AVIATION DATA ANALYSIS USING APACHE PIG

===============================================================

The U.S. Department of Transportation’s (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights appears in DOT’s monthly Air Travel Consumer Report, published about 30 days after the month’s end, as well as in summary tables posted on this website. Summary statistics and raw data are made available to the public at the time the Air Travel Consumer Report is released.

**Input datasets for POC:**

Airports.csv

DelayedFlights.csv -> https://drive.google.com/uc?id=0B\_Qjau8wv1KoWTVDUVFOdzlJNWM&export=download

**Problem Statement 1**

**Find out the top 5 most visited destinations.**

**Input Pig Commands 1:**

REGISTER '/home/cloudera/Downloads/piggybank-0.17.0.jar';

df1 = LOAD '/home/cloudera/chhaya/PigProject2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

ap1 = LOAD '/home/cloudera/chhaya/PigProject2/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

df2 = FILTER df1 BY ($23 == 'N') AND ($18 IS NOT NULL);

df3 = FOREACH df2 GENERATE $18;

da0 = JOIN df3 BY $0,ap1 BY $0;

da1 = FOREACH da0 GENERATE $3,$4,$5,$0;

da2 = GROUP da1 BY ($0,$1,$2);

da3 = FOREACH da2 GENERATE group as travel\_destination, COUNT(da1.$3) as cnt;

da4 = ORDER da3 BY cnt desc;

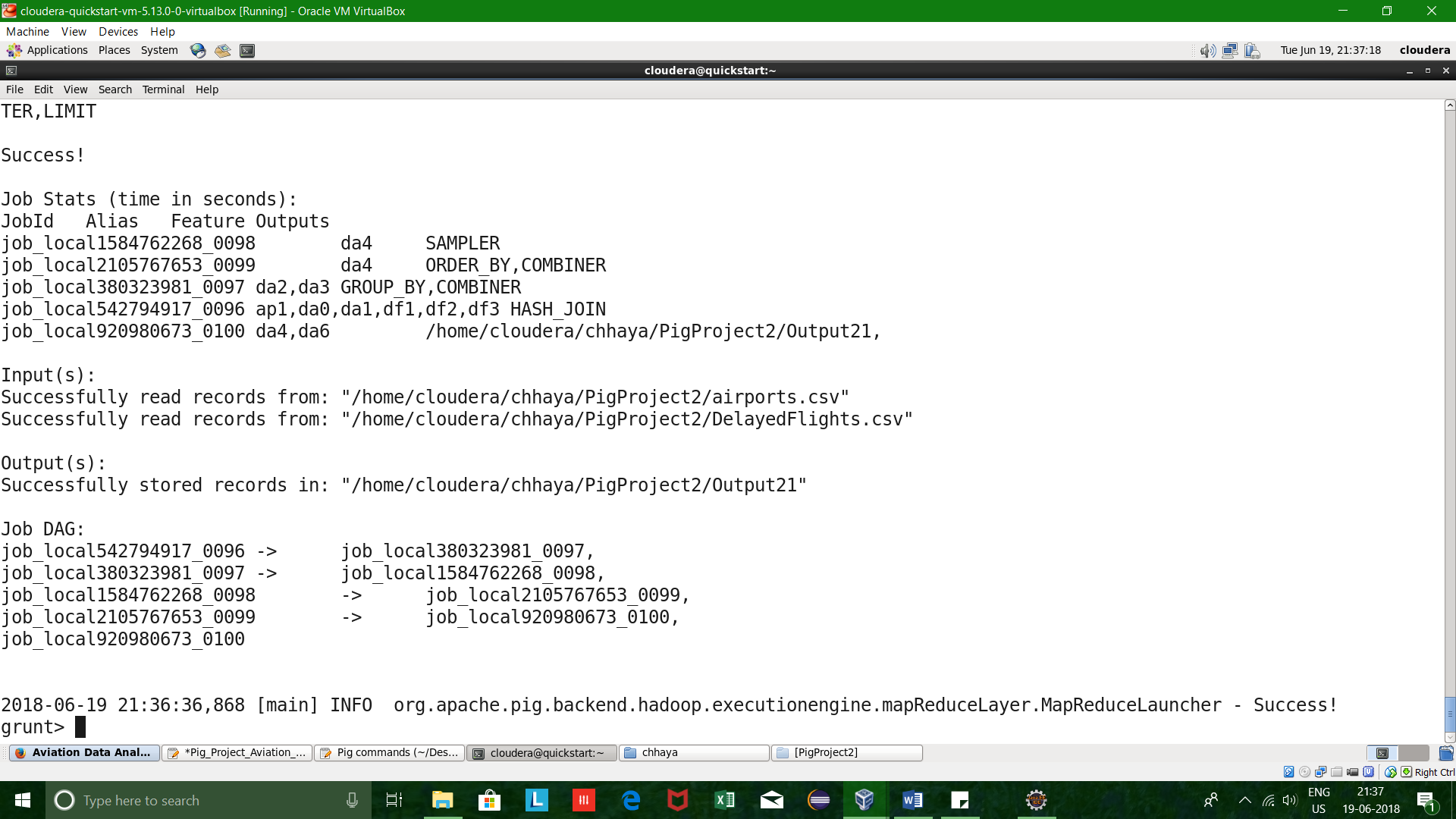
da5 = LIMIT da4 5;

da6 = FOREACH da5 generate travel\_destination;

DUMP da6;

STORE da6 INTO '/home/cloudera/chhaya/PigProject2/Output21' using PigStorage('\t');

**Output Screenshots 1:**



================================================================

**Problem Statement 2**

**Which month has seen the most number of cancellations due to bad weather?**

**Input Pig Commands 2:**

REGISTER '/home/cloudera/Downloads/piggybank-0.17.0.jar';

df1 = LOAD '/home/cloudera/chhaya/PigProject2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

ap1 = LOAD '/home/cloudera/chhaya/PigProject2/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

df2 = FOREACH df1 GENERATE $2 as month, $22 AS cancel\_status, $23 as cancel\_reason;

df3 = FILTER df2 BY (cancel\_status == '1') AND (cancel\_reason == 'B');

df4 = FOREACH df3 GENERATE month, cancel\_status;

df5 = GROUP df4 by month;

df6 = FOREACH df5 GENERATE group as mon, COUNT(df4.month) as total\_cancellations\_monthwise;

df7 = ORDER df6 bY total\_cancellations\_monthwise desc;

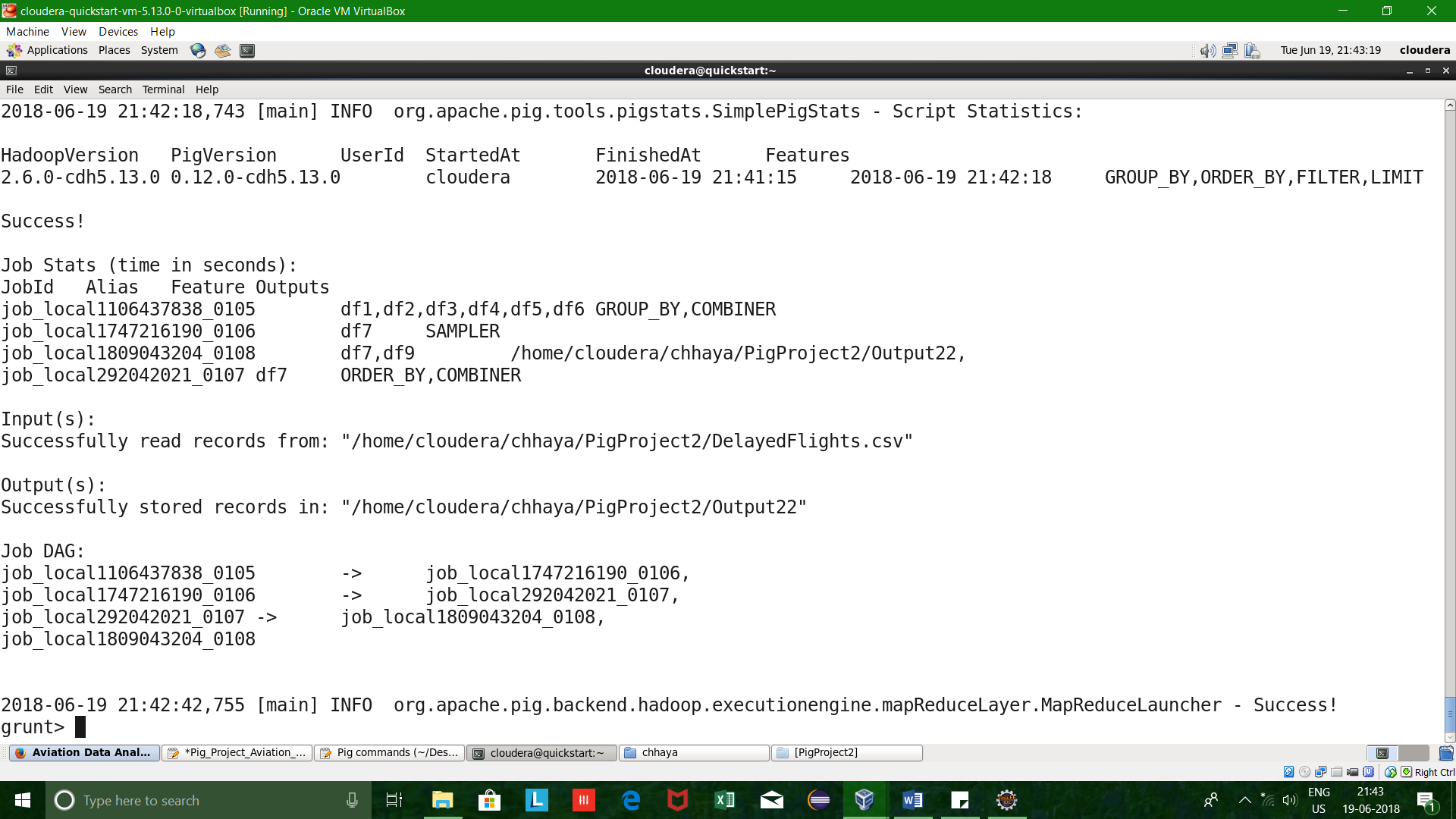
df8 = LIMIT df7 1;

df9 = FOREACH df8 generate mon;

DUMP df9;

STORE df9 INTO '/home/cloudera/chhaya/PigProject2/Output22' using PigStorage('^');

**Output Screenshots 2:**



================================================================

**Problem Statement 3**

**Top ten origins with the highest AVG departure delay.**

**Input Commands 3:**

REGISTER '/home/cloudera/Downloads/piggybank-0.17.0.jar';

df1 = LOAD '/home/cloudera/chhaya/PigProject2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

ap1 = LOAD '/home/cloudera/chhaya/PigProject2/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

df2 = FOREACH df1 generate $17 as origin, $16 as depdelay;

df3= GROUP df2 BY origin;

df4= FOREACH df3 GENERATE group as org, AVG(df2.depdelay) as avgdelay;

df5 = ORDER df4 by avgdelay desc;

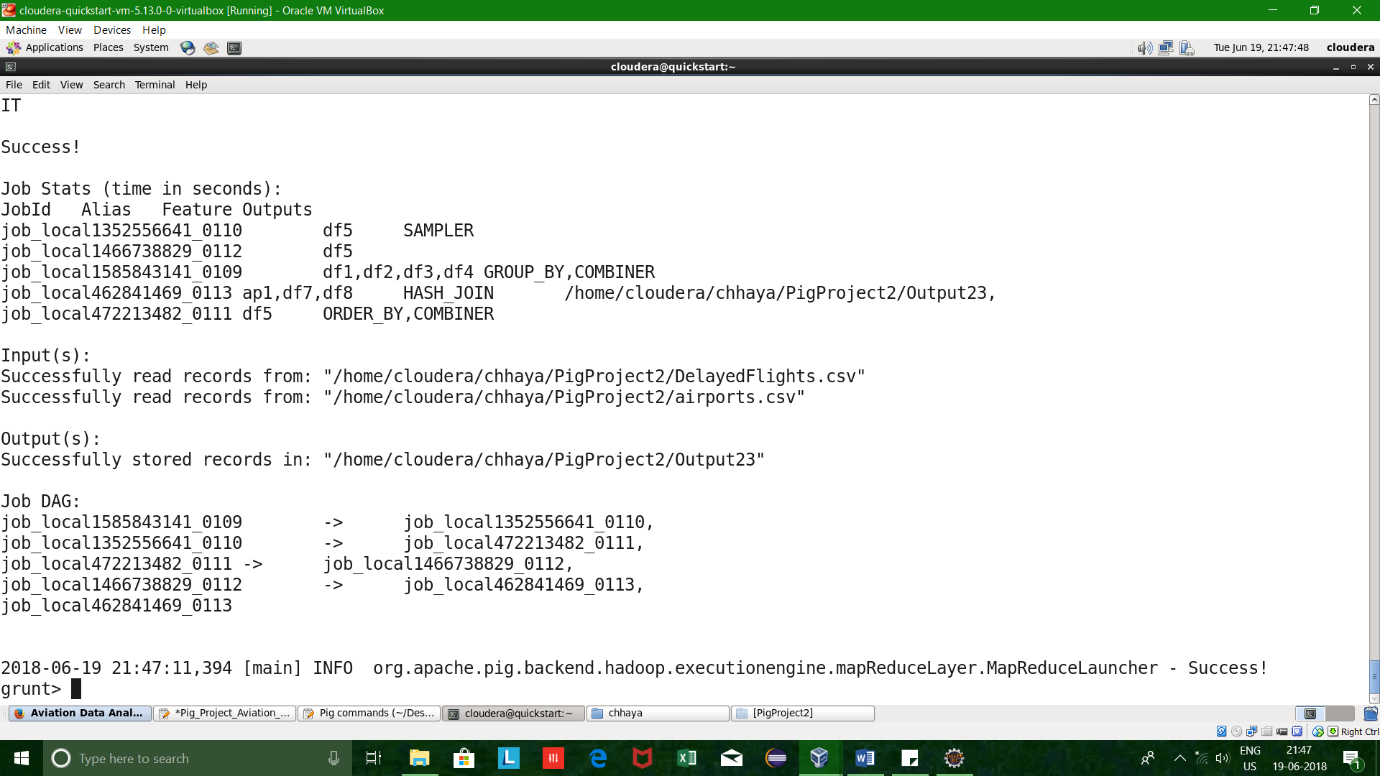
df6 = limit df5 10;

df7 = JOIN df6 by org, ap1 by $0;

df8 = FOREACH df7 GENERATE $4,$5,$6,$1;

STORE df8 INTO '/home/cloudera/chhaya/PigProject2/Output23' using PigStorage('\t');

**Output Screenshots 3:**



================================================================

**Problem Statement 4**

**Which route (origin & destination) has seen the maximum diversion?**

**Input Commands 4:**

REGISTER '/home/cloudera/Downloads/piggybank-0.17.0.jar';

df1 = LOAD '/home/cloudera/chhaya/PigProject2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

ap1 = LOAD '/home/cloudera/chhaya/PigProject2/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

df2 = FOREACH df1 generate $17 as origin, $18 as dest, $24 as diversion;

df3 = FILTER df2 by diversion =='1';

df4 = GROUP df3 by (origin, dest);

df5 = foreach df4 generate group as orig\_dest, SUM(df3.diversion) as total\_diversions;

df6 = ORDER df5 by total\_diversions desc;

df7 = LIMIT df6 1;

df8 = FOREACH df7 GENERATE FLATTEN(orig\_dest) as (orig,dest);

df9 = join df8 by orig, ap1 by $0;

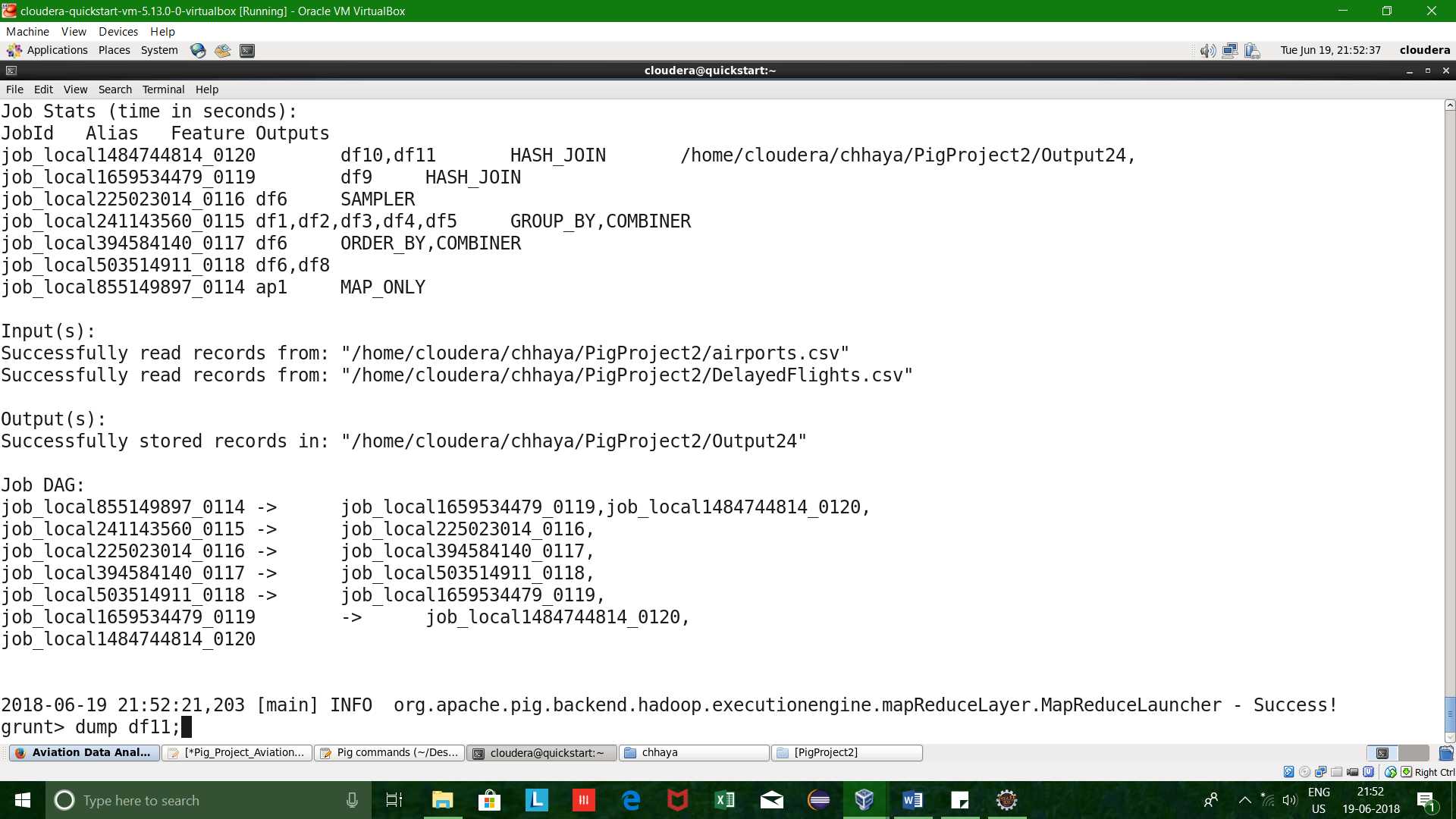
df10 = join df9 by $1, ap1 by $0;

df11 = FOREACH df10 GENERATE $4, $11;

STORE df11 INTO '/home/cloudera/chhaya/PigProject2/Output24' using PigStorage('-');

dump df11;

**Output Screenshots 4:**



================================================================