1) **No of users of each states?**

users = LOAD '/user/cloudera/users.csv' USING PigStorage (',') AS (user\_id:chararray, name:chararray, state:chararray);

users\_group = Group users by state;

count\_user = foreach users\_group Generate FLATTEN(users.state) , COUNT(users.user\_id) as cnt;

result = DISTINCT count\_user;

dump result;

**2)Top 5 states with number of users?**

users = LOAD '/user/cloudera/users.csv' USING PigStorage (',') AS (user\_id:chararray, name:chararray, state:chararray);

users\_group = Group users by state;

count\_user = foreach users\_group Generate FLATTEN(users.state) , COUNT(users.user\_id) as cnt;

count\_user\_flat = DISTINCT count\_user;

result = ORDER count\_user\_flat BY cnt DESC;

finalresult = LIMIT result 5;

Dump finalresult;

**3) tweets that containword "Happy" ordered by tweet\_id?**

users = LOAD '/user/cloudera/tweets.csv' USING PigStorage (',') AS (tweet\_id:chararray, tweet:chararray, username:chararray);

users\_happy = filter users By tweet matches '.\*Happy.\*';

result = foreach users\_happy generate tweet,tweet\_id;

final\_result = order result by tweet\_id;

Dump final\_result;

**4) No.of users in CA with more than 1 tweet?**

tweets = LOAD '/user/cloudera/tweets.csv' USING PigStorage (',') AS (tweet\_id:chararray, tweet:chararray, username:chararray);

users = LOAD '/user/cloudera/users.csv' USING PigStorage (',') AS (user\_id:chararray, name:chararray, state:chararray);

tw\_us = join tweets by username, users by user\_id;

tw\_us\_CA = FILTER tw\_us by state == 'CA';

Grp = group tw\_us\_CA by username;

res = foreach Grp generate flatten(tw\_us\_CA.username), COUNT(tw\_us\_CA.tweet)as cnt;

result = distinct res;

final\_result = FILTER result BY cnt>1;

DUMP final\_result;

**5) Top 20 most active users?**

tweets = LOAD '/user/cloudera/tweets.csv' USING PigStorage (',') AS (tweet\_id:chararray, tweet:chararray, username:chararray);

users = LOAD '/user/cloudera/users.csv' USING PigStorage (',') AS (user\_id:chararray, name:chararray, state:chararray);

t\_u = join tweets by username, users by user\_id;

t\_u\_g = group t\_u by username;

res = foreach t\_u\_g generate (t\_u.username) , COUNT(t\_u.tweet)as cnt;

result = FILTER res by cnt > 1;

finresult = DISTINCT result;

DUMP finresult;

**6) No. of users that posted no tweets?**

tweets = LOAD '/user/cloudera/tweets.csv' USING PigStorage (',') AS (tweet\_id:chararray, tweet:chararray, username:chararray);

users = LOAD '/user/cloudera/users.csv' USING PigStorage (',') AS (user\_id:chararray, name:chararray, state:chararray);

total = join tweets by username, users by user\_id;

res = group total by username;

result = foreach res generate flatten(total.username), COUNT(total.tweet)as cnt;

finres = FILTER result BY cnt ==0;

DUMP finres;

**7)custom Udf to replace "Happy with "Sad" in all tweets?**

StringUDF.java:

package StringReplace;

import java.io.IOException;

import org.apache.pig.EvalFunc;

import org.apache.pig.data.Tuple;

public class StringUDF extends EvalFunc<String>{

public String exec(Tuple input) throws IOException{

if(input == null || input.size() == 0)

return null;

String str = (String)input.get(0);

String myOutput = str.replaceAll("Happy", "Sad");

return myOutput;

}

}