**STEP 2: TRANSFORM DATA BY USING PIG**

**Upload all the 5 csv files obtained from Stack Exchange into the local file system.**

hadoop fs -mkdir /hadoop6 **/\*create a new directory in hadoop file system\*/**

hadoop fs -put QueryResults1.csv QueryResults2.csv QueryResults3.csv QueryResults4.csv QueryResults5.csv /hadoop6**/\*Transfer all the csv files to hadoop using the -put command\*/**

hadoop fs -ls /hadoop6**/\*check list of files in hadoop after uploading\*/**

hadoop fs -mkdir /hadoop7

**/\*Load each csv file into separate variables using CSVExcelStorage\*/**

load\_csv11 = LOAD '/hadoop6/QueryResults1.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') as (Id:int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:datetime,DeletionDate:datetime,Score:int,ViewCount:int,Body:chararray,OwnerUserId:int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:chararray,LastEditDate:datetime,LastActivityDate:datetime,Title:chararray,Tags:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:datetime,CommunityOwnedDate:datetime);

load\_csv12 = LOAD '/hadoop6/QueryResults2.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') as (Id:int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:datetime,DeletionDate:datetime,Score:int,ViewCount:int,Body:chararray,OwnerUserId:int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:chararray,LastEditDate:datetime,LastActivityDate:datetime,Title:chararray,Tags:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:datetime,CommunityOwnedDate:datetime);

load\_csv13 = LOAD '/hadoop6/QueryResults3.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') as (Id:int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:datetime,DeletionDate:datetime,Score:int,ViewCount:int,Body:chararray,OwnerUserId:int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:chararray,LastEditDate:datetime,LastActivityDate:datetime,Title:chararray,Tags:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:datetime,CommunityOwnedDate:datetime);

load\_csv14 = LOAD '/hadoop6/QueryResults4.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') as (Id:int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:datetime,DeletionDate:datetime,Score:int,ViewCount:int,Body:chararray,OwnerUserId:int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:chararray,LastEditDate:datetime,LastActivityDate:datetime,Title:chararray,Tags:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:datetime,CommunityOwnedDate:datetime);

load\_csv15 = LOAD '/hadoop6/QueryResults5.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') as (Id:int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:datetime,DeletionDate:datetime,Score:int,ViewCount:int,Body:chararray,OwnerUserId:int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:chararray,LastEditDate:datetime,LastActivityDate:datetime,Title:chararray,Tags:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:datetime,CommunityOwnedDate:datetime);

load\_merge = UNION load\_csv11,load\_csv12,load\_csv13,load\_csv14,load\_csv15; **/\*Combine the data from the 5 csv files into load\_merge using UNION command\*/**

Store\_Data1 = FOREACH load\_merge generate Id,PostTypeId,Score,ViewCount,REPLACE(REPLACE(Body, '\n',''),',','') As text,OwnerUserId,OwnerDisplayName;

**/\*Clean the data present in Store\_Data1 by replacing commas(‘,’) , new line character present in the body column\*/**

STORE Store\_Data1 INTO 'hadoop7/' USING PigStorage(',');

**/\*Store the cleaned data from the 5 csv files present in Store\_Data1 into hadoop7 directory using STORE command\*/**

**STEP 3 HIVE QUERIES**

CREATE DATABASE CloudAssignment1;

create external table cloud\_assignmentt(id int,PostTypeId int,Score int,ViewCount int,Body string,OwnerUserId int,OwnerDisplayName string) row format delimited fields terminated by ',';

load data inpath '/hadoop7/\*' into table cloud\_assignmentt;

**/\*Create a table cloud\_assignmentt in hive terminal and load the cleaned records from /hadoop7 \*/**

3.I. **The top 10 posts by score**

SELECT Id,Score from cloud\_assignmentt SORT BY Score DESC LIMIT 10;

3.II. **The top 10 users by post score**

SELECT OwnerUserId,OwnerDisplayName,sum(Score) as TOTAL from cloud\_assignmentt GROUP BY OwnerUserId,OwnerDisplayName ORDER BY TOTAL DESC LIMIT 10;

3.III. **The number of distinct users, who used the word “Hadoop” in one of their posts**

select COUNT (DISTINCT OwnerUserId) from cloud\_assignmentt where lower (Body) like ‘%hadoop%’;