**Big data**

Due to evolution of the technology the data generation has increased so much,and so the data storage and processing has also become complex.

* IOT
* Social media
* Other factors

Big data is the term for collection of data sets so large and complex that it becomes difficult to process using on-hand database system tools or traditional data processing applications.

5V’s of Big data classification

* Volume – Defines the volume of the data in terms of size.
* Variety – Defines the various types of data
* Velocity – Defines the speed the data is being generated by different sources.
* Varacity – The inconsistency in the data that is generated during the process.
* Value – Deals about the value of the data.

**Big Data Analysis:**

The challenge was to deal the data storage and processing the data in order to generate useful business.

Hence came up with the concepts of the Big data Analysis which deals with the storage and processing.

* Cost effective big data storage
* Effective and fast way of processing data.
* Producing useful results from the data.

DATA ENGINEER

Starts with Data ends with Decision.

Gathering data:

* SFTP
* API

SFTP

Establishing a secured way to transfer the data from the application server to the destination server.

Shell script

API

Code that requests for the data from the database of the application for the data by API call , by which the data is pulled to the database.

Request library from python

Phase -2

Transforming data

* Audit column
* Time and date formatting

Phase -3

The process of writing logics based on business requirement.

Using pandas,Os library of python

Phase -4

The table loading from which the data can be used by the Power Bi or data scientists for drawing out the decision.

**Types of Data**

* Structured data
* Unstructured data
* Semi Structured data

**Structured data**

These can be seen as the data that is stored in proper table as in excel sheet.

These tables are managed by the RDBMS

**Unstructured data:**

The data that has unstructured data type such as the images,videos,voice records.

**Semi structured:**

The semi structured data contains both unstructured and structured data such as in xml files.

**Data formats:**

The different file formats in which the data can be stored.

CSV

JSON

PARAQUET

ORC

AVRO

Row based data format – JSON , CSV,AVRO

Column based – ORC,paraquet

**HADOOP ECOSYSTEM**

Problem : To store and process the Big data

Since there are many sources of data hence it is very important to effectively store and manage the data in more effective way to run a business.

**The Distributed Concept of storing data.**

The concept of Hadoop that deals with the solving the problems of storing and processing the big data.

The method of handling the data by breaking them into parts based on specific memory capacity that is predefined this method is called as HDFS (Hadoop distributed file system).

**HDFS**

The HDFS stands for the Hadoop distributed system,

This process divides the file from the single file which leads to the parallel processing of the files and increase in the speed of performance of the task.

It follows the replication method to increase the fault tolerance of data due to malfunctionality of the data node.

The multiple copies of the same data block will be kept in different nodes to recover the data in case of any losses.

It has 2 nodes

* Name node
* Data node

Name node stores

* Meta data
* Records of the permissions

Functionalities of name node

* Secondary node
* FS Image
* Editlog

Secondary Node- It keeps track of all the changes that are made in the name node

FSimage – It contains the complete state of the file system name space since the start of the name node

Editlogs – It contains the recent modifications made to the file system with respect to the most recent Fsimage.

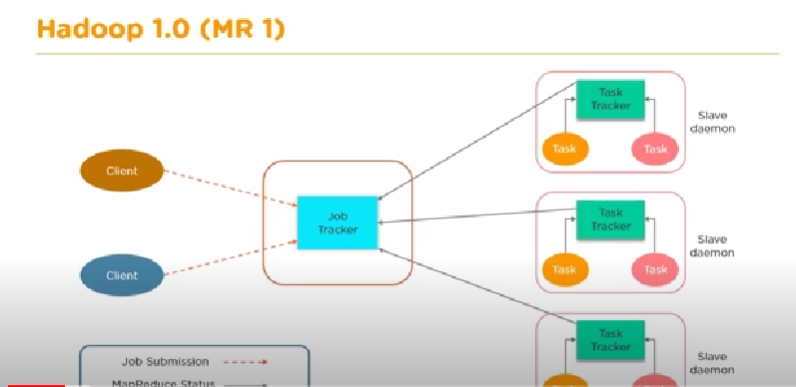
Heart beat - The data node sends the heart beat to name node every 3 seconds to ensure that it is functioning well

Data node

* Contains the actual data
* Sends the heart beats to the name node

**YARN**

**Yet another resource negotiator**



This image shoes the Hadoop version 1-

* Job tracker
* Task Tracker

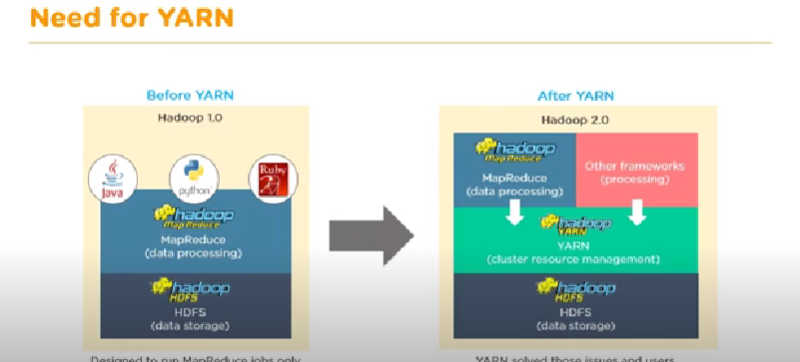
The job tracker had the Functionalities of –

Managing the resources

Allocating the tasks to the task tracker

Limitation:

* Scalability
* Availability issue
* Resource utilization
* Limitations in running non map reduce tasks like graph methods



**Components of YARN**

**Resource manager**

* **Scheduler**
* **Application Manager**

**Node Manager**

* **Container**
* **Application master**

Scheduler

* Responsible for allocating resources to various running applications
* Does not perform monitoring or tracking of status for the applications

Applications Manager

* Responsible for accepting job submissions
* Negotiating the first container for executing the application specific applications master

Node Manager

Contains the container that contains the resources such as cpu, memory

App master runs the task and submits it to the app master in resource manager.

The work flow

1. Client submits the task to the resource manager
2. The resource manager has the scheduler and app manager to negotiage with the container and allocate the containers from the node manager
3. The app master utilizes the container to perform the tasks and then submits it to the app manager

The node manager constantly sends the heart beat to the resource manager regarding the availability of the memory.

**Map Reduce:**

**Map reduce performs the processing of large data sets in distributed and parallel manner.**

It consists of two process:

* Input
* Splitting
* Mapping
* Shuffling
* Reducing
* Final result

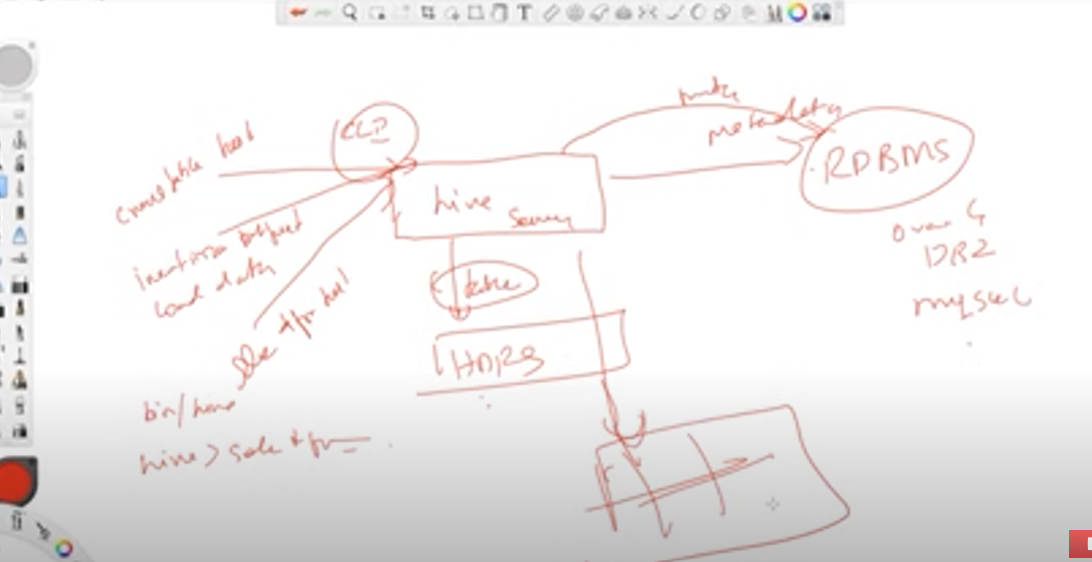
**HIVE**

Hive is a data ware house system which is used for quering and analysing large datasets stored in HDFS.

Hive do not have any database, it uses the HDFS for storing.

It uses the map reduce but uses sql for quering instead of using the java for doing the process.

Hive stores the meta data in RDBMS format



Hive always stores the meta data in RDBMS, it has an embedded database called derbee called remote meta store.

Using of embedded Derbee will raise a problem when you will be using the multiple node,since each time the request will be going to different nodes so it would be difficult to work with.