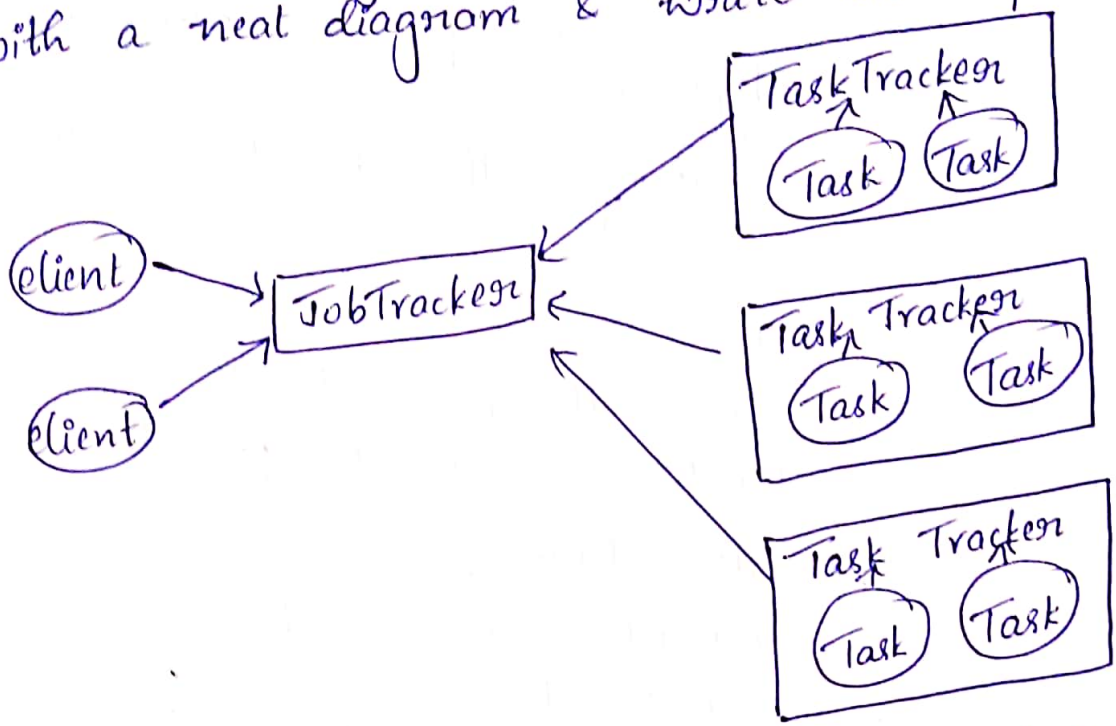
7] a] Illustrate the working model of Mapreduce programming with a neat diagram & write the steps.



1. Input dataset is split into multiple pieces of data.
2. Framework creates master & worker processes & executes them remotely.
3. Map function uses algorithm to extract only those data that are present on their server & generates key/value pair.
4. Map worker uses partition function to decide which reduce worker should get output of specified mapper.
5. Reducer in turn contacts mapper that provides its output which are shuffled & sorted.
6. Reducer function is called for every unique key.
7. The master transfers control to the user progran

b) List & explain 3 major components of YARN.

1. Resource manager:
   → Manages resource for scheduling different comp. ing applications.
   → Co-ordinating with scheduler & Application manager.

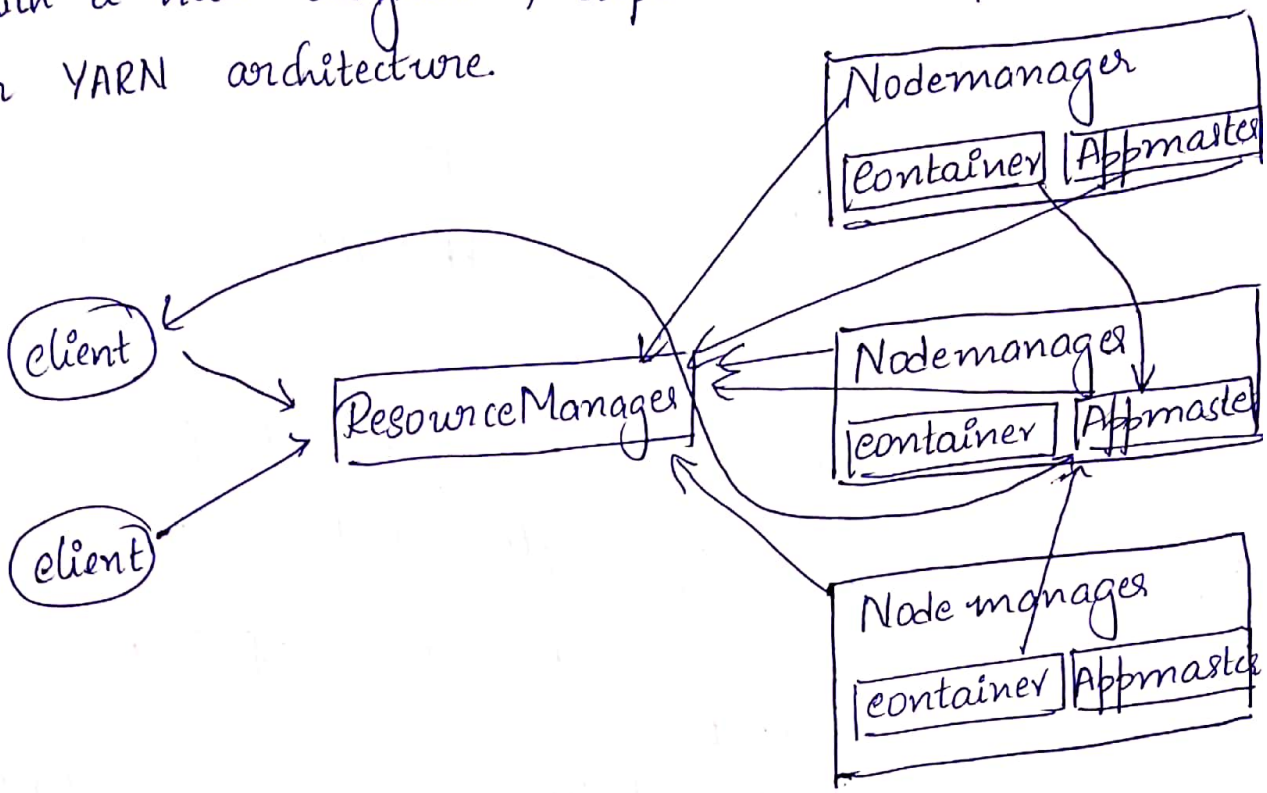
   Scheduler:
   → Schedules the jobs submitted.
   → allocates resources to applications.

   Application Manager:
   → Accept job submission from client.
   → Negotiates first container for executing application specific task with suitable application master.

2. Node Manager:
   → Managing & Executing containers.
   → sending heartbeats to resource manager.

3. Application Master:
   → Negotiating suitable resource container or slave node from resource manager.
   → Working with Nodemanager to execute task.

a) With a neat diagram, explain the steps involved in YARN architecture.



1. A client program submits the application which includes specifications to launch application specific app-master.
2. Resource manager launches application master by assigning suitable container.
3. Application master, on boot up registers with Resource Manager.
4. Application master negotiates appropriate resource container via resource-request protocol.
5. Application master launches container by sending container launch specification to node-manager.

6. Node manager executes application & provi statuses.

7. During application execution, client can direct communicate with application master to get the status.

8. Application master detaches itself from resource manager after task completion.

b) List & explain the applications of MapReduce.

1. Distributed GREP:
   → used to find pattern in large no. of files.
   → map function takes input as (inputfile, line) & generate a key, value pair if a match is found.

2. Geospatial Query processing:
   Google Maps uses this to solve problems like.
   → given an intersection, find connecting roads to it
   → Rendering tiles on map.

3. LIDAR data.

Local gridding algorithm utilizes elevation information from LIDAR measurements to compute elevation of each grid.

Q.10

a) Write a short note on piglatin statements & identifiers with examples.

1. Pig Latin statements are basic constructs to process data using Pig.

2. Pig Latin statement is an operator.

3. An operator in PigLatin takes a relation as input & yields another relation as output.

4. Pig Latin statements include schemas & expressions to process data.

5. Pig Latin statements end with semi colons.

Pig Latin script:

1. LOAD statement reads data from file system.
2. DUMP or STORE to display/store result.

Ex:-

```
A = load 'student' (rollno, name, gpa);
A = filter A by gpa > 4.0;
STORE A INTO 'myreport'.
```

Pig Latin identifiers:

Identifiers should begin with a letter & should be followed by letters, numbers, & underscores.

Valid Identifier	Y	A1	A1_2014	Sample
Invalid Identifier	5	Sales$	Sales%	_sales.

b) Briefly explain RC file format with example.

RCFile stores data in Column Oriented Manner which ensures that Aggregation operation is not expensive. RCFile partitions the table first horizontally & then vertically to serialize data.

Table with 4 columns.

C1	C2	C3	C4
11	12	13	14
21	22	23	24
31	32	33	34
41	42	43	44
51	52	53	54

Table with 2 groups.

Row Group 1

C1	C2	C3	C4
11	12	13	14
21	22	23	24
31	32	33	34

Row Group 2

C1	C2	C3	C4
41	42	43	44
51	52	53	54

Table in RCFile format.

RowGroup1

11 21 31;
12 22 32;
13 23 33;
14 24 34;

RowGroup2.

41 51;
42 52;
43 53;
44 54;

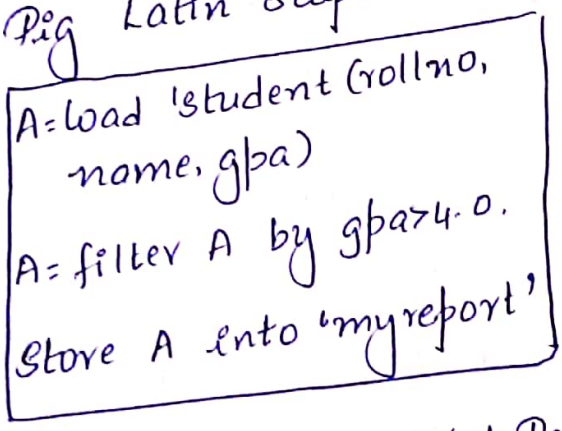
List out the key features of pig. Explain anatomy of pig.

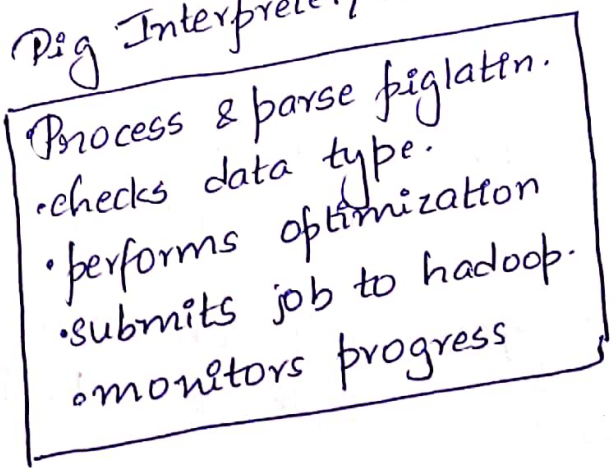
1. It provides a language called "PigLatin" to express data flows.
2. It allows users to develop their own functions.
3. Pig Latin contains operators for many of traditional data operations such as join, sort etc.

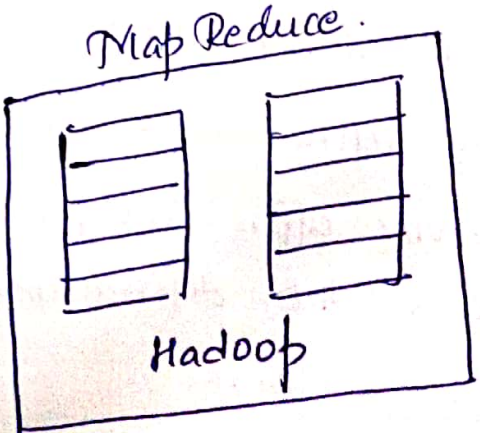
Anatomy of Pig:

Components of Pig:

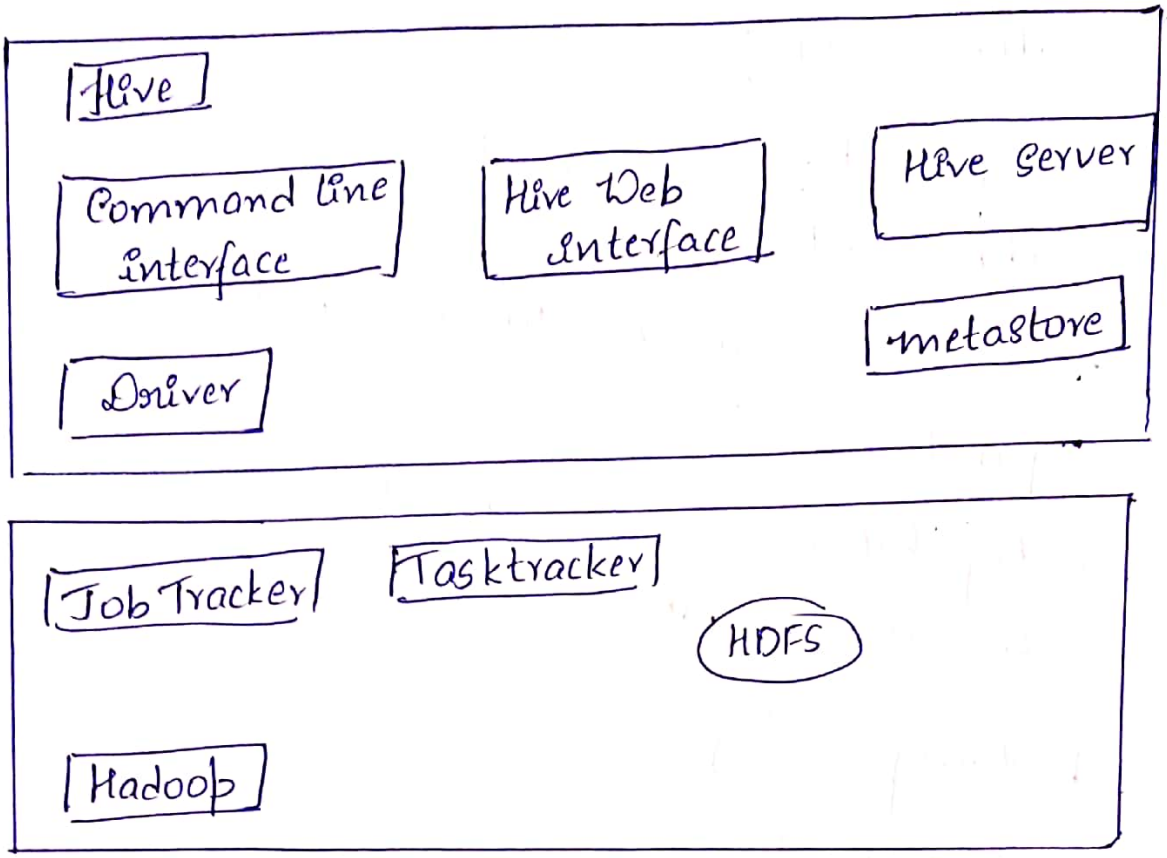
1. Data flow Language. (Pig Latin)
2. Interactive Shell (Grunt)
3. Pig interpreter & Execution Engine.







b) Illustrate hive architecture with a neat diagram.



Command line interface: commonly used interface to interact with hive.

Hive Web interface: Graphical user interface used to execute queries.

Hive Server: optional server to submit to submit hive jobs from remote client.

JDBC/ODBC: one can write java code to connect to hive & submit jobs on it.

Driver: Hive queries are sent to driver for compilation, optimization & execution

Metastore: Contains definations & mappings to the data.

- Metastore service: offers interface to hive.
- Databases: Stores data definations, mappings to the data & others.