NAME: ADITYA GUPTA

STUDENT ID: 19210195

ASSIGNMENT: 1

**STEP 1: Downloaded the data from stackexchange website.**

**STEP 2: Uploaded the data to GCP Cluster with master node. Then took SSH.**

**STEP 3: Loaded data with pig using below commands :**

REGISTER /usr/lib/pig/piggybank.jar

DEFINE CSVLoader org.apache.pig.piggybank.storage.CSVLoader();

data1 = LOAD '/96850.csv' using org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'NOCHANGE','SKIP\_INPUT\_HEADER') as (Id:int, PostTypeId:chararray, AcceptedAnswerId:chararray, ParentId:chararray, CreationDate:chararray, DeletionDate:chararray, Score:int, ViewCount:int, Body:chararray, OwnerUserId:chararray, OwnerDisplayName:chararray, LastEditorUserId:chararray, LastEditorDisplayName:chararray, LastEditDate:chararray, LastActivityDate:chararray, Title:chararray, Tags:chararray, AnswerCount:chararray, CommentCount:chararray, FavoriteCount:chararray, ClosedDate:chararray, CommunityOwnedDate:chararray);

data2 = LOAD '/57000-96850.csv' using org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'NOCHANGE','SKIP\_INPUT\_HEADER') as (Id:int, PostTypeId:chararray, AcceptedAnswerId:chararray, ParentId:chararray, CreationDate:chararray, DeletionDate:chararray, Score:int, ViewCount:int, Body:chararray, OwnerUserId:chararray, OwnerDisplayName:chararray, LastEditorUserId:chararray, LastEditorDisplayName:chararray, LastEditDate:chararray, LastActivityDate:chararray, Title:chararray, Tags:chararray, AnswerCount:chararray, CommentCount:chararray, FavoriteCount:chararray, ClosedDate:chararray, CommunityOwnedDate:chararray);

data3 = LOAD '/40900-57000.csv' using org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'NOCHANGE','SKIP\_INPUT\_HEADER') as (Id:int, PostTypeId:chararray, AcceptedAnswerId:chararray, ParentId:chararray, CreationDate:chararray, DeletionDate:chararray, Score:int, ViewCount:int, Body:chararray, OwnerUserId:chararray, OwnerDisplayName:chararray, LastEditorUserId:chararray, LastEditorDisplayName:chararray, LastEditDate:chararray, LastActivityDate:chararray, Title:chararray, Tags:chararray, AnswerCount:chararray, CommentCount:chararray, FavoriteCount:chararray, ClosedDate:chararray, CommunityOwnedDate:chararray);

data4 = LOAD '/32000-40900.csv' using org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'NOCHANGE','SKIP\_INPUT\_HEADER') as (Id:int, PostTypeId:chararray, AcceptedAnswerId:chararray, ParentId:chararray, CreationDate:chararray, DeletionDate:chararray, Score:int, ViewCount:int, Body:chararray, OwnerUserId:chararray, OwnerDisplayName:chararray, LastEditorUserId:chararray, LastEditorDisplayName:chararray, LastEditDate:chararray, LastActivityDate:chararray, Title:chararray, Tags:chararray, AnswerCount:chararray, CommentCount:chararray, FavoriteCount:chararray, ClosedDate:chararray, CommunityOwnedDate:chararray);

Merge = UNION data1, data2, data3, data4;

F = FOREACH Merge generate Id,PostTypeId,Score,ViewCount,REPLACE(REPLACE(Body, '\n',''),',','') As Body, OwnerUserId,OwnerDisplayName;

Store F into '/Output/' using PigStorage(',');

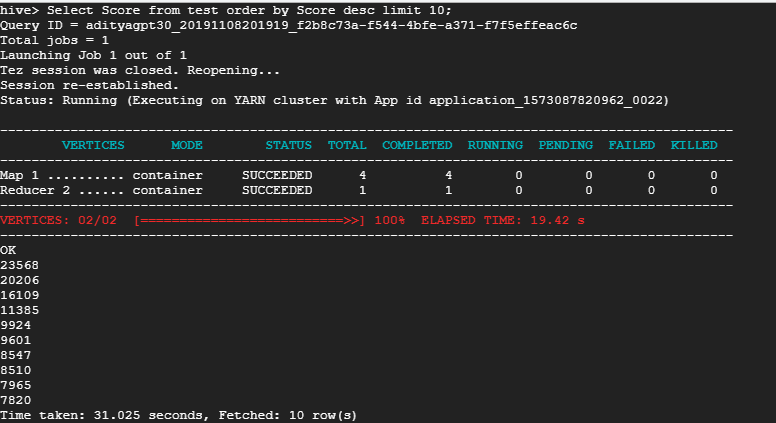
**STEP 4: Then using HIVE**

create external table test (Id int,PostTypeId int,Score int,ViewCount int,Body string,OwnerUserId int,OwnerDisplayName string) row format delimited fields terminated by ',';

LOAD DATA INPATH '/Output' OVERWRITE INTO TABLE test;

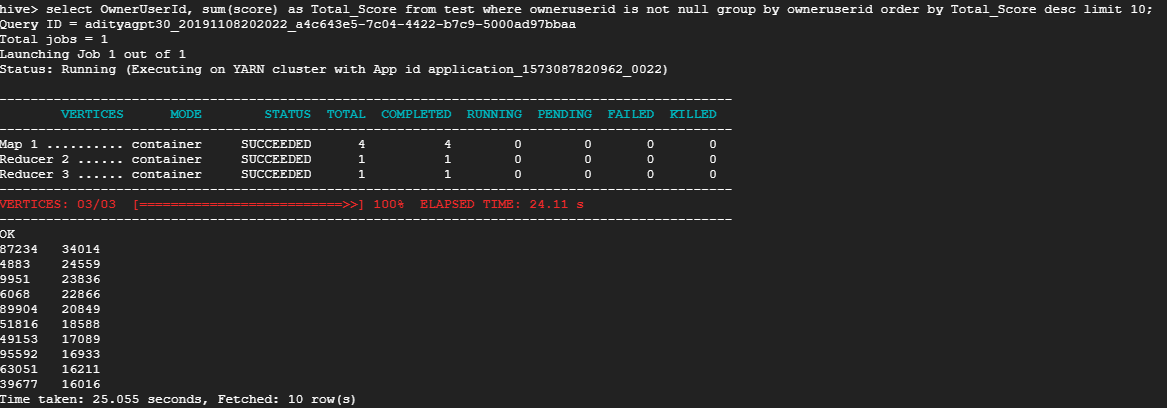
**4.1: The top 10 posts by score:**

hive> Select Score from test order by Score desc limit 10;



**4.2: The top 10 users by post score:**

select OwnerUserId, sum(score) as Total\_Score from test where owneruserid is not null group by owneruserid order by Total\_Score desc limit 10;



**4.3: The number of distinct users, who used the word “Hadoop” in one of their posts:**

select count(distinct (owneruserid)) from test where body like '%hadoop%';

