Hadoop Hive Mini Project:

# Assignment:

**Analyse Employee Dataset using Hadoop**

1. Using Hadoop command move all those employees data into HDFS directory "/user/your\_user\_name/employees\_data" directory
2. Create an external Hive table "employees\_Table" representing this "employees\_data". This table will have 4 fields id,age,gender,role and salary.
3. create a new bucketed table "Consultant\_Table\_Bucket" having 4 buckets on the field salary. This table should store the data into columnar format ORC
4. Insert all those employees whose salary is greater than 5000 into bucketed table "Consultant\_Table\_Bucket" from "employees\_Table" table. While inserting into "Consultant\_Table\_Bucket" table you need to convert "consultant" role into "BigData Consultant" role.

  6. Write a Hive query to find out Max, min salary of "BigData Consultant" from the  "Consultant\_Table\_Bucket" table

# Solution:

**Step 1:** *Creation of files in HDFS file system*

hdfs dfs -put Consultantdata.txt /user/arun.kpselvam\_gmail/employees\_data

**Step 2:** *Creating new database in hive*

create database jun17d;

**Step 3:** *Using existing database*

use jun17d;

**Step 4:** *Creation of external table without partition*

CREATE EXTERNAL TABLE employees\_Table(

id INT,

age INT,

gender STRING,

role STRING,

salary BIGINT

)

ROW FORMAT delimited fields terminated BY '|'

STORED AS textfile

LOCATION '/user/arun.kpselvam\_gmail/emp1d';

**Step 5:** *Loading data from consultants.txt in to employees\_data table*

load data inpath '/user/arun.kpselvam\_gmail/employees\_data' into table employees\_Table;

**Step 6:** *Enabling Bucketing feature*

SET hive.enforce.bucketing=true;

**Step 7:** *Creation of bucketing table*

CREATE TABLE Consultant\_Table\_Bucket (

id INT,

age INT,

gender STRING,

role STRING,

salary INT)

clustered by (salary) into 4 buckets

ROW FORMAT delimited fields terminated BY '|'

STORED AS orc;

**Step 8:** *Insertion of employee data based on conditions*

INSERT OVERWRITE TABLE Consultant\_Table\_Bucket

SELECT id, age, gender, CASE WHEN role = 'consultant' THEN 'BigData Consultant' ELSE role END as role,salary

FROM employees\_Table

WHERE salary > 5000;

**Step 9:** *Query to fetch maximum salary value for big data consultant*

SELECT max(salary)

FROM Consultant\_Table\_Bucket

WHERE role = 'BigData Consultant';

**Step 10:** *Query to fetch maximum salary value for big data consultant*

SELECT min(salary)

FROM Consultant\_Table\_Bucket

WHERE role = 'BigData Consultant';

# Results:

