Big Data Hadoop Project submission on Social Media

**Question – 1 Top 10 most commonly used tags in this data set**

**Solution – 1**

**Pig Script**

questions\_data = LOAD '/user/biadmin/Social/Social\_Media.csv’ USING PigStorage('\_') AS (c1:chararray,c2:chararray,c3:chararray,c4:chararray,tags:chararray,c6:chararray,c7:chararray,c8:chararray,c9:chararray,c10:chararray,c11:chararray);

tags\_data = FOREACH questions\_data GENERATE tags;

tags\_seperated = FOREACH tags\_data GENERATE TOKENIZE(tags);

tags = FOREACH tags\_seperated GENERATE FLATTEN($0) AS tag;

tags\_group = GROUP tags BY tag;

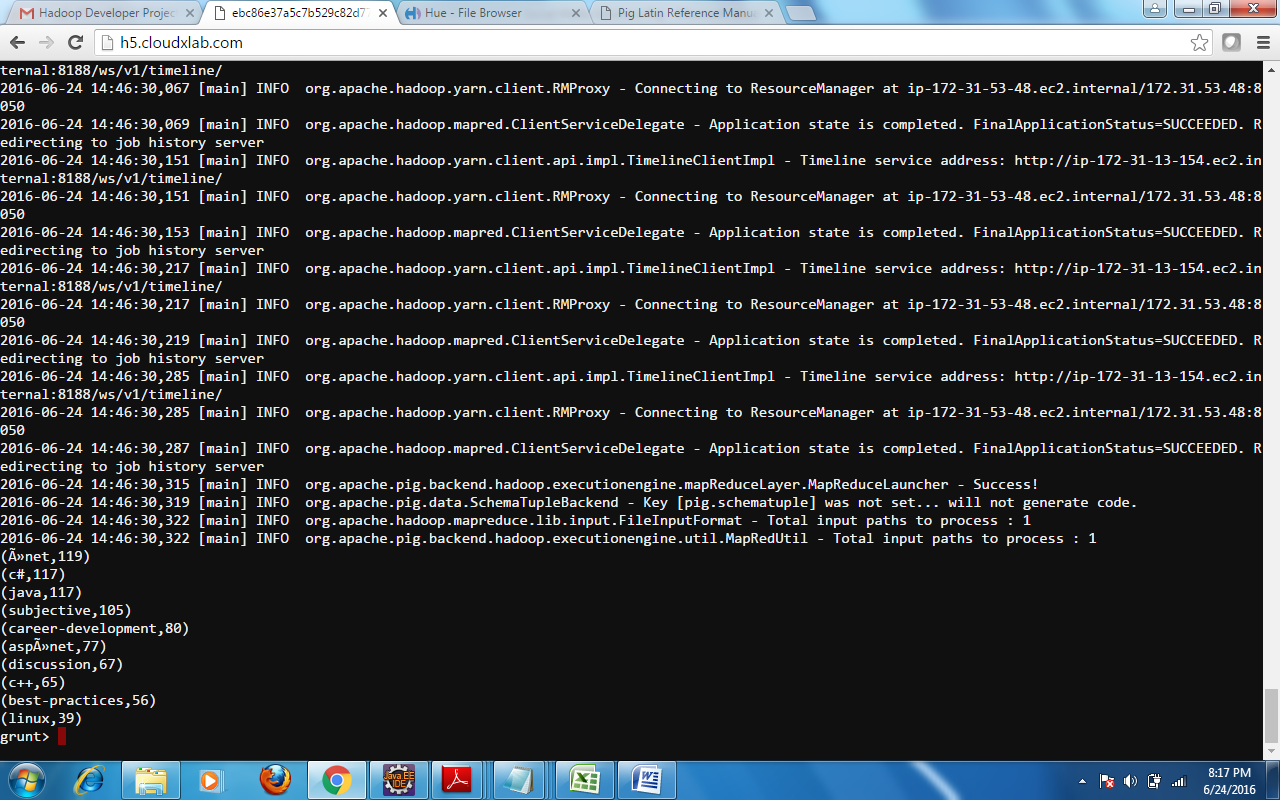
tags\_count = FOREACH tags\_group GENERATE group, COUNT(tags) as cnt;

tags\_sorted = ORDER tags\_count BY cnt DESC;

top10\_tags = LIMIT tags\_sorted 10;

DUMP top10\_tags;

**Output –**



**Question – 2 Average time to answer questions**

**Solution – 2**

**Pig Script –**

questions\_data = LOAD '/user/biadmin/Social/Social\_Media.csv’ USING PigStorage('\_') AS (qid:chararray,c2:chararray,c3:chararray,c4:chararray,c5:chararray,c6:chararray,c7:chararray,c8:chararray,c9:chararray,c10:chararray,at:long);

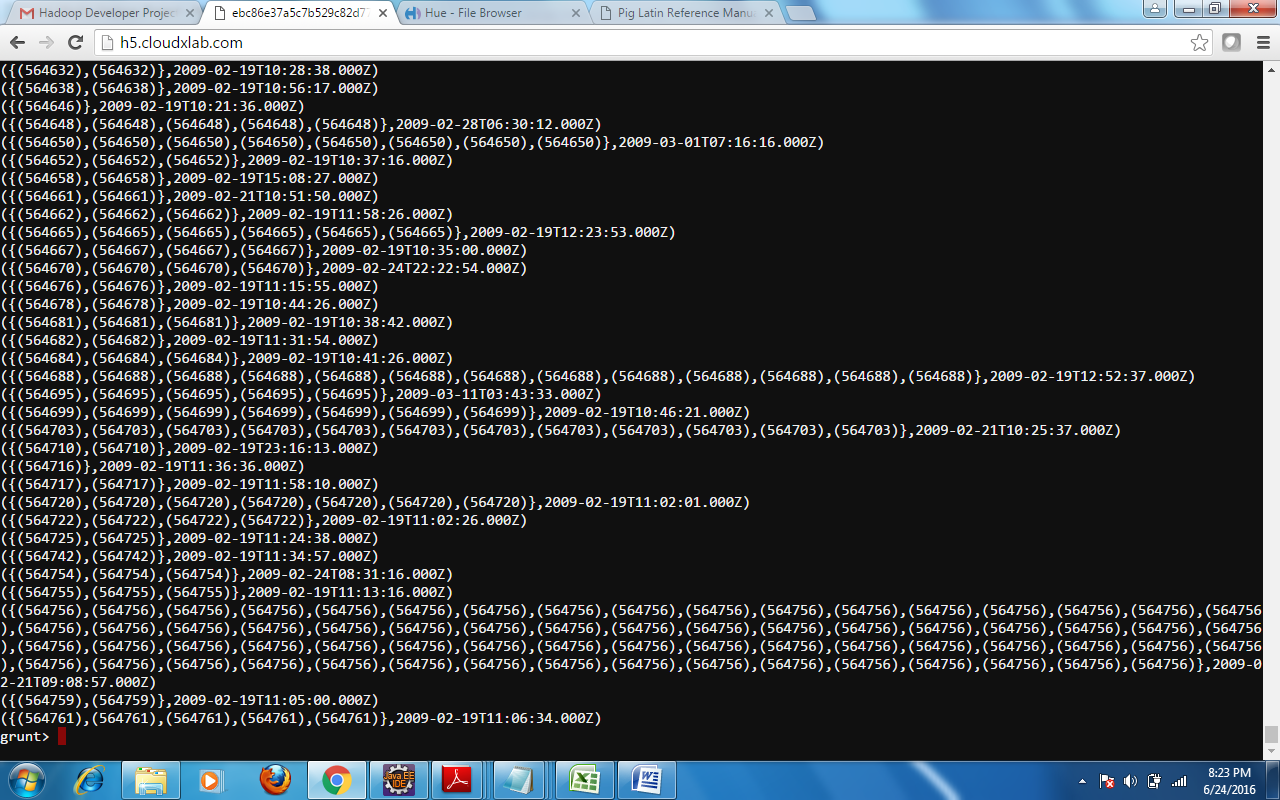
group\_qid = GROUP questions\_data BY qid;

avg\_qtime = FOREACH group\_qid GENERATE group, questions\_data.qid as qid, AVG(questions\_data.at) as avg;

avg\_qtime1 = FOREACH avg\_qtime GENERATE qid, ToDate((long)avg\*1000) as avg\_ans\_time;

DUMP avg\_qtime1;

**Output –**



**Question – 3 Number of questions which got answered within 1 hour** '/user/ebc86e37a5c7b529c82d77dbb24392/VinayT/Datafiles/answers.txt'

**Solution – 3**

**Pig Script –**

questions\_data = LOAD '/user/biadmin/Social/Social\_Media.csv’ USING PigStorage('\_') AS (qid:chararray,c2:chararray,c3:chararray,c4:chararray,c5:chararray,c6:chararray,c7:chararray,c8:chararray,c9:chararray,c10:chararray,at:long);

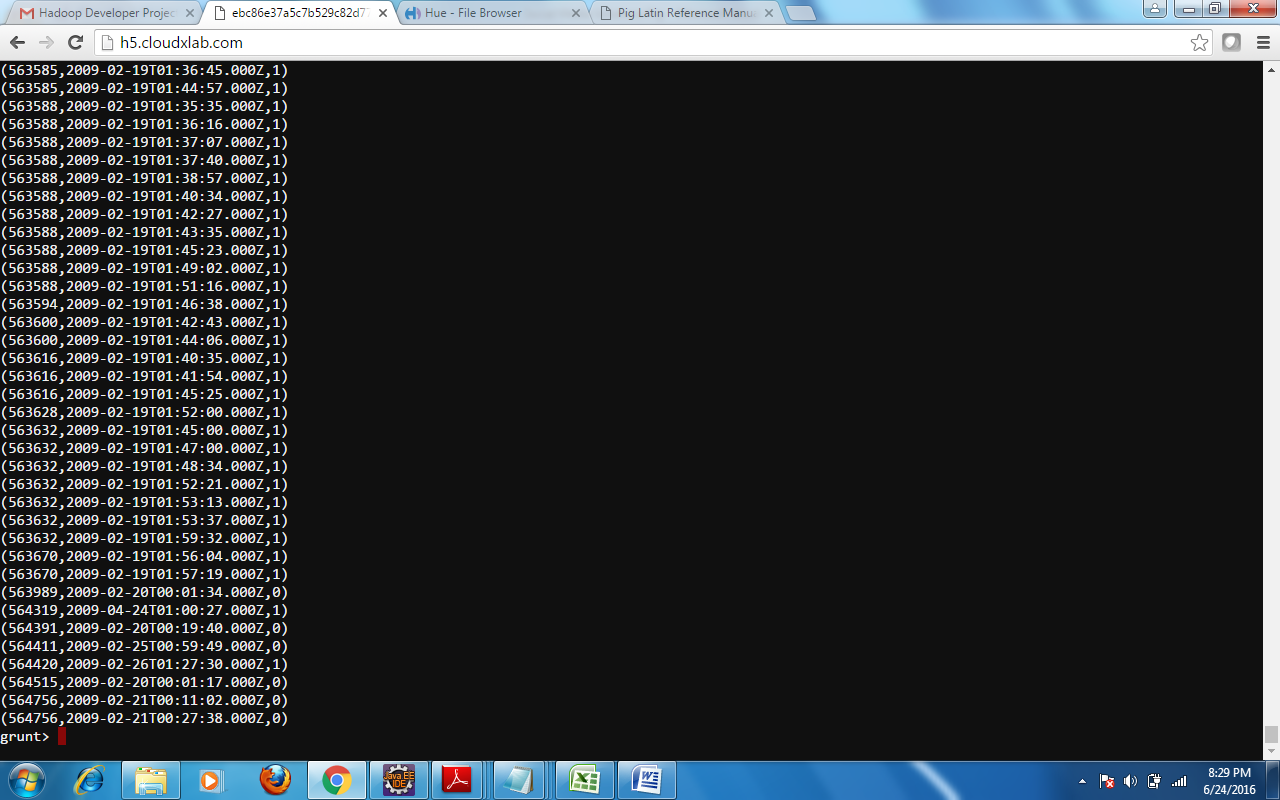
qid\_at = FOREACH questions\_data GENERATE qid as q, ToDate((long)at\*1000) as time;

qid\_gethour = FOREACH qid\_at GENERATE q as q, time as time, GetHour(time) as hour;

qid\_hourless1 = FILTER qid\_gethour by hour <= 1;

DUMP qid\_hourless1;

Output –



**Question – 4 Tags of questions which got answered within 1 hour**

**Solution – 4**

**Pig Script –**

questions\_data = LOAD '/user/biadmin/Social/Social\_Media.csv’ USING PigStorage('\_') AS (qid:chararray,c2:chararray,c3:chararray,c4:chararray,tags:chararray,c6:chararray,c7:chararray,c8:chararray,c9:chararray,c10:chararray,at:chararray);

tags\_data\_qid\_time = FOREACH questions\_data GENERATE tags, qid as q, ToDate((long)at\*1000) as time;

tags\_seperated\_with\_hour = FOREACH tags\_data\_qid\_time GENERATE TOKENIZE(tags), q as q, GetHour(time) as hour;

tags\_flatten = FOREACH tags\_seperated\_with\_hour GENERATE FLATTEN($0) AS tag, q as q, hour as hour;

tags\_lessthanonehour = FILTER tags\_flatten by hour <= 1;

tags\_order = ORDER tags\_lessthanonehour by tag;

DUMP tags\_order;

**Output –**

