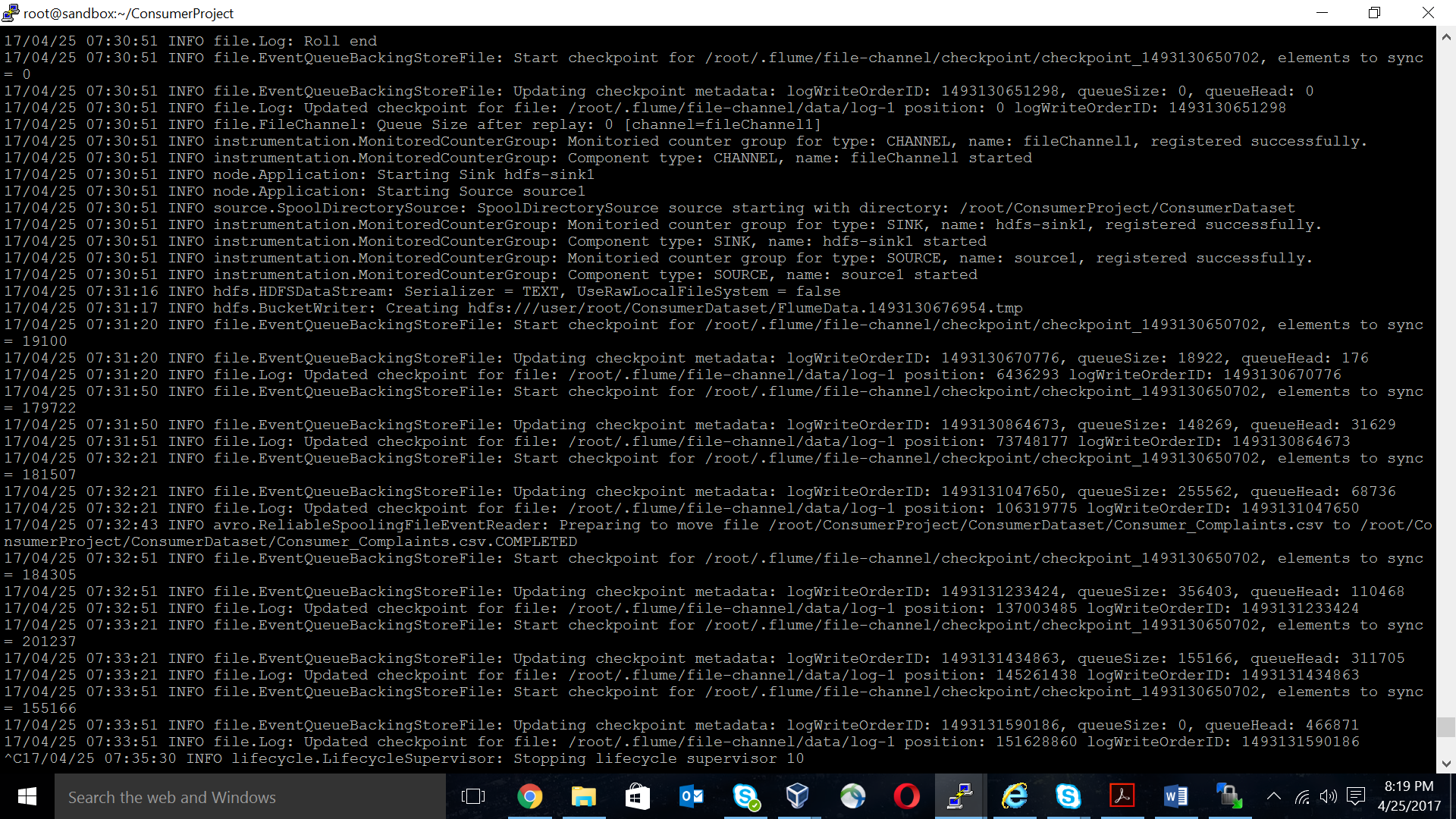
**SESSION 21-24: PROJECT II**

**USA Consumer Forum Data Analysis**

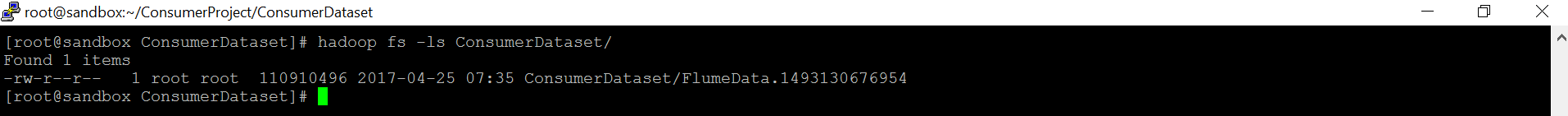
**Transferring Data into HDFS using Flume:**

A spool directory is setup at /root/ConsumerProject/ConsumerDataset. Flume agent is setup in configuration file /root/ConsumerProject/ConsumerFlume.conf. The spool directory will transfer the dataset from local file system to HDFS (/user/root/ConsumerDataset).

flume-ng agent –n agent2 –f /root/ConsumerProject/ConsumerFlume.conf

The above command creates FlumeData file at /user/root/ConsumerDataset in HDFS.





**Processing Data through PIG:**

**The aim of this project is to analyze performance of various companies on aspects like:**

-- Common commands for all problems

REGISTER /usr/local/pig/lib/piggybank.jar

data = LOAD '/user/root/ConsumerDataset/' USING

org.apache.pig.piggybank.storage.CSVExcelStorage() AS (rdate:chararray,

product:chararray, subproduct:chararray, issue:chararray, subissue:chararray,

csnarrative:chararray, cpresponse:chararray, company:chararray, state:chararray,

zip:chararray, submitvia:chararray, sdate:chararray, status:chararray,

tresponse:chararray, csdisputed:chararray, complaintID:chararray);

1. **Write a pig script to find number of complaints which got timely response.**

filter\_data = FILTER data BY (tresponse == 'Yes');

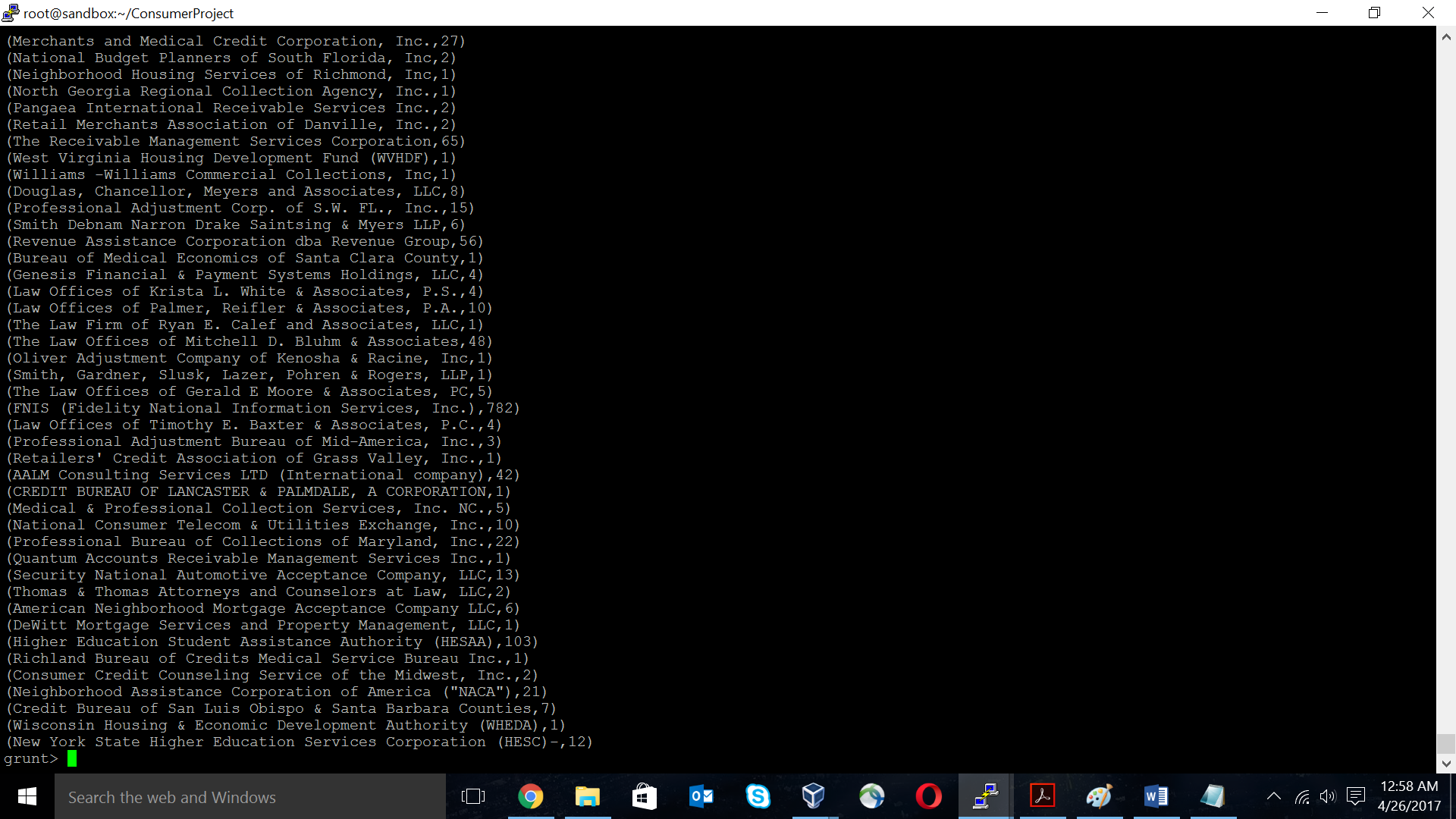
grp\_data = GROUP filter\_data BY company;

result = FOREACH grp\_data GENERATE group as company,

COUNT(filter\_data.company);

STORE result INTO ‘/user/root/output1/’ USING PigStorage(‘\t’);

dump result;



1. **Write a pig script to find number of complaints where consumer forum forwarded the complaint same day they received to respective company.**

filter\_data = FILTER data BY (sdate == rdate);

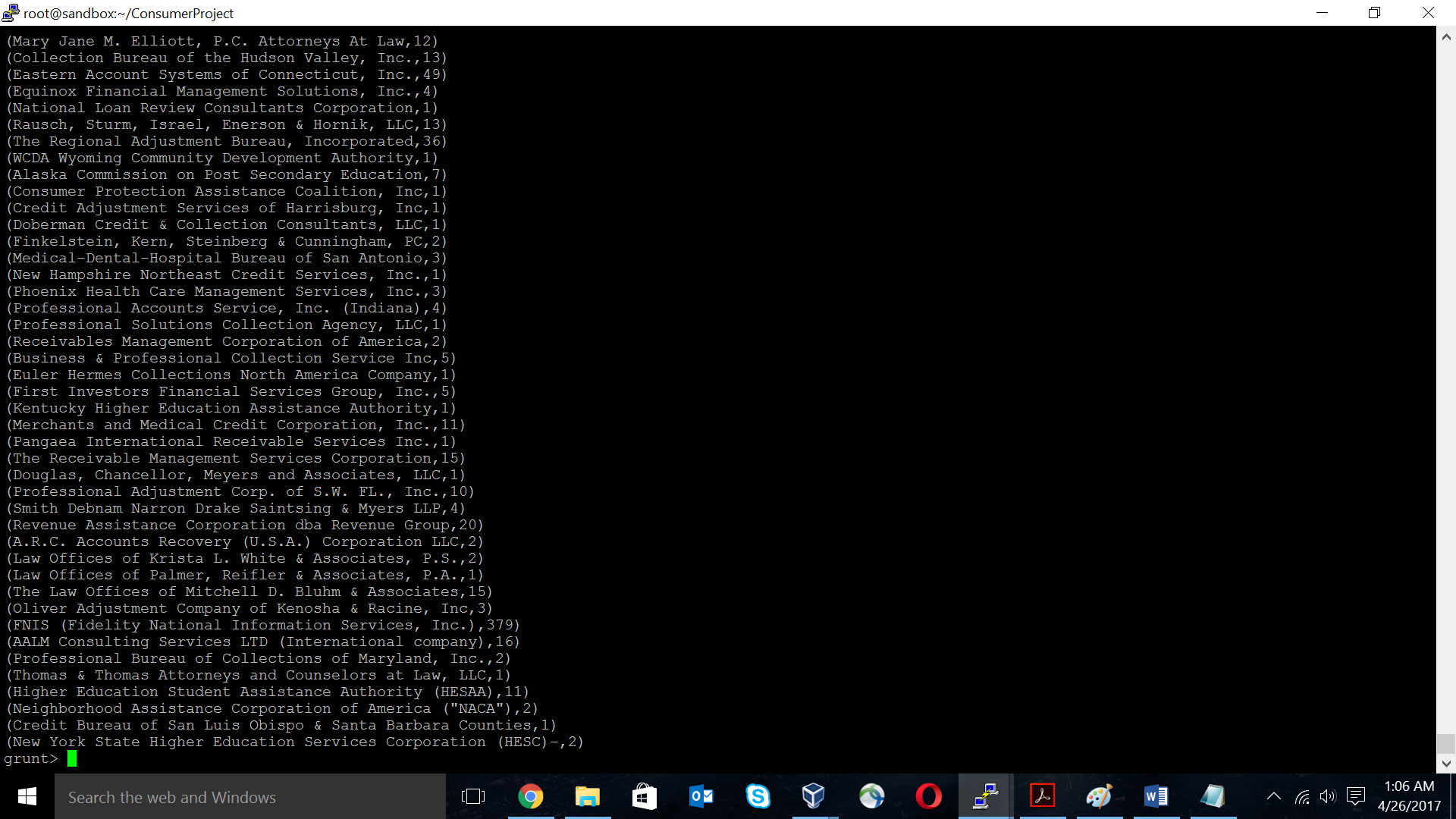
fdata\_group = GROUP filter\_data BY company;

result = FOREACH fdata\_group GENERATE group as company,

COUNT(filter\_data.company) as count;

STORE result INTO ‘/user/root/output2/’ USING PigStorage(‘\t’);

dump result;



1. **Write a pig script to find list of companies topping in complaints chart (companies with maximum number of complaints).**

grouped\_data = GROUP data BY company;

complaint\_count = FOREACH grouped\_data GENERATE group as cmpny,

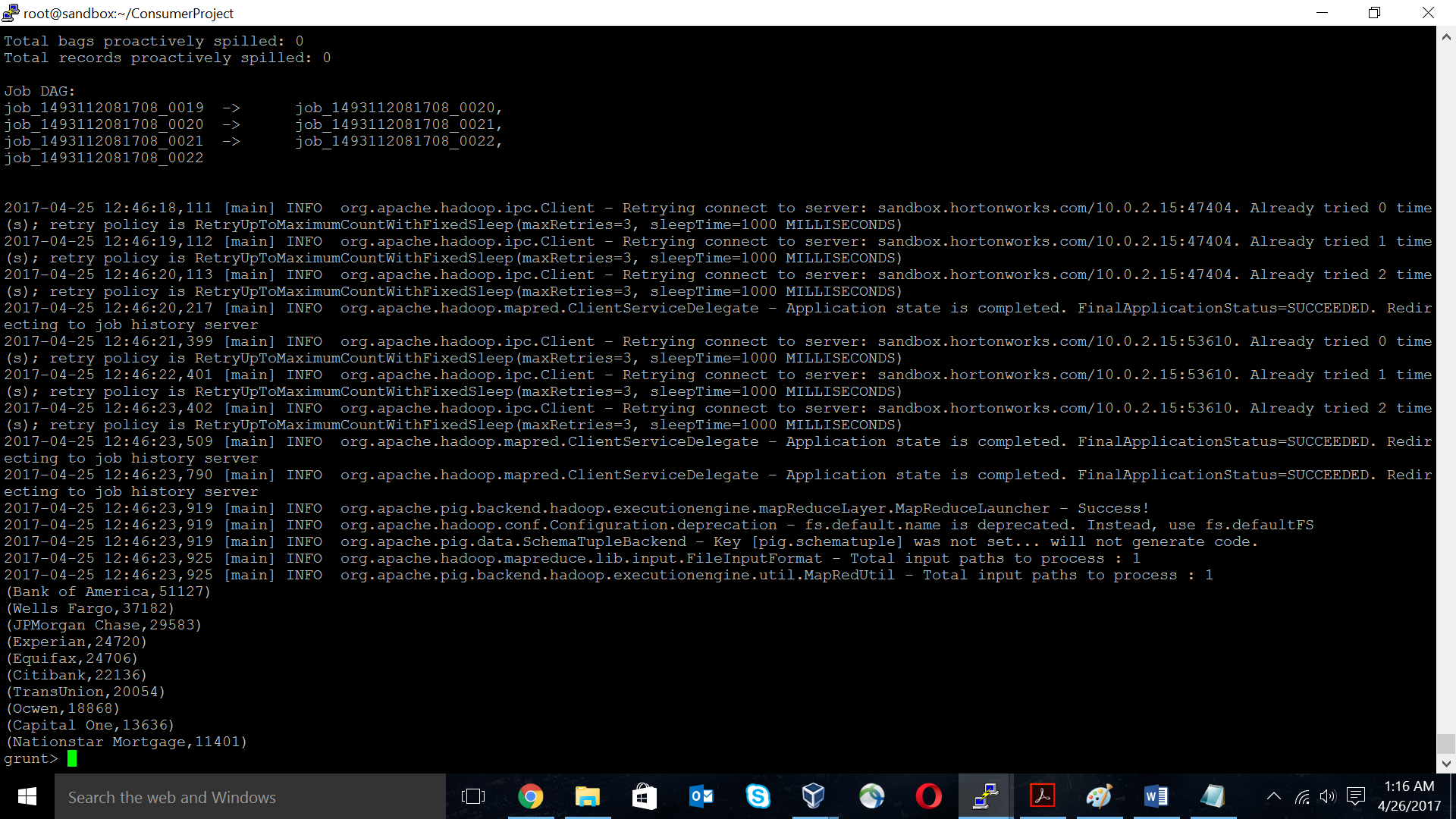
COUNT(data.complaintID) as count;

sorted\_company\_list = ORDER complaint\_count BY count DESC;

result = LIMIT sorted\_company\_list 10;

STORE result INTO ‘/user/root/output3’ USING PigStorage(‘\t’);

dump result;



1. **Write a pig script to find number of complaints filed with product type “Debt Collection” for the year 2015.**

filter\_data = FILTER data BY ((SUBSTRING(rdate,6,10) == '2015' OR

SUBSTRING(rdate,5,9) == '2015' OR SUBSTRING(rdate,4,8) == '2015') AND

(product == 'Debt collection'));

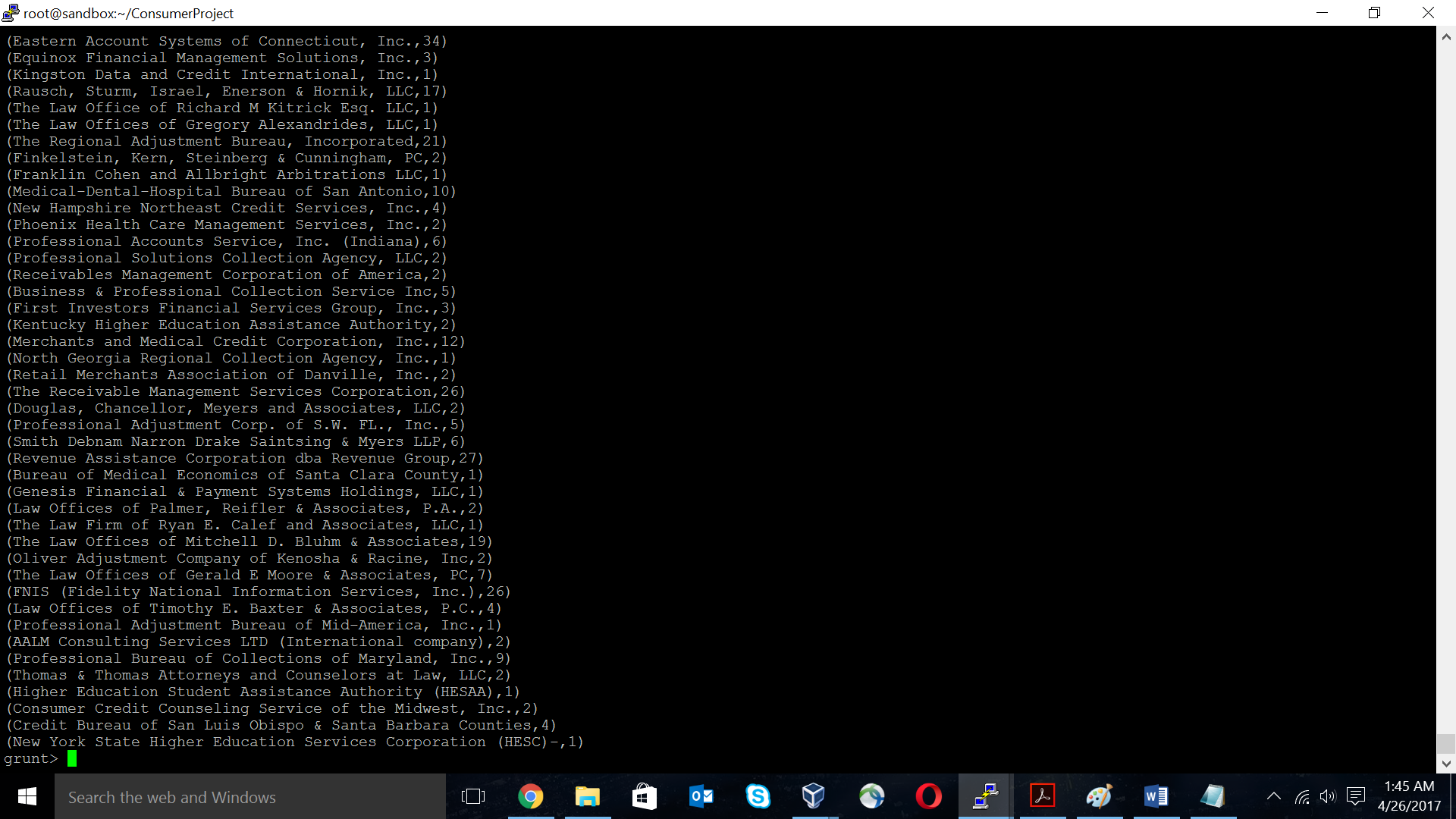
group\_data = GROUP filter\_data BY company;

result = FOREACH group\_data GENERATE group as company,

COUNT(filter\_data.product) as count;

STORE result INTO ‘/user/root/output4’ USING PigStorage(‘\t’);

dump result;



**Exporting output files into RDBMS:**

The output of the above 4 queries were stored in 4 separate output directories in HDFS under /user/root, namely output1, output2, output3 and output4 respectively.

In order to export these to RDBMS, we first need to create 4 tables in RDBMS which can obtain these outputs.

mysql –u root

mysql> create database ConsumerOutput;

mysql> use ConsumerOutput;

mysql> create table output1 (name varchar(100), count int);

mysql> create table output2 (name varchar(100), count int);

mysql> create table output3 (name varchar(100), count int);

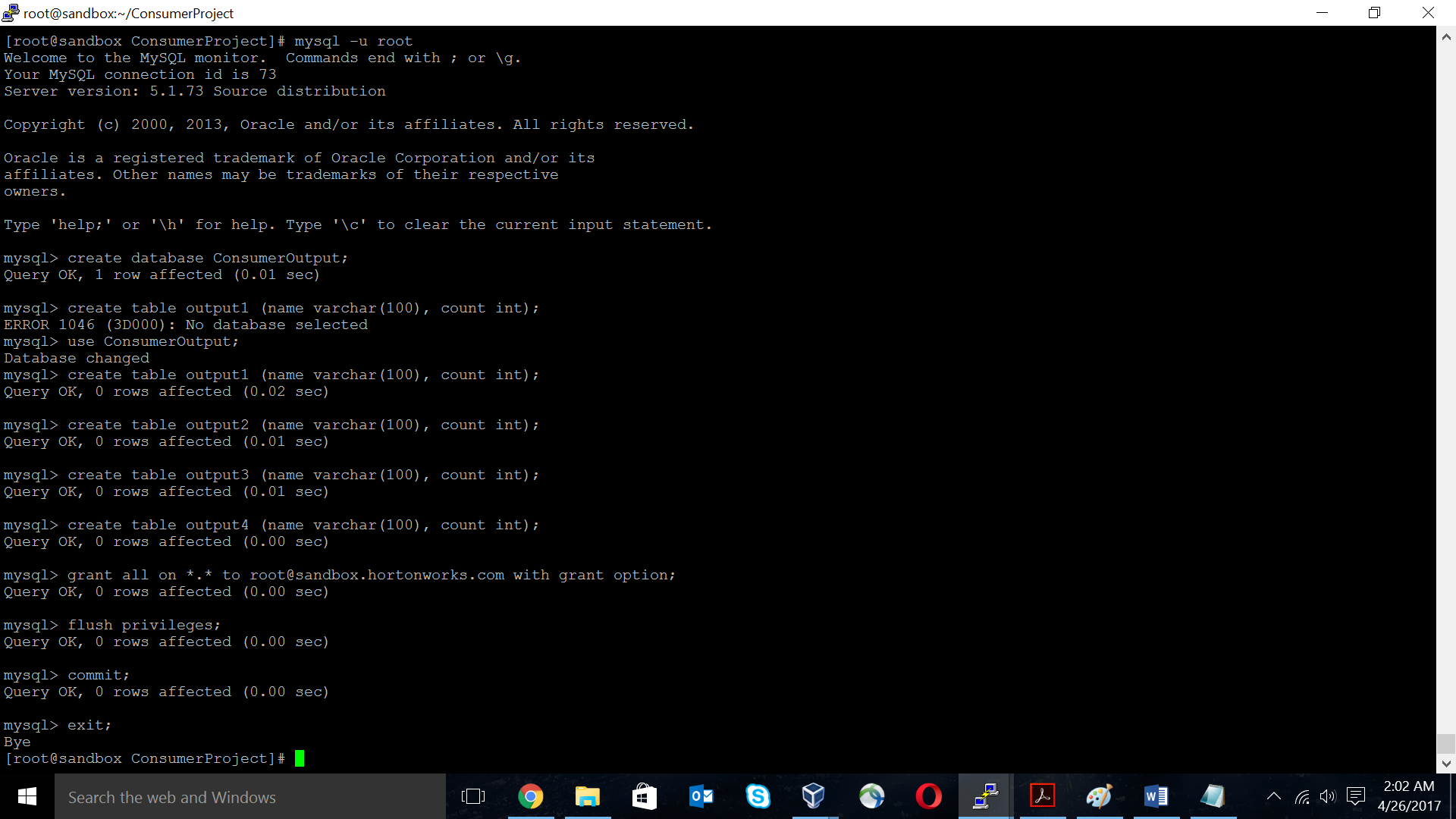
mysql> create table output4 (name varchar(100), count int);

mysql> grant all on \*.\* to [root@sandbox.hortonworks.com](mailto:root@sandbox.hortonworks.com) with grant option;

mysql> flush privileges;

mysql> commit;

mysql> exit;



Now, we will transfer the outputs in HDFS to the above created tables in RDBMS using SQOOP.

sqoop export --connect jdbc:mysql://localhost/ConsumerOutput --username 'root' --table 'output1' --export-dir '/user/root/output1/' --input-fields-terminated-by '\t' -m 1 --columns name,count;

sqoop export --connect jdbc:mysql://localhost/ConsumerOutput --username 'root' --table 'output2' --export-dir '/user/root/output2/' --input-fields-terminated-by '\t' -m 1 --columns name,count;

sqoop export --connect jdbc:mysql://localhost/ConsumerOutput --username 'root' --table 'output3' --export-dir '/user/root/output3/' --input-fields-terminated-by '\t' -m 1 --columns name,count;

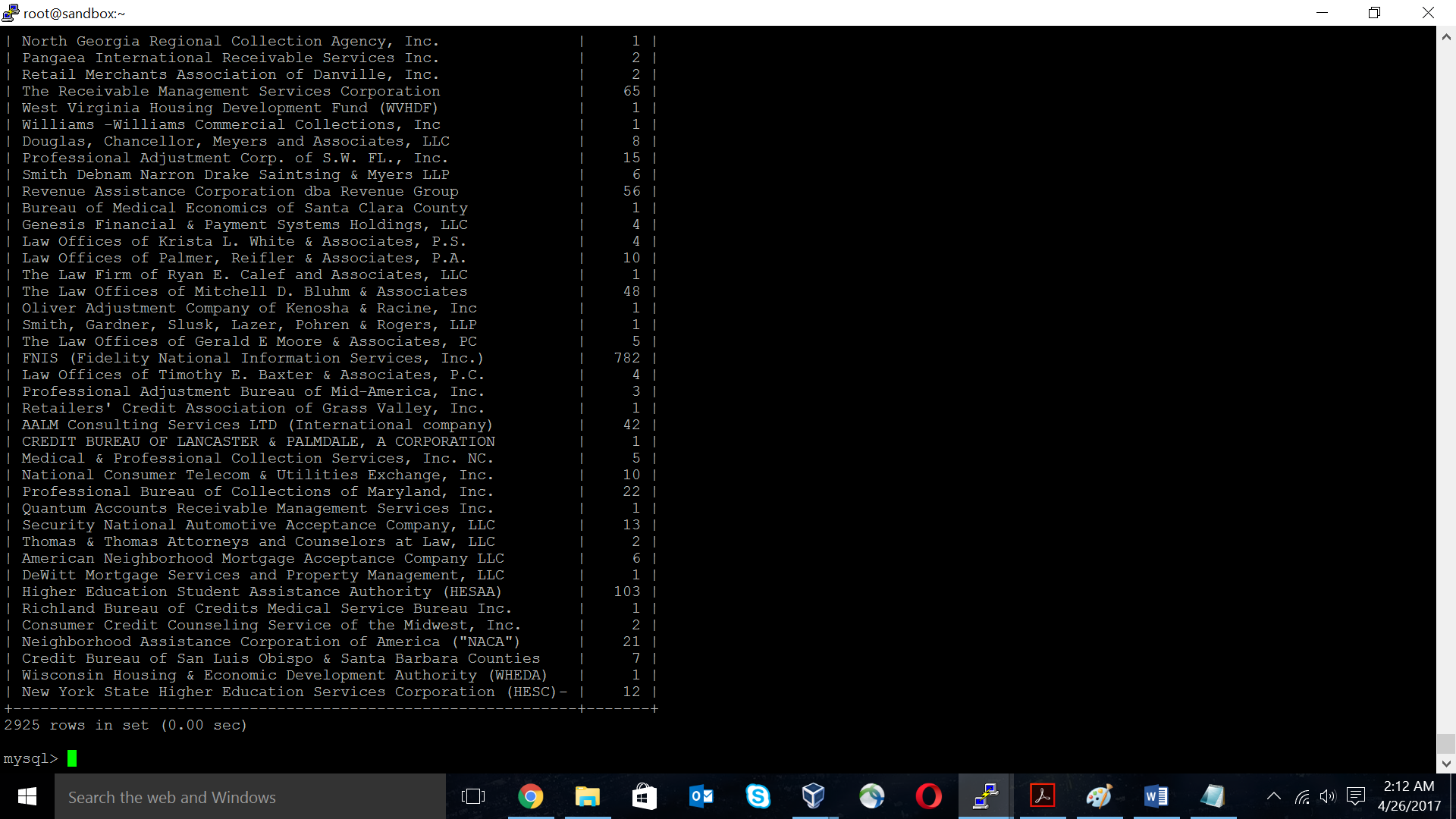
sqoop export --connect jdbc:mysql://localhost/ConsumerOutput --username 'root' --table 'output4' --export-dir '/user/root/output4/' --input-fields-terminated-by '\t' -m 1 --columns name,count;

**Verification of output in RDBMS:**

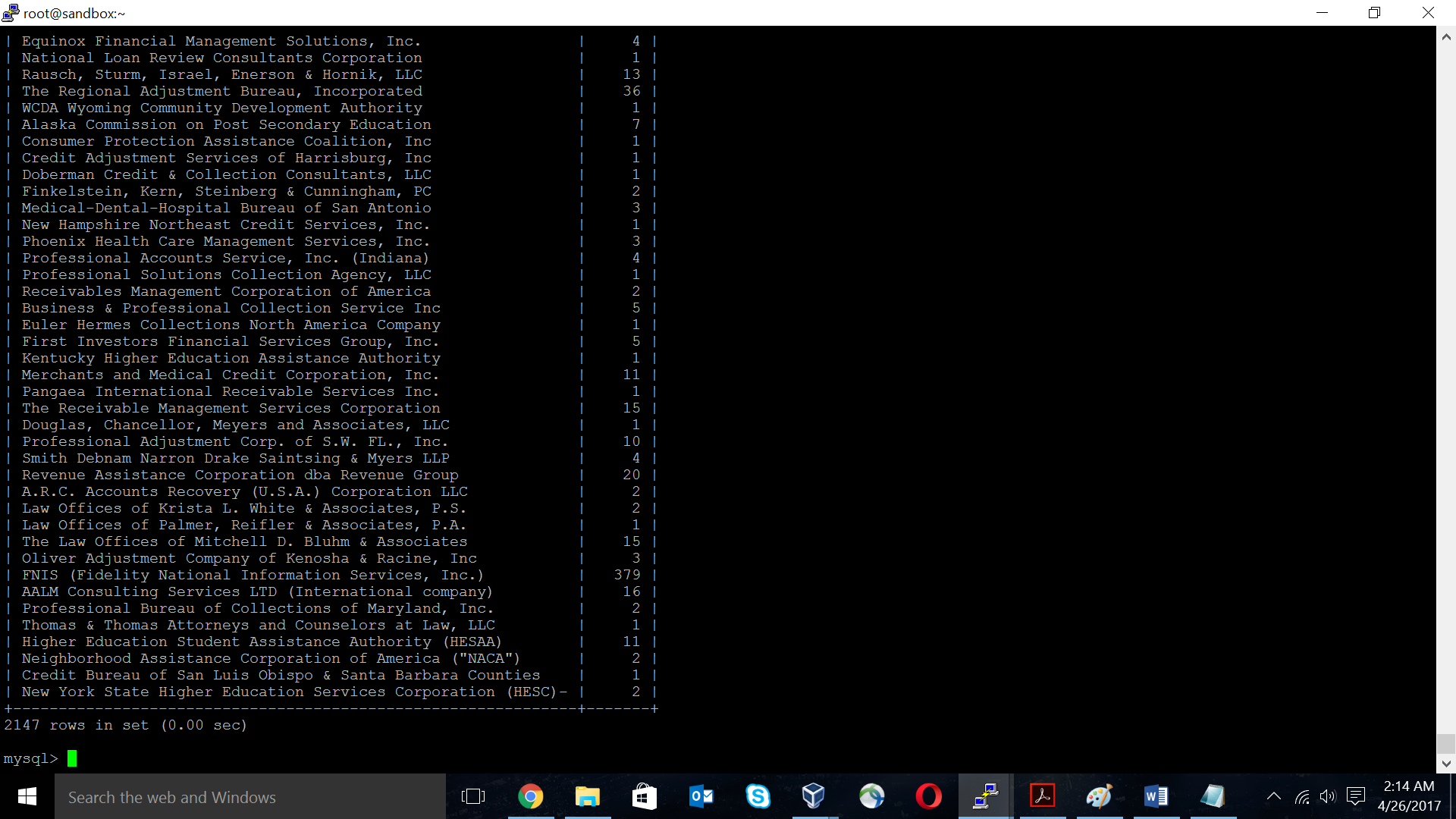
mysql –u root

mysql> use ConsumerOutput;

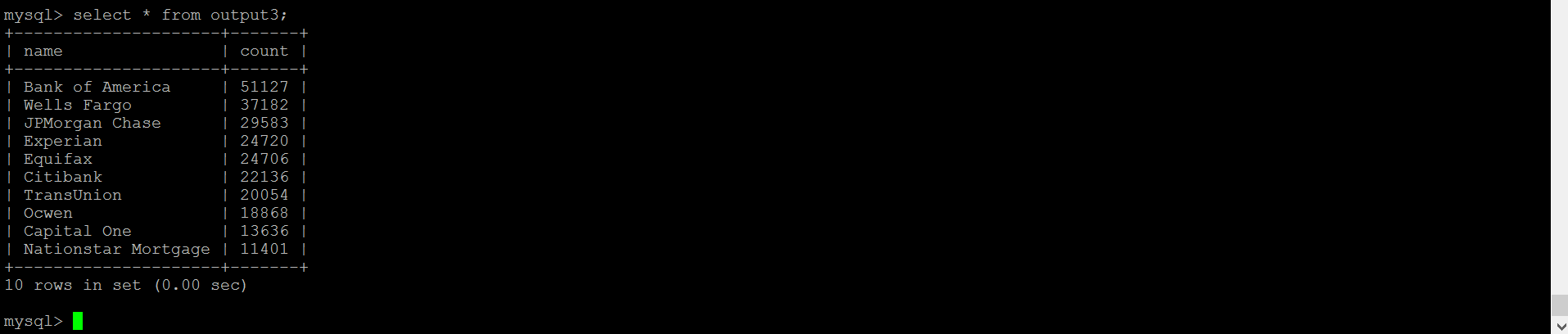
mysql> select \* from output1;



mysql> select \* from output2;



mysql> select \* from output3;



mysql> select \* from output4;

