Big Data Analytics
Assignment

Y. Keerthi
19131A05R3
CSE-4

1) What is Big Data? What are features and challenges of Bigdata

A) Bigdata refers to datasets that could not be percieved, acquired, managed and processed by traditional IT and software/hardware tools within a tolerable time.

→ It is a collection of large datasets that can't be processed using traditional computing techniques.

Features: 5V's of big data

i) Volume: It is a huge amount of data

→ To determine value of data, Size of data plays a very crucial rde. If volume of data is very large then it is actually considered as Bigdata. This means whether particular data can actually be considered as Bigdata or not is dependent upon volume of data

ii) Velocity: It refers to high Speed of accumulation of data.

→ In Bigdata velocity, data flows in from sources like machines networks, social media, Phones etc.

→ There is a massive & continous flow of data. This determines potential of data that how fast data generated & processed.

iii) Variety: It refers to nature of data i.e Structured, Semi Structured, Unstructured data.

→ It also refers to heterogenous sources.

→ variety is basically arrival of data.

iv) Veracity: It refers to inconsistencies and uncertainity in data.

→ Bigdata is also variable because of multitude of data dimensions resulting from multiple disparate data types.

V) Value: After having 4V's into account there comes one more V which is value.

→ Data itself no use or importance but it needs to be converted into something valuable to extract information

Challenges:

i) Sharing and Accessing data

→ Perhaps most frequent challenge in bigdata efforts is inaccesability of data sets from external sources.

→ It include need for inter and intra institutional legal documents.

ii) privacy and security

→ It is another most important challenge. Most of the organisation are unable to maintain regular checks due to large amounts of data generation.

→ It should be necessary to perform security checks and observations in real time because it is more beneficial.

iii) Analytical challenge:

→ There are some huge analytical challenge in Bigdata which arise some questions like how to deal with problem if data volume gets too large.

iv) Technical challenges

Quality of data: when there is a collection of large amount of data and storage of data, it comes at cost-

Fault tolerance: It is extremly hard, involving intricate algorithms.

Scalability: Bigdata projects can grow and evolve rapidly. The scalability issue has lead towards cloud computing.

2) what are developments of Bigdata? Explain in detail.

A) A project called Hadoop was born in 2005. Hadoop is very important technology in field of bigdata. It provides a Software Framework for distributed Storage and processing of bigdata using MapReduce programming model. Many countries around world and some research institutes have conducted some pilot projects on Hadoop and achieved series of results

Between 2012 and 2015, many governments and companies around world including UN, Published series of ideas or outlines of action to promote development of bigdata. After that bigdata has entered highspeed developing phase. ~~and~~

3) Distinguish between Bigdata, Hadoop, & cloud.

A)

Features	Bigdata	Hadoop
Definition	It refers to large volume of both structured and unstructured data.	It is framework to handle and process large volume of bigdata
Significance	It has no significance until it is processed and utilized to generate revenue.	It is a tool that makes bigdata more meaningful by procesing data.
Storage	It is very difficult to store bigdata because it comes in structured and unstructured form	Apache Hadoop HDFS is capable of storing bigdata
Accessability	When it comes to accessing, it is very difficult.	It is very fast when compared to other tools.

Big data	cloud computing
1) It refers to data which is huge in size and also Increasing rapidly.	1) It refers to on demand availability of computing resources over internet.
2) It includes Structured, unstructured, Semistructured data.	2) It includes Iaas, Paas, Saas
3) Volume, velocity, variety, veracity, value are considered as 5 important features.	3) On demand availability of IT resources, broad network access are considered
4) Purpose is to organise and extract useful information from it and use that for improvement	4) Purpose is to store and process in cloud or availing remote IT services.
5) low cost.	5) cost effective, robust
6) Social media data, e-commerce data, IoT sensor data included	6) Includes AWS, IBM cloud etc.

4) Explain how IoT related to Bigdata?

A) IoT describes network of physical objects that are embedded systems with sensors, software and other technologies for Purpose of connecting and exchanging data with other devices.

IoT	Big data
1) It is a global System of interrelated computing devices that are able to sense, collect and exchange data over internet.	1) It is described as large sets of data generated from variety of sources that are so large to process using traditional techniques
2) The concept to provide inter-connection between devices to create smart environment thereby making machines smart enough to bypass human Intermediaries	2) The concept to find insights in new and emerging types of data and content that lead to better decisions and strategic business moves
3) It collects, analyzes, processes data streams in realtime.	3) Data streams ~~with human~~ not subjected to process realtime.
4) It involves analysing machine generated data.	4) It deals with human-generated data.
5) It is about simultaneously track and monitor assets	5) It is more into collecting and accumulating huge data for analysis.

5) Explain the following i) Cloud preliminary ii) IoT preliminary

Cloud preliminary:

→ It refers to on demand availability of computing resources over internet. These resources includes servers, storage, databases over internet and all resources can be used as per requirement of customer.

→ It is evolved from distributed computing, Parallel Computing and Grid computing.

→ In general sense, It means delivery and use mode of services. In narrow sense, It means delivery and use mode of IT infrastructure.

IOT preliminaries :

→ The basic idea of IOT is to connect different objects in real world.

→ IOT deemed as extension of Internet and is an important part of future Internet.

→ IOT has features : i) various terminal equipments
ii) Automatic data acquisition
iii) Intelligent terminal.

Architecture of IOT

Sensors / device → collecting data through Sensors.

↓

Connectivity → wifi / Bluetooth

↳ transfer to cloud.

Internet

↓

Data preprocessing → using BDA tools and algorithms

↓ ↓

User Interface → output

Unit - II

1) What is Hadoop? Explain Architecture of Hadoop.

A) Hadoop is an open source software framework used for storing and processing bigdata in a distributed manner on large clusters of commodity hardware.

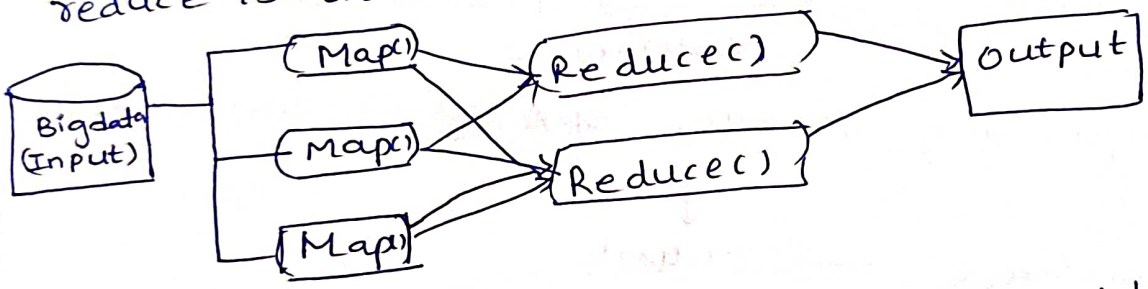
Hadoop Architecture : It consists of 4 component.
i) Map Reduce ii) HDFS iii) YARN iv) common utilities or Hadoop common

i) Map Reduce :

→ It is like an algorithm or data structure that is based on YARN framework.

→ Major feature is to perform distributed processing in Parallel in Hadoop cluster.

→ It has 2 tasks divided phase wise.
In 1st phase, Map is utilised and in next phase, reduce is utilised.

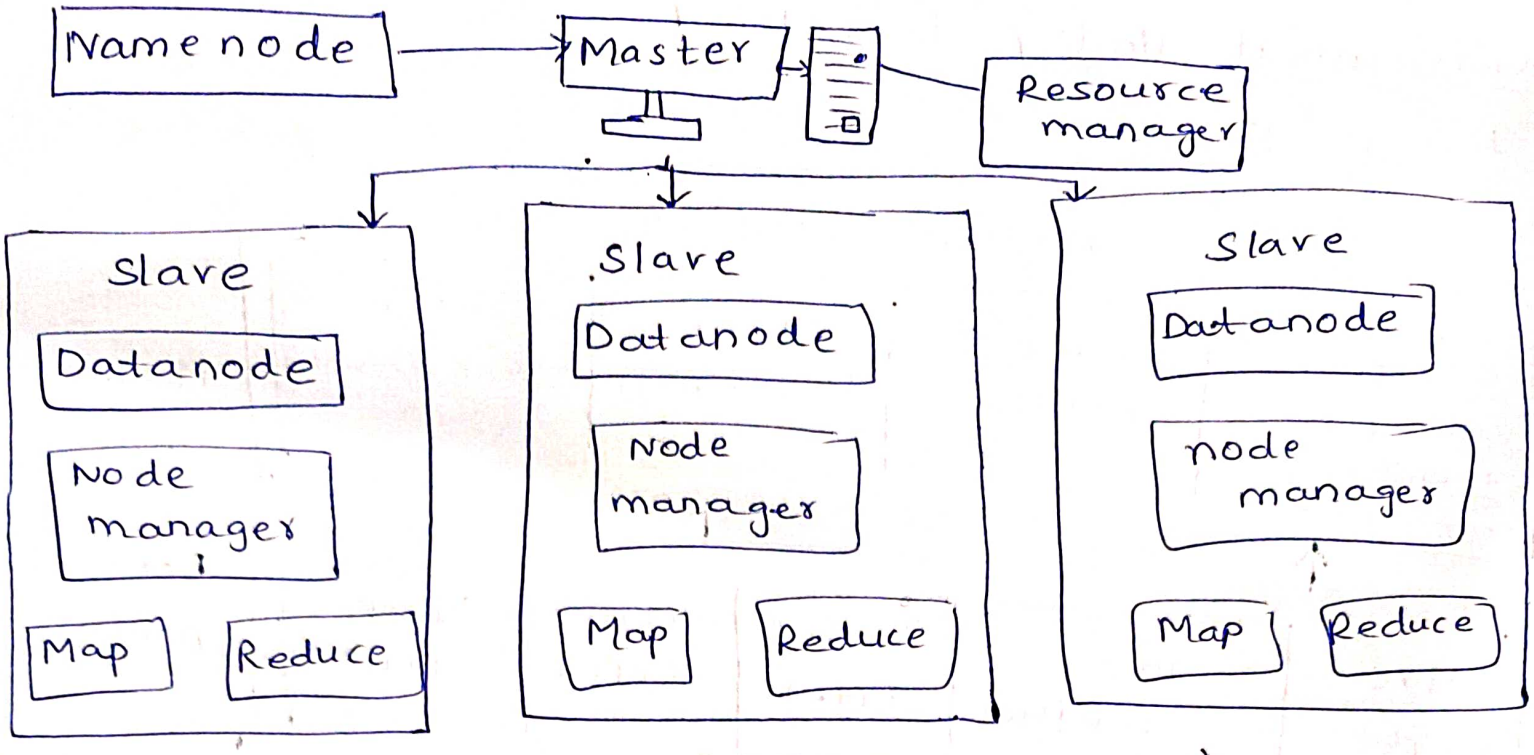


ii) HDFS :
HDFS in Hadoop provides fault tolerance and high availability to storage layer and other devices present in that Hadoop cluster.

Name node : It works as Master in Hadoop cluster that guides datanode. Mainly used to store metadata.

Datanode : It works as slave. Mainly utilised for storing data in Hadoop cluster.

→ Name node instructs datanodes with operations like delete, create, replicate etc.



iii) YARN (yet Another Resource Negotiator)

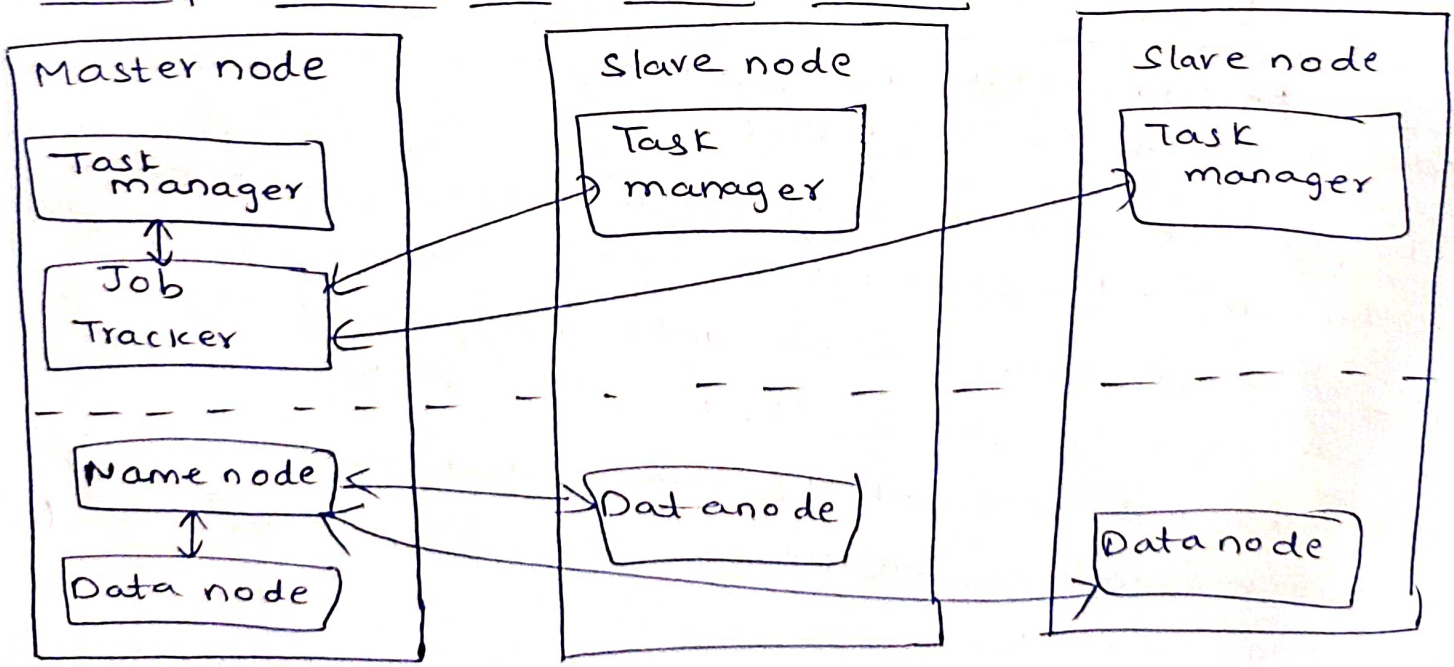
→ It is a framework on which MapReduce works.

→ It performs 2 operations.

Job scheduling → divides Big task into Small jobs

Resource Management → manage all resources that are available for running Hadoop cluster.

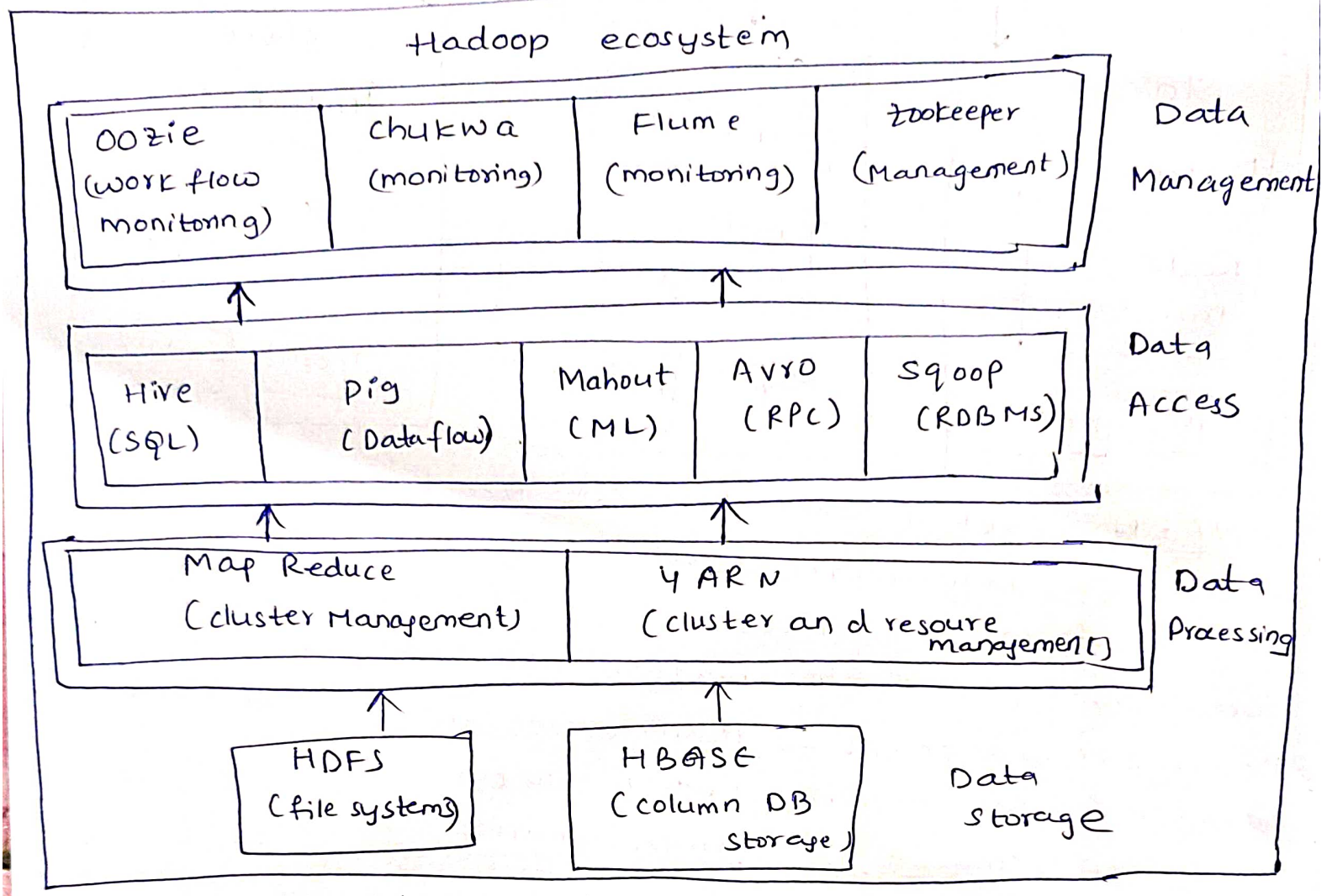
iv) Hadoop common or common utilities :



→ These utilities are used by HDFS, YARN and MapReduce for running cluster.

2) Write about Hadoop Ecosystem?

A) Hadoop Ecosystem :



Oozie : Performs task of a scheduler, thus scheduling jobs and binding togethers as single unit.

Zookeeper : It overcame huge issue of management of coordination and Syncronization among resources of Hadoop.

Hive : It performs reading and writing of large datasets. Its query language is HQL. It is highly Scalable.

Pig : Basicalled developed by Yahoo which works on Pig Latin Language which is Similar to SQL. It does ~~not~~ the work of executing commands.

Mahout : Allows Machine learnability to a System or Applications. It provides various Libraries or functionalities

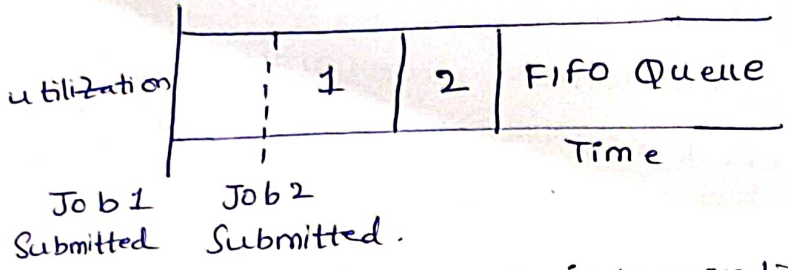
3) What is FIFO, Fair, Capacity schedulers?

A) There are 3 types of schedulers in Hadoop

i) FIFO scheduler ii) Capacity scheduler iii) Fair scheduler

i) FIFO scheduler : First In First Out.

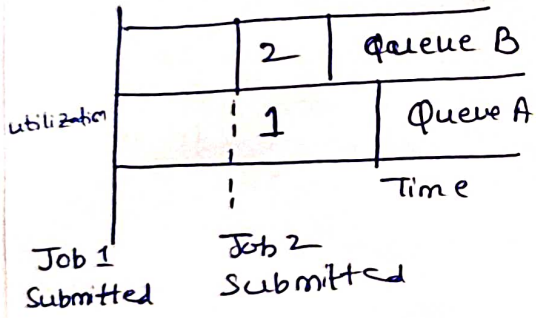
→ Tasks that comes first will be served first. This is default scheduler we use in Hadoop.

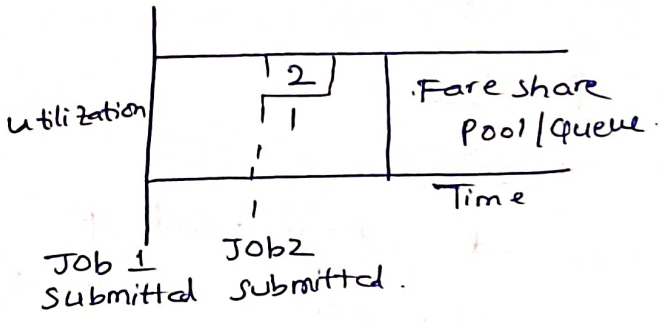


→ Once, job scheduled, no intervention allowed.

ii) Capacity scheduler : we have multiple job queues for scheduling our tasks. Each job queues has own slots to perform task.

→ It has 3 components that are root, parent, leaf.





iii) Fair Scheduler : It is very much similar to capacity Scheduler. The priority of Job is kept in consideration

→ It takes scheduling decisions on basis of memory.

→ we configure it to work with CPU.

→ The major thing in Fair scheduler, whenever high Priority job arises in same queue, task processed in parallel by replacing some portion from already dedicated slots.

4) What is HIVE? Explain

A) Hive is a data warehouse and an ETL tool which provides SQL-like interface between user and Hadoop distributed file system. Hive is interface between HDFS and user.

Components of Hive

i) HCatalog -

It is a Hive component and is a table as well as a store management layer for Hadoop.

→ It enables user, along with various data processing tools like pig and MapReduce which enables to read and write on grid easily.

ii) WebHCat -

→ It provides a service which can be utilized by user to run Hadoop's.

→ Map Reduce, Pig, Hive tasks or functions Hive metadata operations with HTTP interface.

Modes of Hive :

i) Local Mode

→ It is used, when Hadoop is built under pseudo mode which has only one data node.

ii) Map Reduce Mode

→ It is used, when Hadoop is built with multiple data nodes and data is divided across various nodes.

5) What is Pig? Write down Execution modes of PIG?

A) Pig is a high-level platform or tool which is used to process large datasets.

→ It provides high level scripting language known as Pig Latin Language.

→ Pig Engine has 2 types of execution environment. i-e Local Execution environment in a single JVM and distributed environment in a Hadoop cluster.

Modes :

i) Local Mode :

→ In this mode of execution, we need single machine and all files are installed and run using localhost and filesystem.

ii) Map Reduce Mode

→ It is the default mode of Apache Pig Grunt shell.

→ In this mode, we need to load data in HDFS and then we perform operation.

Pig Execution Modes :

A user can execute Apache pig Latin scripts in 3 ways as below :

i) Interactive Mode (Grunt shell):

→ In this mode, user can interactively run Apache Pig using Grunt shell. To invoke Grunt shell, run Pig command.

→ users can submit commands and get a result there only.

ii) Batch node :

→ In this mode, a user can run Apache pig in batch mode by creating pig Latin script file and running it from local or Mapreduce mode.

iii) Embedded Mode :

→ We can define our own functions called as UDF (user defined functions