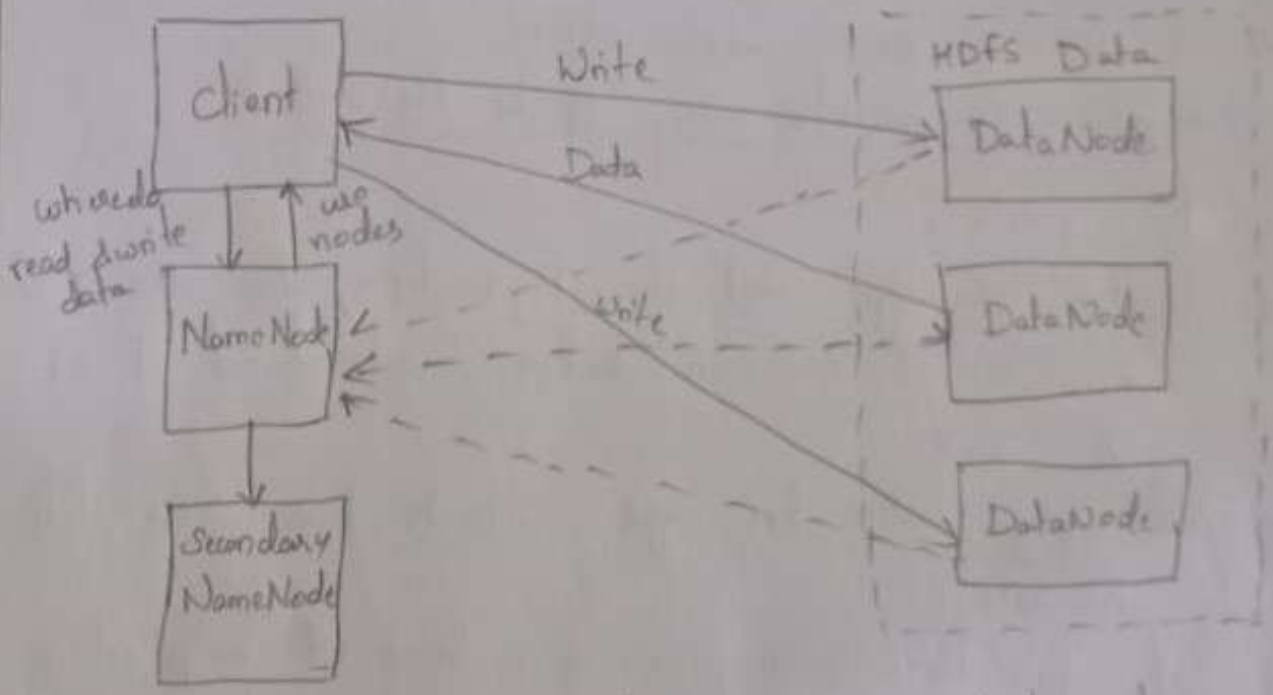
Topic: Components

Unit No. : 2
Lecture No.: 9
Link to Session Planner (SP) S.No. of SP
Date Conducted:
Page No. 9

HDFS Uses Commands, Hadoop Ecosystem components using Apache Pig: 2.10

Components:

- The design of HDFS is based on two types of nodes: a NameNode and multiple dataNode.
- In a basic design a single NameNode manages all the metadata needed to store and retrieve the actual data from the DataNodes.



Various system roles in HDFS deployment

HDFS User commands:

dfs → runs a file system command on the file systems supported in Hadoop.

namenode -format → format the DFS file system

Secondary namenode → run the DFS secondary namenode.

namenode → run the DFS namenode.

journalnode → run the DFS journalnode

zkfc → run the zk failover controller daemon

datanode → run a DFS datanode

dfsadmin → run a DFS admin client

haadmin → run a DFS HA admin client

headmin → run a DFS admin client.

Hadoop Ecosystem Components using Apche Pig: 2.10 :

- Pig has two parts
  Pig Latin → the language and the Pig runtime, for the eneries environment. You can better understand it as Java & JVM.
- It supports pig latin language, which has SQL like command structure.
- 10 lines of pig latin = approx. 200 lines of mapReduce Java code.
- the compiler internally converts pig latin to mapReduce.
  It provides a sequential set of mapReduce jobs and that's an abstract.
- PIG was initially developed by Yahoo.
- It gives you a platform for building data flow for ETL processing and analysis huge datasets.

Class Note:

Subject : BDT

Faculty : V. Divya

Topic : Hive, Flume, Sqoop

Unit No. : 2
Lecture No. : 10
Link to Session
Planner (SP) S.No. ... of SP
Date Conducted:
Page No. : 10

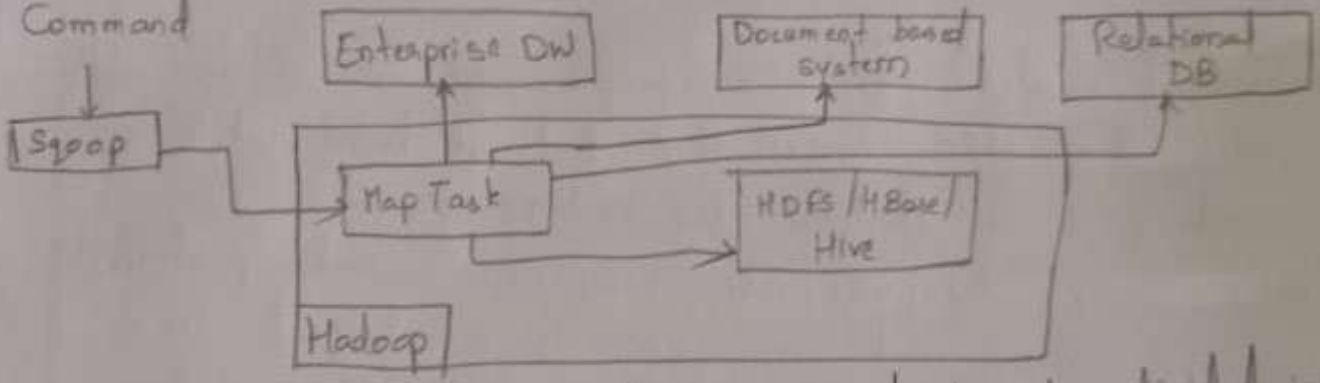
Hive :-

- facebook created HIVE for people who are fluent with SQL. Thus HIVE makes them feel at home while working in a Hadoop Eco system
- HIVE is a datawarehouse component which perform reading, writing and managing large datasets in a distributed environment using SQL like interface.
- HIVE + SQL = HQL
- The query language of HIVE is called Hive query language (HQL) which is very similar like SQL.
- It has 2 basic components : Hive command line and JDBC/ODBC drivers
- The Hive command line interface is used to execute HQL commands.
- Java Database Connectivity (JDBC) and Object Database Connectivity (ODBC) is used to establish connection from datastorage.

Sqoop

Major difference between flume and Sqoop is

- Flume only ingests unstructured data or semi structured data into HDFS
- While sqoop can import as well as export structured data from RDBMS & Enterprise datawarehouses to HDFS & vice versa



- When sqoop command submitted, main task gets divided into sub tasks which is handled by individual map Task internally.

Flume

- Flume is a service which helps in ingesting unstructured and semi structured data into HDFS.
- It gives us a solution which is reliable and distributed and helps us in collecting, aggregating and moving large amount of datasets.
- It helps us to ingest online streaming data from various sources like network traffic, social media, email messages, log files etc in HDFS.