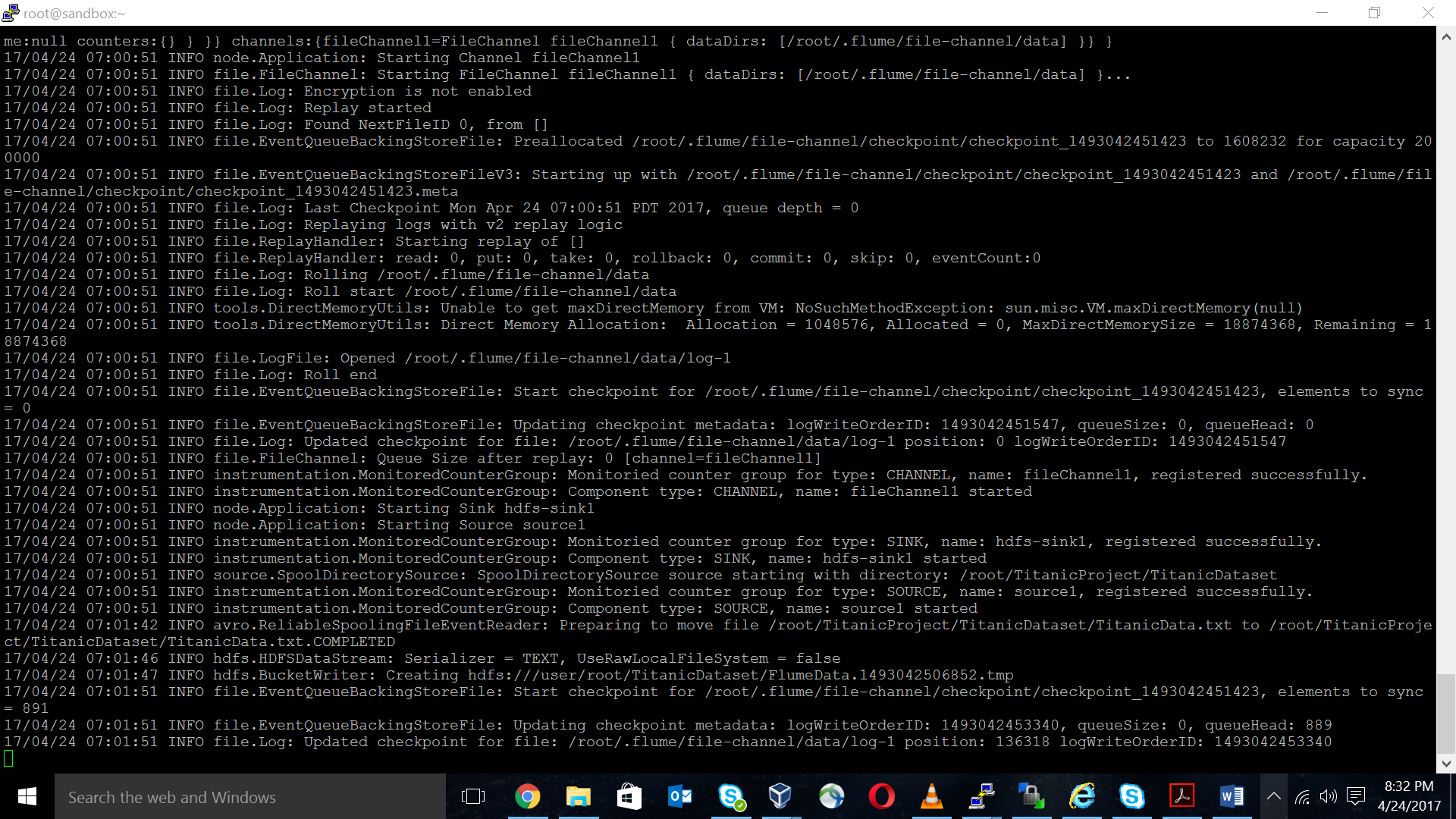
**SESSION 10-11: PROJECT I**

**Titanic Data Analysis**

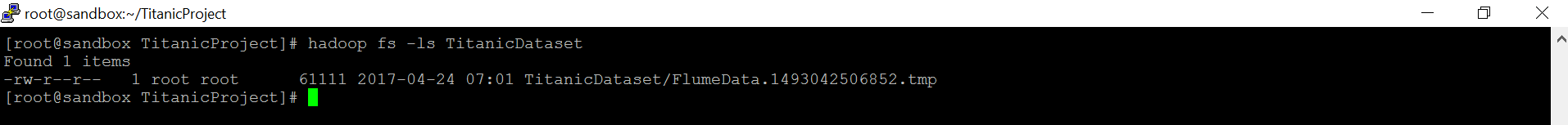
**Transferring Data into HDFS using Flume:**

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

flume-ng agent –n agent2 –f /root/TitanicProject/TitanicFlume.conf

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





**Processing Data through PIG:**

data = LOAD '/user/root/TitanicDataset' USING PigStorage(',') AS (pid:int, survive:int, pclass:int,

name:chararray, sex:chararray, age:int, sibsp:chararray, parch:chararray,

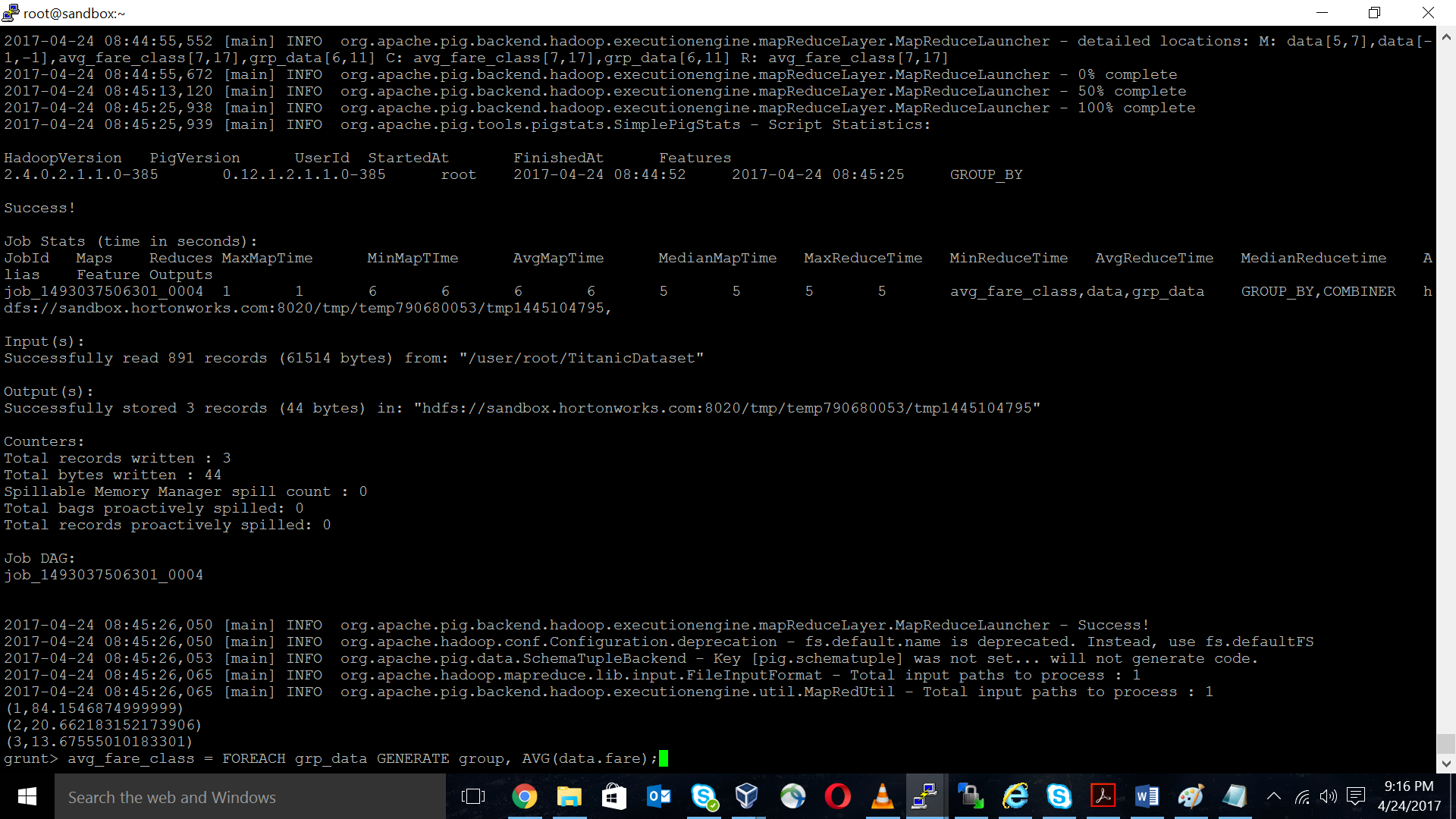
ticket:chararray, fare:double, cabin:chararray, embark:chararray);

1. **Find the average fare of each class.**

grp\_data = GROUP data BY pclass;

avg\_fare\_class = FOREACH grp\_data GENERATE group, AVG(data.fare);

dump avg\_fare\_class;



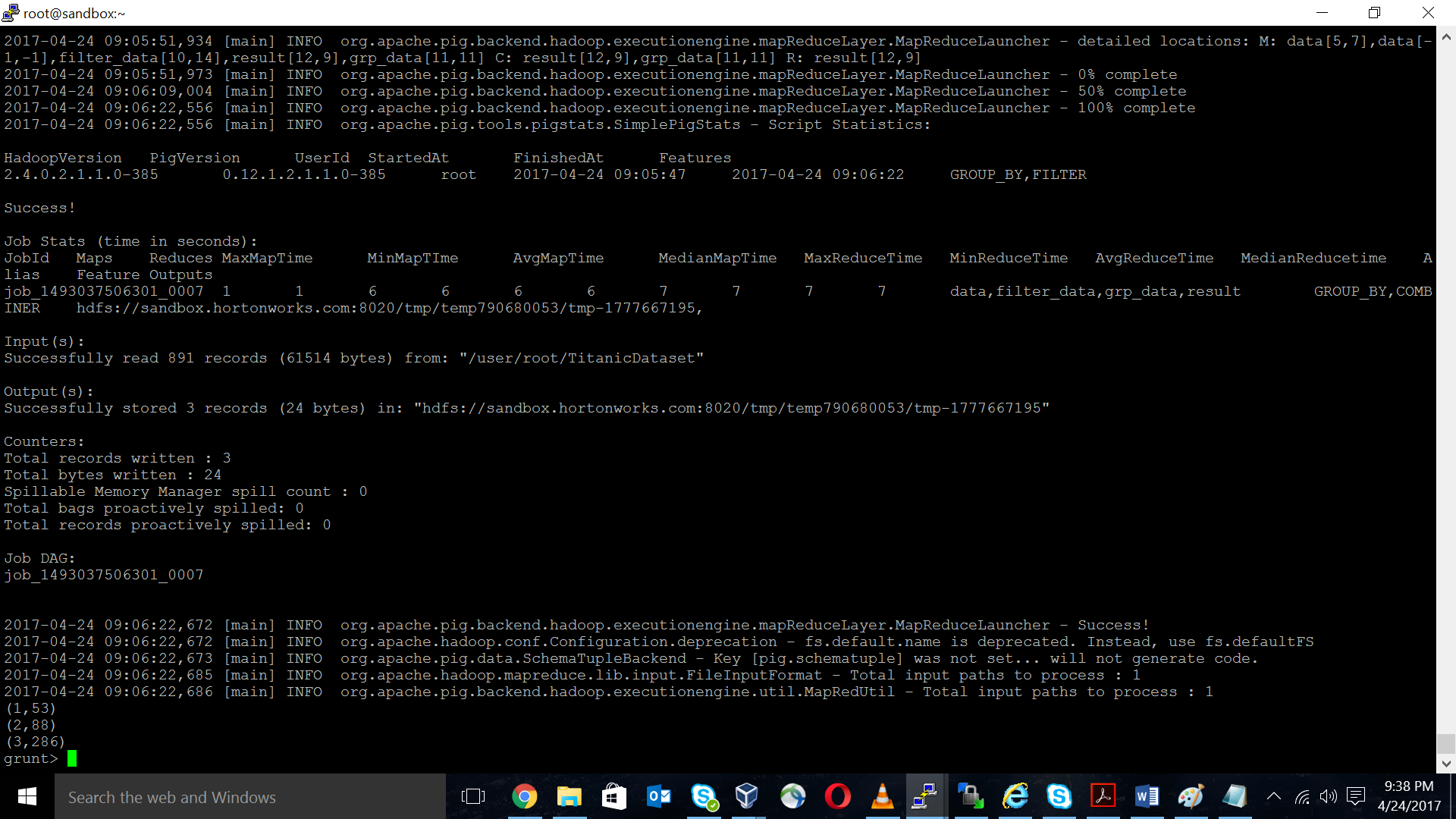
1. **Find the number of people alive in each class and are embarked in Southampton.**

filter\_data = FILTER data BY (survive == 0 AND embark == 'S');

grp\_data = GROUP filter\_data BY pclass;

result = FOREACH grp\_data GENERATE group, COUNT(filter\_data.pid);

dump result;



1. **Find the number of male and female people that died in each class.**

male\_deaths = FILTER data BY (survive == 1 AND sex == 'male');

female\_deaths = FILTER data BY (survive == 1 AND sex == 'female');

male\_group = GROUP male\_deaths BY pclass;

female\_group = GROUP female\_deaths BY pclass;

male\_result = FOREACH male\_group GENERATE group as pclass,

COUNT(male\_deaths.pid) as mcount;

female\_result = FOREACH female\_group GENERATE group as pclass,

COUNT(female\_deaths.pid) as fcount;

join\_result = JOIN male\_result BY pclass, female\_result BY pclass;

result = FOREACH join\_result GENERATE male\_result::pclass, male\_result::mcount,

female\_result::fcount;

dump result;

