1. **What is BigData? What are the features &challenges of BD?**

Def: Big data refers to the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time.

In 2010, Apache Hadoop defined big data as “datasets which could not be captured, managed, and processed by general computers within an acceptable scope.

Big data is a collection of large datasets that cannot be processed using traditional computing techniques.

**Advantages of BD:**

* Cost Savings
* Better decision-making
* Better Sales insights
* Increased Productivity
* Improved customer service.

**5V’s of big data:**

1. **Volume:**

* The name ‘Big Data’ itself is related to a size which is enormous.
* Volume is a **huge amount of data**.
* To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a ‘Big Data’. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
* Hence while dealing with Big Data it is necessary to consider a characteristic ‘Volume’.

**2. Velocity:**

* Velocity refers to the **high speed of accumulation of data.**
* In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
* There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
* *Example:* There are more than 3.5 billion searches per day are made on Google. Also, FaceBook users are increasing by 22%(Approx.) year by year.

**3. Variety:**

* It refers to **nature of data** that is structured, semi-structured and unstructured data.
* It also refers to heterogeneous sources.
* Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.

**4. Veracity:**

* It refers to **inconsistencies and uncertainty in data**, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
* Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
* *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

**5. Value:**

* After having the 4 V’s into account there comes one more V which stands for Value.
* Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value is the most important V of all the 5V’s.

**Big Data challenges :**

* ***Sharing and Accessing Data:***
  + Perhaps the most frequent challenge in big data efforts is the inaccessibility of data sets from external sources.
  + It include the need for inter and intra- institutional legal documents.
  + Accessing data from public repositories leads to multiple difficulties.
  + It is necessary for the data to be available in an accurate, complete and timely manner.
* ***Privacy and Security:***
  + It is another most important challenge with Big Data.
  + Most of the organizations are unable to maintain regular checks due to large amounts of data generation. However, it should be necessary to perform security checks and observation in real time because it is most beneficial.
  + There is some information of a person which when combined with external large data may lead to some facts of a person which may be secretive and he might not want the owner to know this information about that person.
* ***Analytical Challenges:***
  + There are some huge analytical challenges in big data which arise some main challenges questions like how to deal with a problem if data volume gets too large?
  + Or how to find out the important data points?
  + Or how to use data to the best advantage?
* ***Technical challenges:***
  + **Quality of data:**
    - When there is a collection of a large amount of data and storage of this data, it comes at a cost.
    - For better results and conclusions, Big data rather than having irrelevant data, focuses on quality data storage.
  + **Fault tolerance:**
    - Fault tolerance is another technical challenge and fault tolerance computing is extremely hard, involving intricate algorithms.
    - Nowadays some of the new technologies like cloud computing and big data always intended that whenever the failure occurs the damage done should be within the acceptable threshold that is the whole task should not begin from the scratch.
  + **Scalability:**
    - Big data projects can grow and evolve rapidly. The scalability issue of Big Data has lead towards cloud computing.
    - It leads to various challenges like how to run and execute various jobs so that goal of each workload can be achieved cost-effectively.
    - It also requires dealing with the system failures in an efficient manner.

1. **Distinguish between BD, Hadoop and Cloud.**

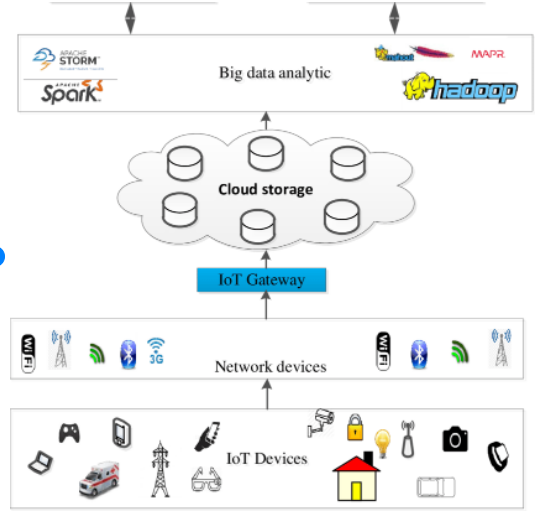
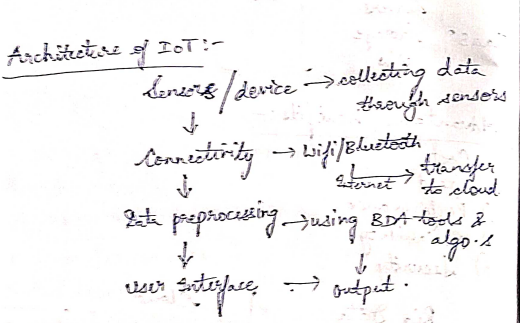
|  |  |  |
| --- | --- | --- |
| Features | Big Data | Hadoop |
| **Definition** | Big Data refers to a large volume of both structured and unstructured data. | Hadoop is a framework to handle and process this large volume of Big data |
| **Significance** | Big Data has no significance until it is processed and utilized to generate revenue. | It is a tool that makes big data more meaningful by processing the data. |
| **Storage** | It is very difficult to store big data because it comes in structured and unstructured form. | Apache Hadoop HDFS is capable of storing big data. |
| **Accessibility** | When it comes to accessing the big data, it is very difficult. | Hadoop framework lets you access and process the data very fast when compared to other tools. |

|  |  |  |
| --- | --- | --- |
| **S.No.** | **BIG DATA** | **CLOUD COMPUTING** |
| 01. | Big data refers to the data which is huge in size and also increasing rapidly with respect to time. | Cloud computing refers to the on demand availability of computing resources over internet. |
| 02. | Big data includes structured data, unstructured data as well as semi-structured data. | Cloud Computing Services includes Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). |
| 03. | Volume of data, Velocity of data, Variety of data, Veracity of data, and Value of data are considered as the 5 most important characteristics of Big data. | On-Demand availability of IT resources, broad network access, resource pooling, elasticity and measured service are considered as the main characteristics of cloud computing. |
| 04. | The purpose of big data is to organizing the large volume of data and extracting the useful information from it and using that information for the improvement of business. | The purpose of cloud computing is to store and process data in cloud or availing remote IT services without physically installing any IT resources. |
| 05. | Distributed computing is used for analyzing the data and extracting the useful information. | Internet is used to get the cloud based services from different cloud vendors. |
| 06. | Big data management allows centralized platform, provision for backup and recovery and low maintenance cost. | Cloud computing services are cost effective, scalable and robust. |
| 07. | Some of the challenges of big data are variety of data, data storage and integration, data processing and resource management. | Some of the challenges of cloud computing are availability, transformation, security concern, charging model. |
| 08. | Big data refers to huge volume of data, its management, and useful information extraction. | Cloud computing refers to remote IT resources and different internet service models. |
| 09. | Big data is used to describe huge volume of data and information. | Cloud computing is used to store data and information on remote servers and also processing the data using remote infrastructure. |
| 10. | Some of the sources where big data is generated includes social media data, e-commerce data, weather station data, IoT Sensor data etc. | Some of the cloud computing vendors who provides cloud computing services are Amazon Web Service (AWS), Microsoft Azure, Google Cloud Platform, IBM Cloud Services etc. |

1. **Explain how IoT is related to big data?**

The Internet of Things (IoT) describes the network of physical objects—“things”—that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet.

IoT Architecture:

| No | IOT | BIG DATA |
| --- | --- | --- |
| 1 | IoT is a global system of interrelated computing devices that are able to sense, collect, and exchange data over the Internet. | Big Data is described as large sets of data generated from a variety of sources that are so large to process using traditional techniques. |
| 2 | The concept is to provide interconnection between devices to create a smart environment thereby making machines smart enough to bypass human intermediaries. | The concept is to find insights in new and emerging types of data and content that lead to better decisions and strategic business moves. |
| 3 | IoT collects, analyzes, and processes data streams in real-time without any delay to make control decisions in an effective manner. | The data streams are not subjected to processing real-time and there is a delay between when the data is collected and when it is processed. |
| 4 | IoT involves analyzing machine-generated data such as sensors in home appliances and so on. | Big Data deals with human-generated data such as social media usage, photos, and videos, etc |
| 5 | IoT is about simultaneously collecting and processing data to make real-time decisions. | Big data is more into collecting and accumulating huge data for analysis afterward. |
| 6 | Using IoT you can track and monitor assets like trucks, engines, HVAC systems, and pumps. You can correct problems as you detect them. | With big data, you can analyze all the information you have about failures and start to uncover the root causes. |

1. **Explain the following:**

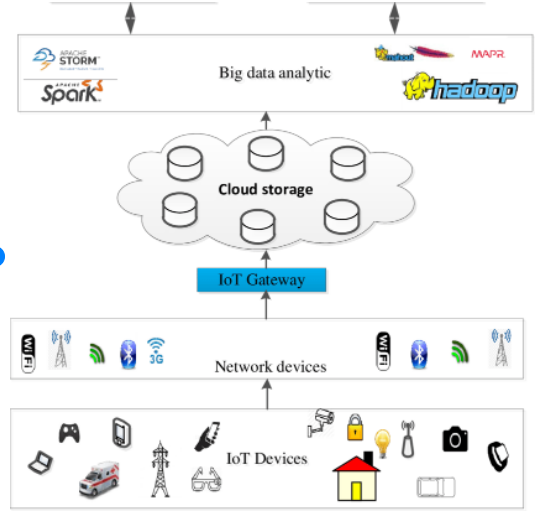
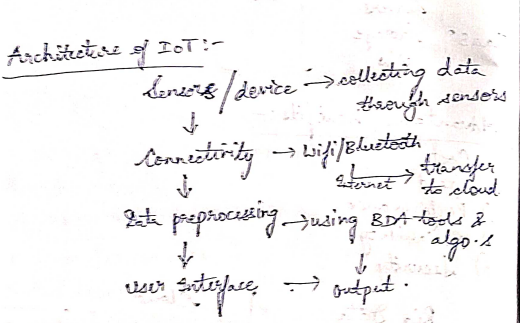
Cloud preliminaries:

* Cloud computing refers to the on demand availability of computing resources over internet. These resources includes servers, storage, databases, software, analytics, networking and intelligence over the Internet and all these resources can be used as per requirement of the customer.
* Cloud Computing is evolved from Distributed Computing, Parallel Computing, and Grid Computing.
* In a narrow sense, cloud computing means the delivery and use mode of IT infrastructure, i.e., acquiring necessary resources through the Internet on-demand.
* In a general sense, cloud computing means the delivery and use mode of services.
* In short, it refers to the case that users access a server through the network in a remote location and then use some services provided by the server.
* Services provided by cloud computing can be described by three service models and three deployment models, including wide network access, resource pool, rapidity, elasticity, and service management, thus meeting the requirements of many applications. Therefore, cloud computing will be instrumental for big data analysis and applications.

IoT Preliminaries:

* + The basic idea of IoT is to connect different objects in the real world (such as RFID, bar code readers, sensors, and mobile phones, etc.,) to exchange information.
  + IoT is deemed as the extension of the Internet and is an important part of the future Internet
  + IoT has the following main features :
    1. Various terminal equipments
    2. Automatic data acquisition
    3. Intelligent terminal.

IoT Architecture:

Unit 2:

**1)What is Hadoop? Explain Architecture of Hadoop.**

[Hadoop](https://www.edureka.co/blog/every-hadoop-component/) is an open-source software framework used for storing and processing Big Data in a distributed manner on large clusters of commodity hardware.

**Hadoop Architecture:**

The Hadoop Architecture Mainly consists of 4 components.

* MapReduce
* HDFS(Hadoop distributed File System)
* YARN(Yet Another Resource Negotiator)
* Common Utilities or Hadoop Common

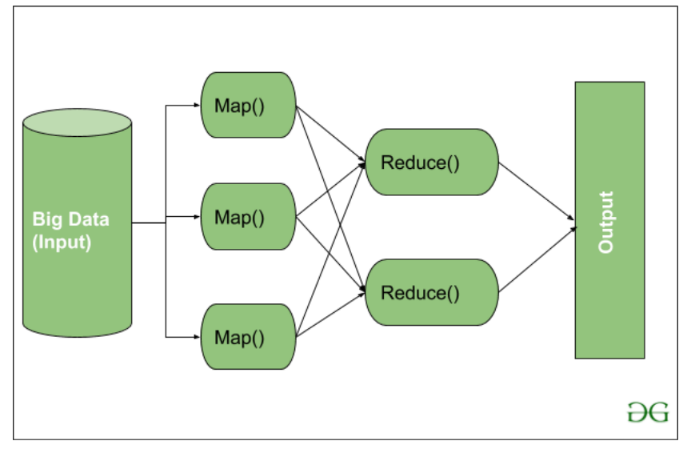
### 1. MapReduce

MapReduce nothing but just like an Algorithm or a data structure that is based on the YARN framework.

The major feature of MapReduce is to perform the distributed processing in parallel in a Hadoop cluster which Makes Hadoop working so fast.

MapReduce has mainly 2 tasks which are divided phase-wise:

In first phase, **Map** is utilized and in next phase **Reduce** is utilized.



Here, we can see that the *Input* is provided to the Map() function then it’s *output* is used as an input to the Reduce function and after that, we receive our final output.

**2. HDFS**

HDFS in Hadoop provides Fault-tolerance and High availability to the storage layer and the other devices present in that Hadoop cluster. Data storage Nodes in HDFS.

* NameNode(Master)
* DataNode(Slave)

Data in HDFS is always stored in terms of blocks.

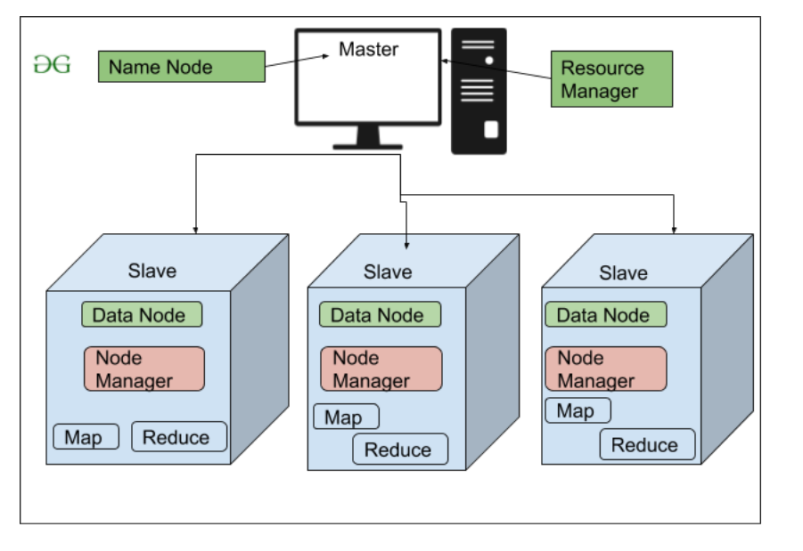
**NameNode:**

* NameNode works as a Master in a Hadoop cluster that guides the Datanode(Slaves).
* Namenode is mainly used for storing the Metadata i.e. the data about the data. Meta Data can be the transaction logs that keep track of the user’s activity in a Hadoop cluster.
* Meta Data can also be the name of the file, size, and the information about the location(Block number, Block ids) of Datanode that Namenode stores to find the closest DataNode for Faster Communication.
* Namenode instructs the DataNodes with the operation like delete, create, Replicate, etc.

**DataNode:**

* DataNodes works as a Slave.
* DataNodes are mainly utilized for storing the data in a Hadoop cluster, the number of DataNodes can be from 1 to 500 or even more than that. The more number of DataNode, the Hadoop cluster will be able to store more data. So it is advised that the DataNode should have High storing capacity to store a large number of file blocks.

**High Level Architecture Of Hadoop**



**3. YARN(Yet Another Resource Negotiator)**

YARN is a Framework on which MapReduce works.

YARN performs 2 operations that are Job scheduling and Resource Management.

The Purpose of Job schedular is to divide a big task into small jobs so that each job can be assigned to various slaves in a Hadoop cluster and Processing can be Maximized.

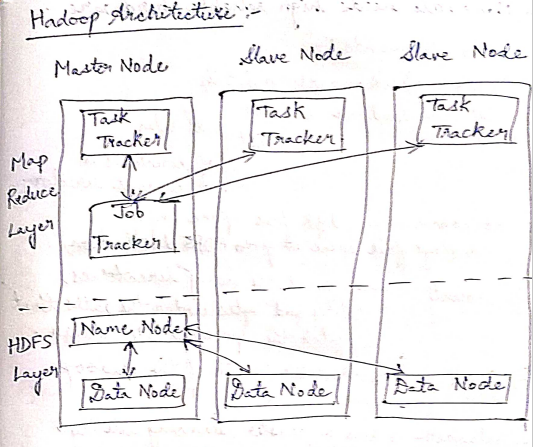
Job Scheduler also keeps track of which job is important, which job has more priority, dependencies between the jobs and all the other information like job timing, etc.

The use of Resource Manager is to manage all the resources that are made available for running a Hadoop cluster.

**Features of YARN**

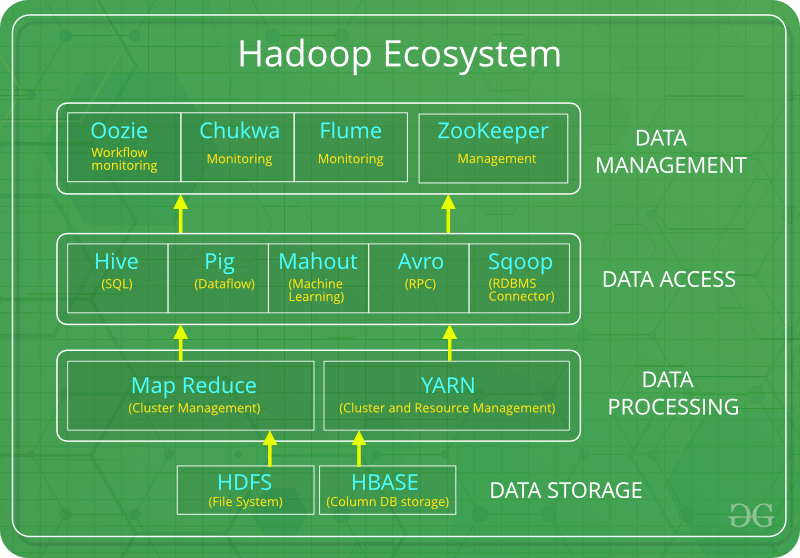
* Multi-Tenancy
* Scalability
* Cluster-Utilization
* Compatibility

**4. Hadoop common or Common Utilities:** Hadoop common or Common utilities are nothing but java library and java files. These utilities are used by HDFS, YARN, and MapReduce for running the cluster.



2)Write about Hadoop Ecosystem.

**Hadoop Ecosystem:**



The components that collectively form a Hadoop ecosystem:

* HDFS: Hadoop Distributed File System
* YARN:  Yet Another Resource Negotiator
* MapReduce:  Programming based Data Processing
* Spark: In-Memory data processing
* PIG, HIVE: Query based processing of data services
* HBase: NoSQL Database
* Mahout, Spark MLLib: [Machine Learning](https://www.geeksforgeeks.org/machine-learning/)algorithm libraries
* Solar, Lucene: Searching and Indexing
* Zookeeper: Managing cluster
* Oozie: Job Scheduling

**NOTE**: HDFS, MapReduce and YARN already described above.

**PIG:**

 Pig was basically developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL.

* It is a platform for structuring the data flow, processing and analyzing huge data sets.
* Pig does the work of executing commands, After the processing, pig stores the result in HDFS.
* Pig Latin language is specially designed for this framework which runs on Pig Runtime. Just the way Java runs on the [JVM](https://www.geeksforgeeks.org/jvm-works-jvm-architecture/).

**HIVE:**

* HIVE performs reading and writing of large data sets. Its query language is called as HQL (Hive Query Language).
* It is highly scalable as it allows real-time processing and batch processing. All the SQL datatypes are supported by Hive.
* Similar to the Query Processing frameworks, HIVE too comes with two components: *JDBC Drivers* and *HIVE Command Line*.
* JDBC, along with ODBC drivers work on establishing the data storage permissions and connection whereas HIVE Command line helps in the processing of queries.

**Mahout:**

* Mahout, allows Machine Learnability to a system or application.
* It provides various libraries or functionalities such as filtering, clustering, and classification. It allows invoking algorithms as per our need with the help of its own libraries.

**Apache Spark:**

It’s a platform that handles all the processing tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc.

* Spark is best suited for real-time data whereas Hadoop is best suited for structured data or batch processing.

**Apache HBase**:   
It’s a NoSQL database which supports all kinds of data. It provides capabilities of Google’s BigTable, thus able to work on Big Data sets effectively.

* At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short quick span of time. At such times, HBase comes handy as it gives us a tolerant way of storing limited data

**Lucene:**

Lucene perform the task of searching and indexing with the help of some java libraries, However, Lucene is driven by Solr.

**Zookeeper:**

Zookeeper overcame huge issue of management of coordination and synchronization among the resources of Hadoop by performing synchronization, inter-component based communication, grouping, and maintenance.

**Oozie:**

Oozie simply performs the task of a scheduler, thus scheduling jobs and binding them together as a single unit.

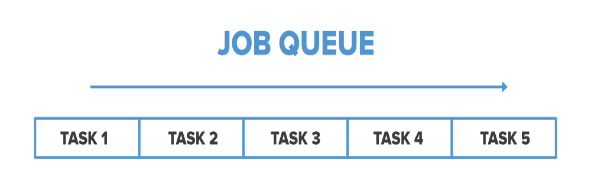
There is two kinds of jobs .i.e Oozie workflow and Oozie coordinator jobs.

**3)What is FIFO, Fair, Capacity Schedulers?**

There are mainly 3 types of Schedulers in Hadoop:

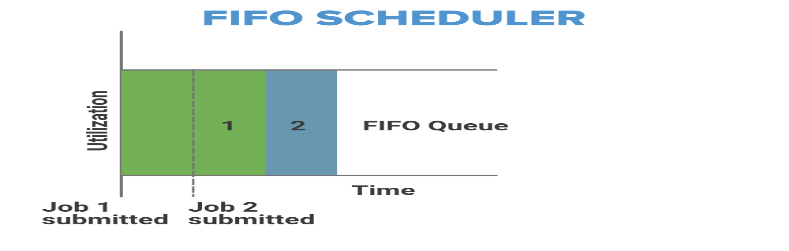
1. FIFO (First In First Out) Scheduler.
2. Capacity Scheduler.
3. Fair Scheduler.

A **Job queue** is nothing but the collection of various tasks that we have received from our various clients.



### 1)FIFO Scheduler

### As the name suggests FIFO i.e. First In First Out, so the tasks that comes first will be served first. This is the default Scheduler we use in Hadoop. In this method, once the job is scheduled, no intervention is allowed. So sometimes the high-priority process has to wait for a long time since the priority of the task does not matter in this method.

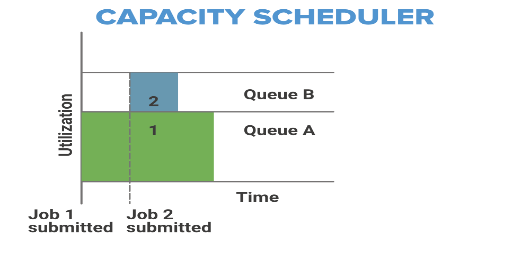
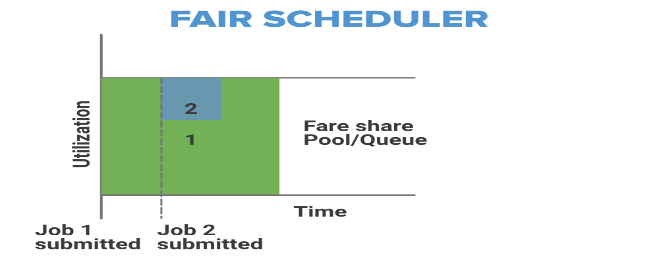


### 2. Capacity Scheduler

In Capacity Scheduler we have multiple job queues for scheduling our tasks. Each job queue has it’s own slots to perform its task.

In case we have tasks to perform in only one queue then the tasks of that queue can access the slots of other queues also as they are free to use.

The capacity Scheduler mainly contains 3 types of the queue that are root, parent, and leaf

### 3. Fair Scheduler

The Fair Scheduler is very much similar to that of the capacity scheduler.

The priority of the job is kept in consideration.

Fair Scheduler takes Scheduling decisions on the basis of memory, we can configure it to work with CPU also.

It is similar to Capacity Scheduler but the major thing is that in Fair Scheduler whenever any high priority job arises in the same queue, the task is processed in parallel by replacing some portion from the already dedicated slots.

**4)What is HIVE? And Explain.**

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

**HIVE is a interface between HDFS and User.**

**Components of Hive:**

1. **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 the grid easily.
2. **WebHCat –**   
   It provides a service which can be utilized by the user to run Hadoop MapReduce, Pig, Hive tasks or function Hive metadata operations with an HTTP interface.

**Modes of Hive:**

1. **Local Mode:**   
   It is used, when the Hadoop is built under pseudo mode which has only one data node.
2. **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.**

Pig is a high-level platform or tool which is used to process the large datasets

It provides a high-level scripting language, known as *Pig Latin Language.*

Pig Engine has two type of execution environment i.e. a *local execution environment* in a single JVM and *distributed execution environment* in a Hadoop Cluster.

**Features of Apache Pig:**

* Easy to learn, read and write. Especially for SQL-programmer, Apache Pig is a boon.
* Apache Pig is extensible so that you can make your own user-defined functions and process.
* Fewer lines of code.
* Pig can handle the analysis of both structured and unstructured data.

**Types of Data Models in Apache Pig:**

It consist of the 4 types of data models as follows:

* **Atom**: It is a atomic data value which is used to store as a string.
* **Tuple**: It is an ordered set of the fields.
* **Bag**: It is a collection of the tuples.
* **Map**: It is a set of key/value pairs.

### Modes:

### 1.Local Mode:

### In this mode of execution, we need a single machine and all files are installed and run using your localhost and file system.

* This mode is used for testing and development purposes. The local mode does not need HDFS or Hadoop.

**2. Mapreduce Mode:**

* Mapreduce is the default mode of the Apache Pig Grunt shell.
* In this mode, we need to load data in HDFS and then we can perform the operation.

## **Apache Pig Execution Methods**

A user can execute Apache Pig Latin scripts in three ways as mentioned below.

* **1. Interactive Mode (Grunt shell)**

In this mode, a user can interactively run Apache Pig using the Grunt shell. To invoke Grunt Shell, run Pig command. Users can submit commands and get a result there only.

### **Batch Mode (Script):** In this mode, a user can run Apache Pig in batch mode by creating a Pig Latin script file and running it from local or MapReduce mode.

### Embedded Mode: We can define our own functions called as UDF(Usser Defined Functions).