**Khanh Nguyen**

**Date: 8/6/2022**

**Project:**

**Distributed Processing Map Reduce Framework and Pig .**

DESCRIPTION

The US Department of Transport collects statistics from all the airlines, which include airline details, airport details, and flight journey details.  
These airlines have a global presence and they operate almost in every country.  
Flight data can help to decide which airline provides better service and find the routes in which flights are getting delayed.  
The data collected is present in the files: flight.csv, airline.csv, and aiport.csv.  
You are hired as a big data consultant to provide important insights.  
Your task is to write MapReduce jobs and provide insights on:

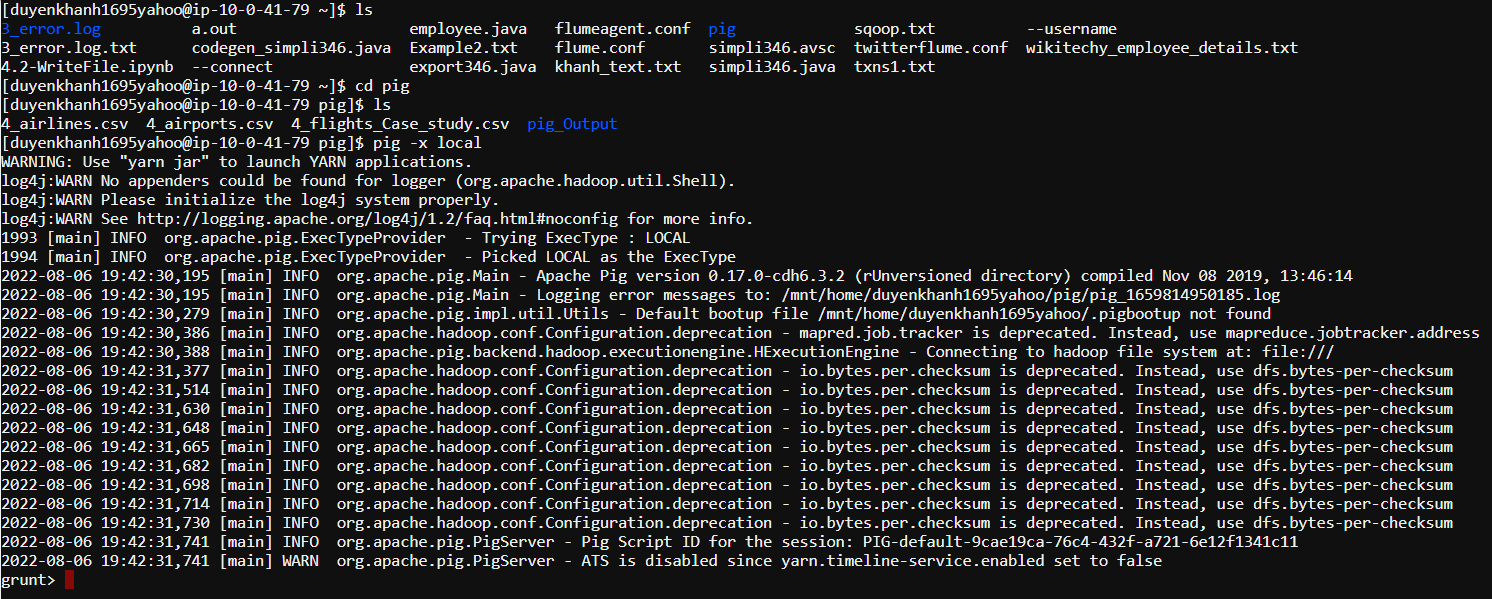
1. The number of flights that get canceled in a month, every year
2. The airline names and the number of times their flights were canceled and diverted
3. The number of times flights were delayed for each airline

Objective: To understand how Map-reduce works in Hadoop and its limitations.

You can download the datasets from [here](https://github.com/Simplilearn-Edu/BDHS/tree/master/Lesson%20end%20project%20Datasets/Lesson-4)

This is a non-gradable project.

**1. Upload .cvs files to local FTP /pig; Launch pig local**



**2. Loading all the necessary datasets:**

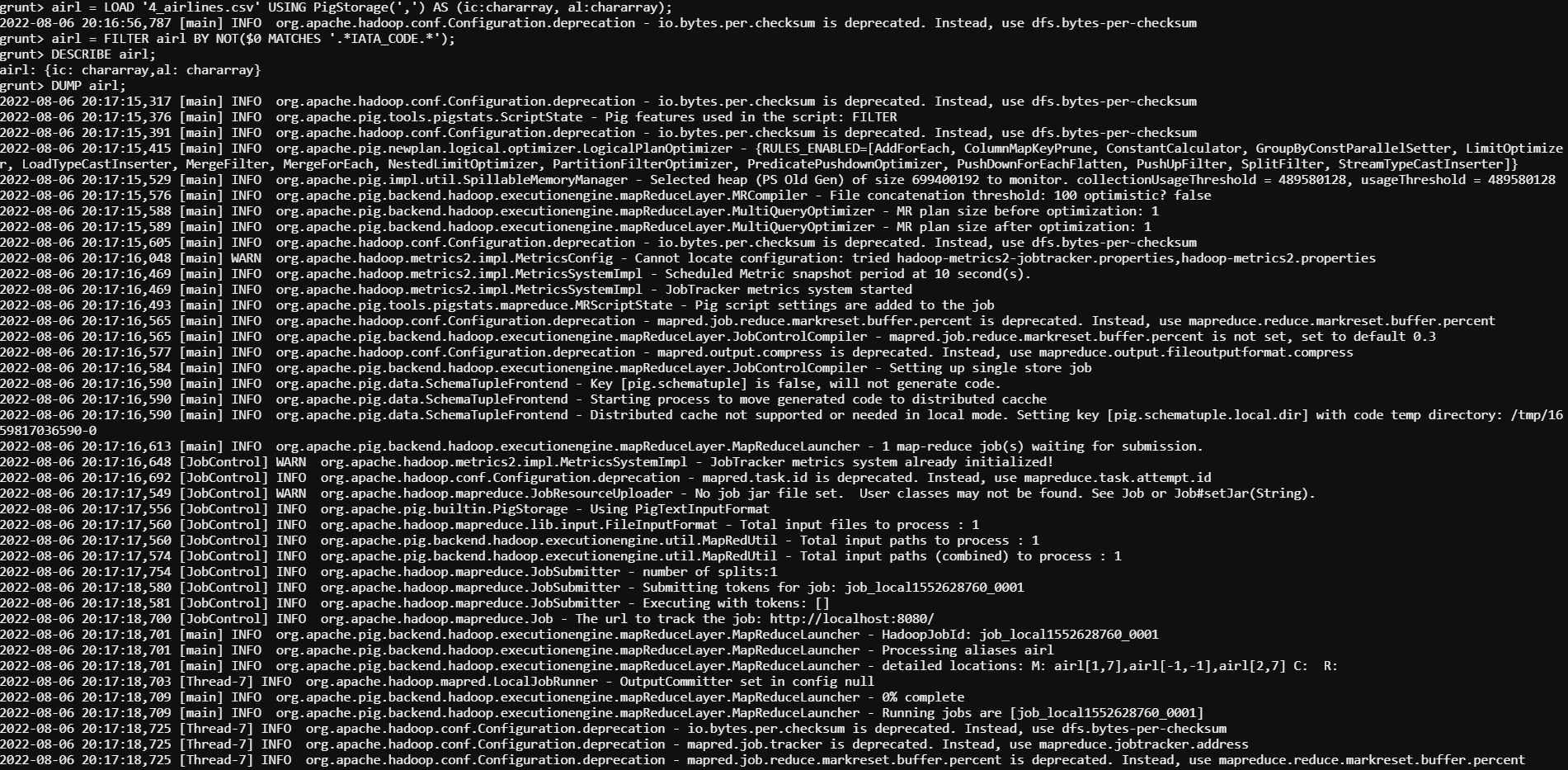
**Airline data:**

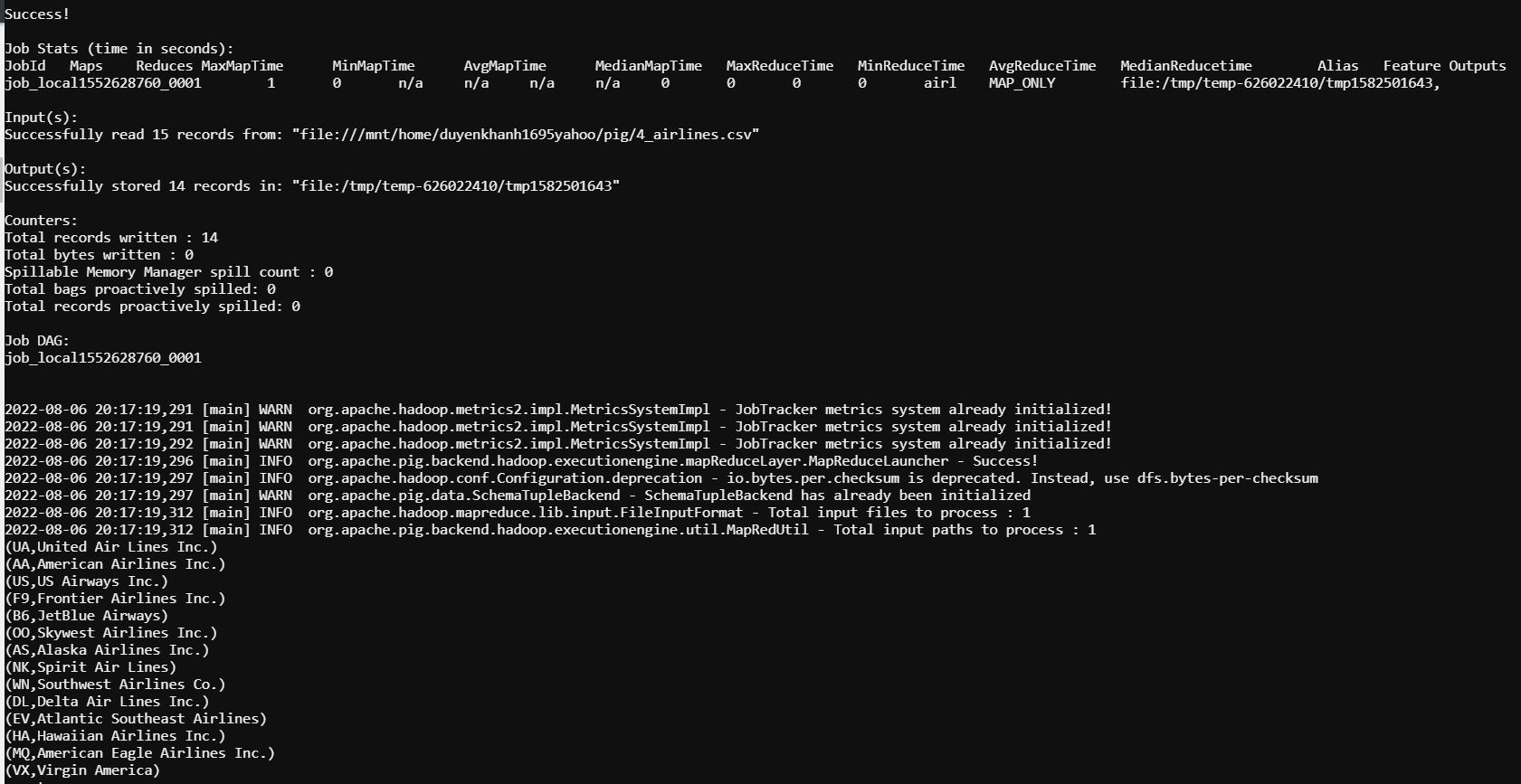
airl = LOAD '4\_airlines.csv' USING PigStorage(',') AS (ic:chararray, al:chararray);

airl = FILTER airl BY NOT($0 MATCHES '.\*IATA\_CODE.\*'); // filter out the header

DESCRIBE airl;

DUMP airl;



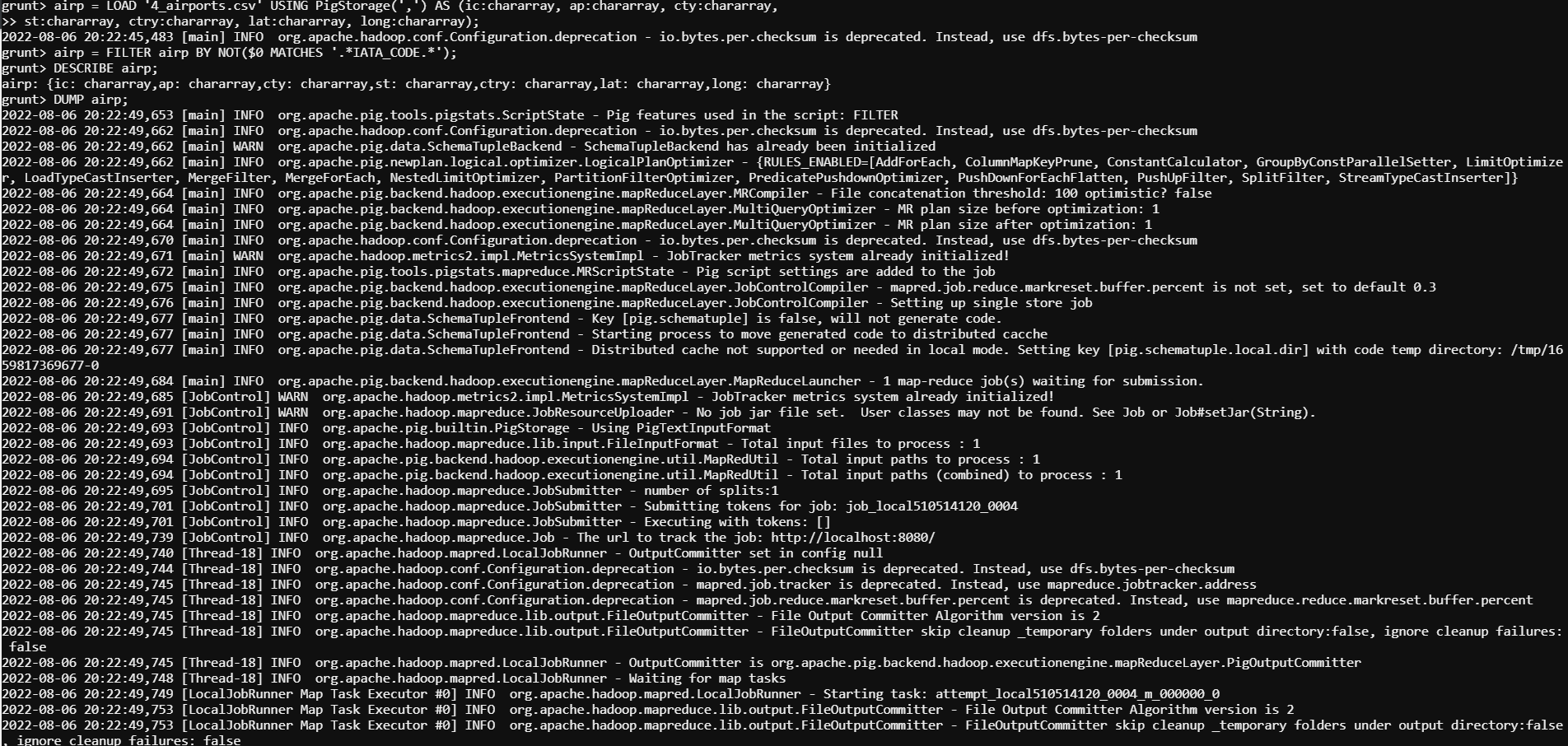


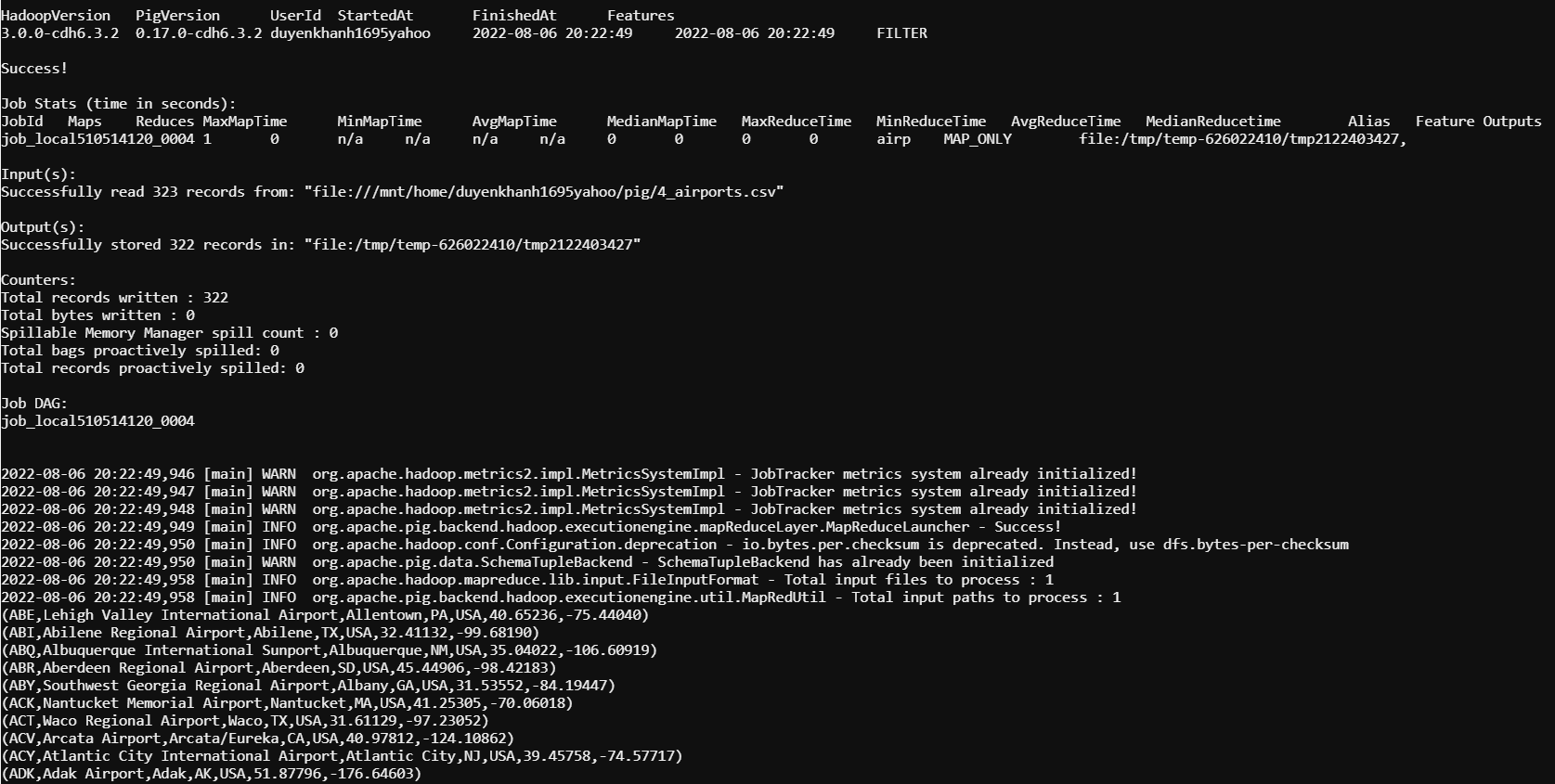
**Airport data:**

airp = LOAD '4\_airports.csv' USING PigStorage(',') AS (ic:chararray, ap:chararray, cty:chararray, st:chararray, ctry:chararray, lat:chararray, long:chararray);

airp = FILTER airp BY NOT($0 MATCHES '.\*IATA\_CODE.\*'); //filter out the header

DESCRIBE airp; DUMP airp;





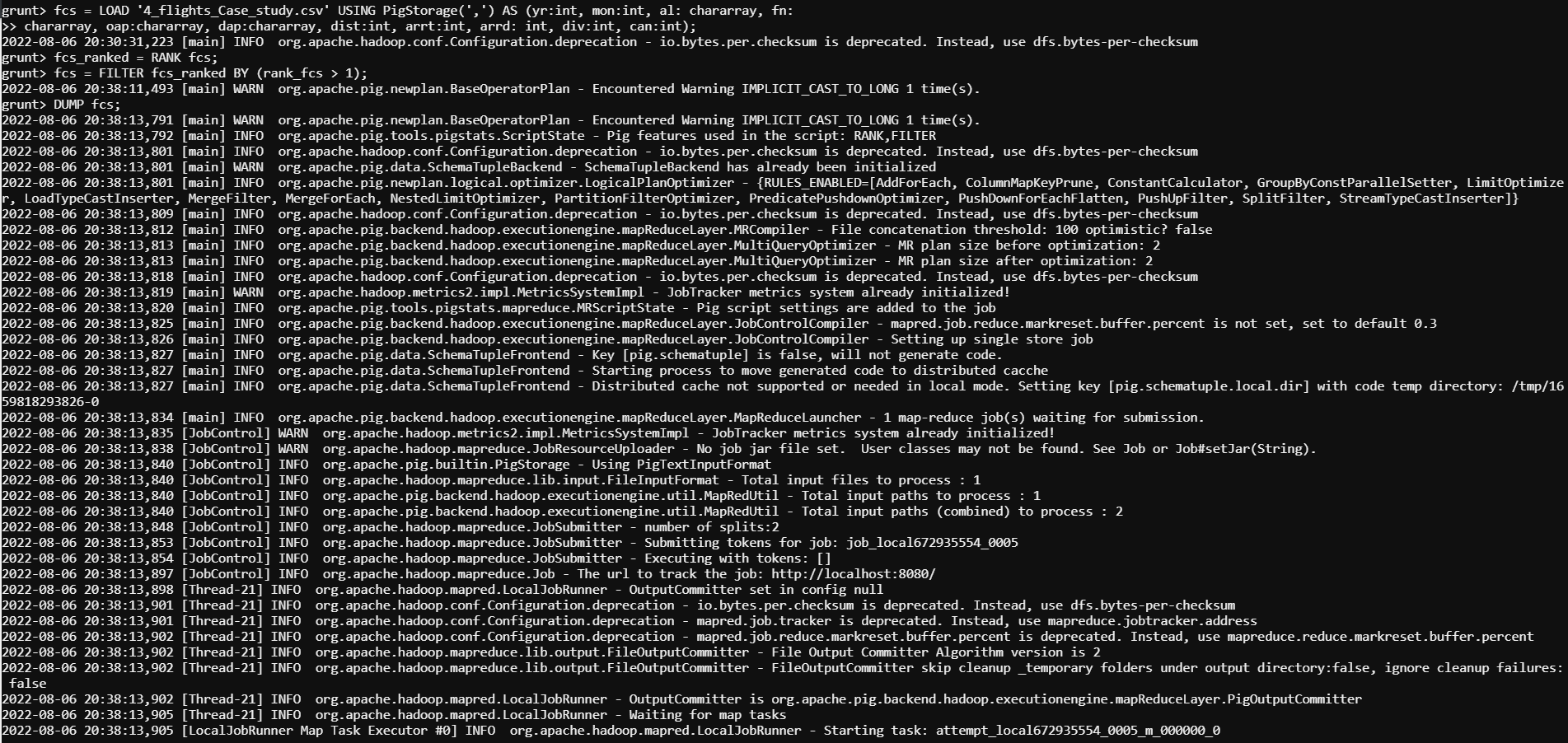
**Flight data:**

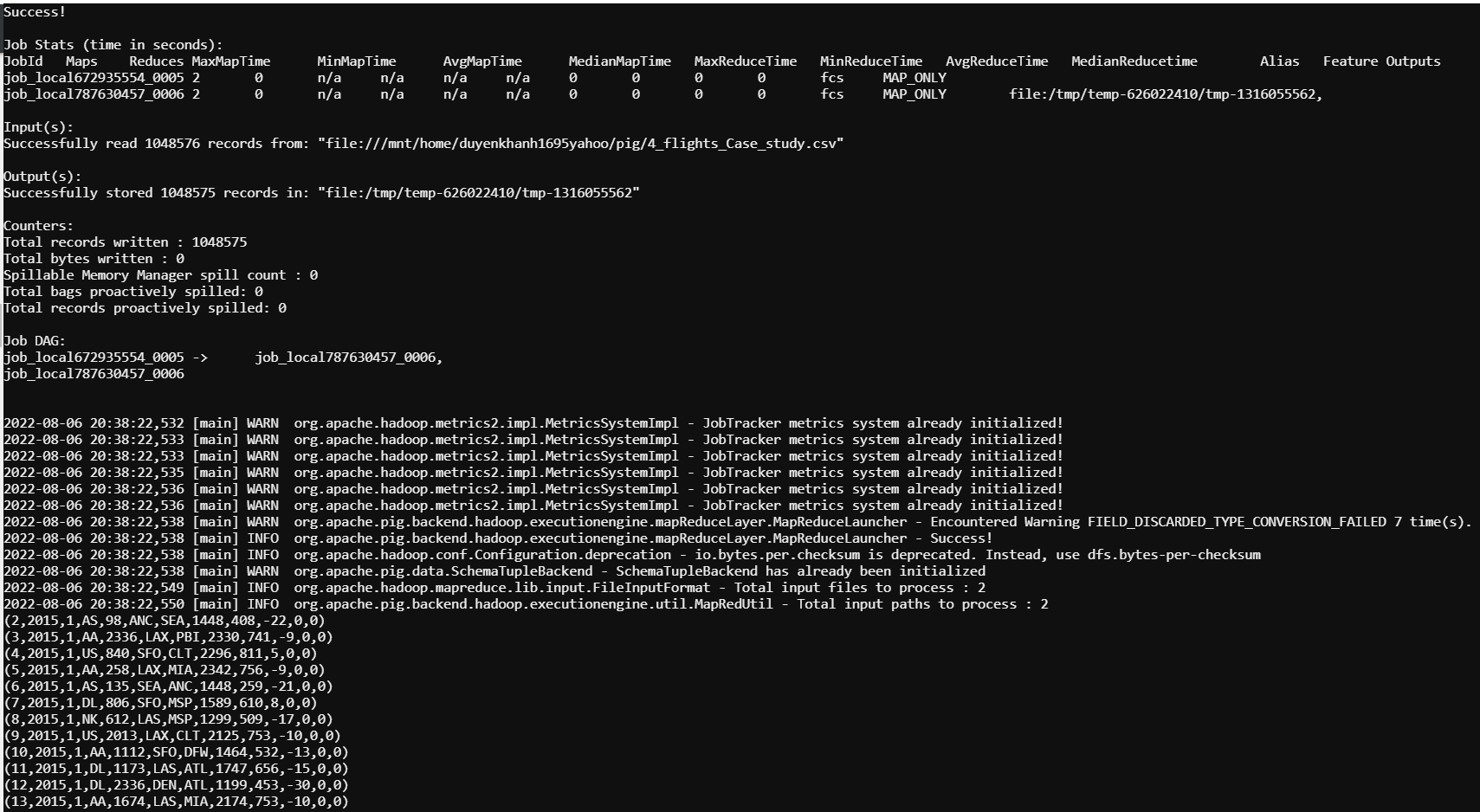
fcs = LOAD '4\_flights\_Case\_study.csv' USING PigStorage(',') AS (yr:int, mon:int, al: chararray, fn: chararray, oap:chararray, dap:chararray, dist:int, arrt:int, arrd: int, div:int, can:int);

fcs\_ranked = RANK fcs;

fcs = FILTER fcs\_ranked BY (rank\_fcs > 1); //filter out header (1st row)

DUMP fcs;



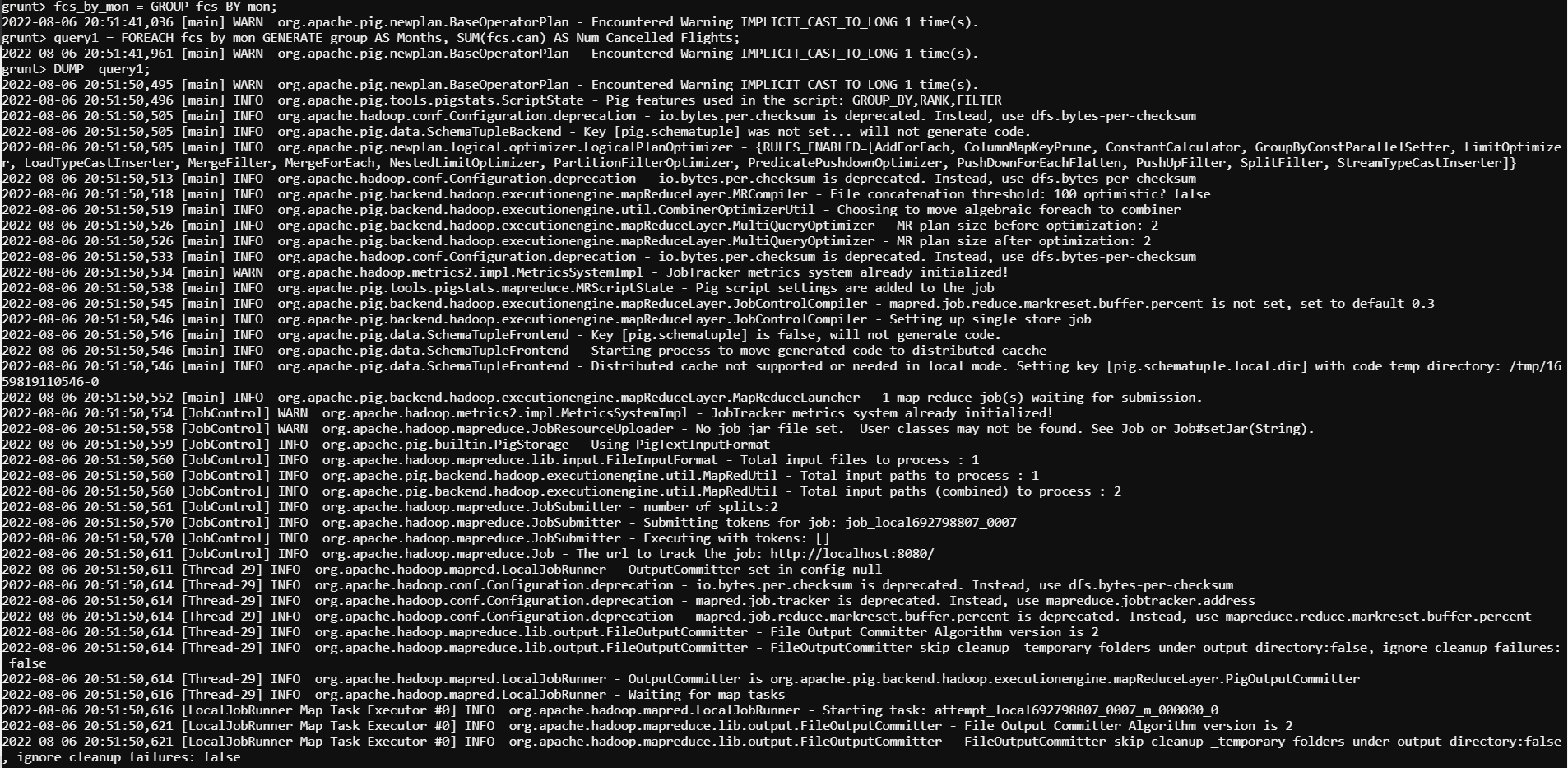


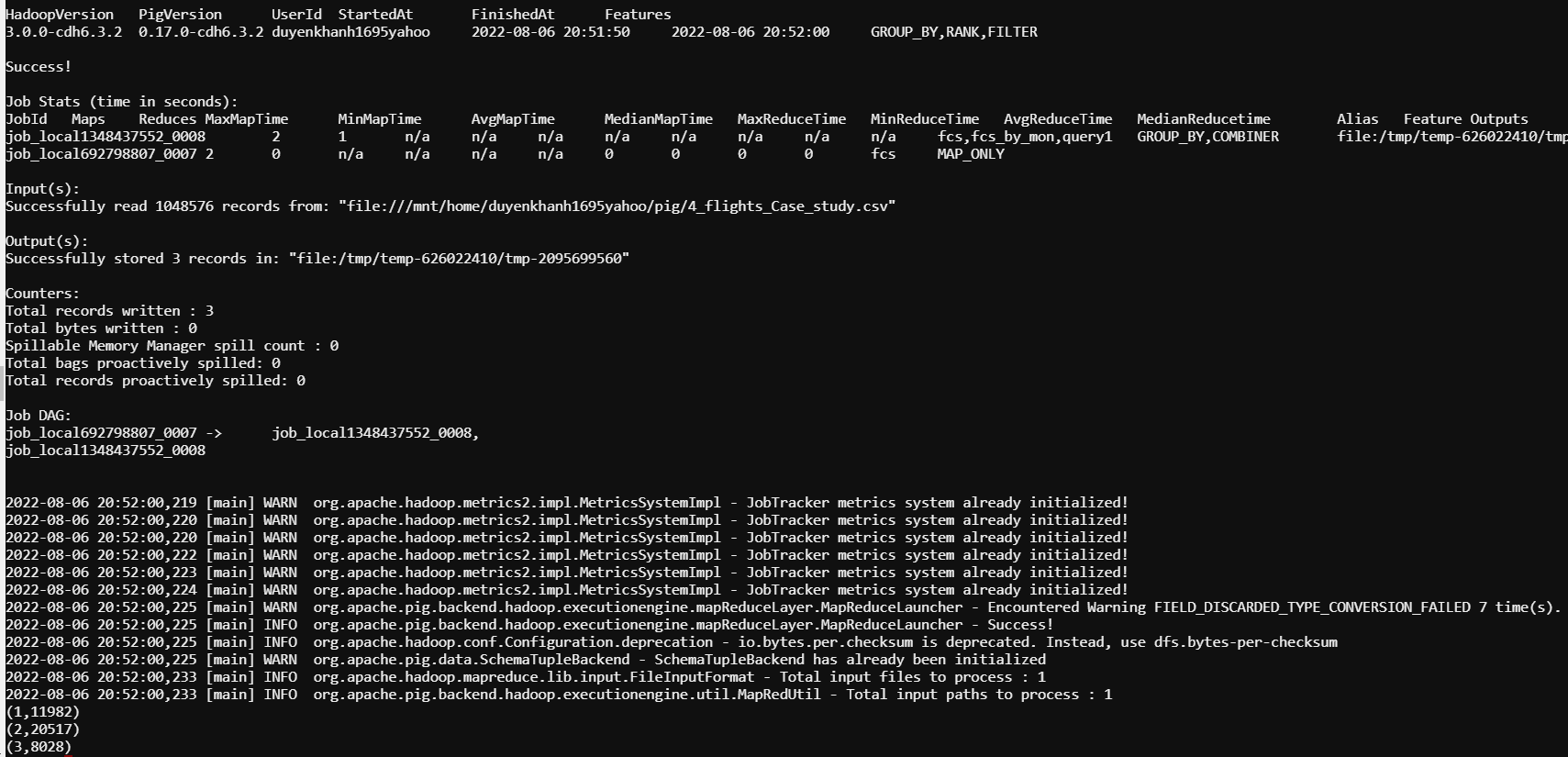
**Executing Pig Queries**

**1. The number of flights that get canceled in a month, every year**

fcs\_by\_mon = GROUP fcs BY mon;

query1 = FOREACH fcs\_by\_mon GENERATE group AS Months, SUM(fcs.can) AS Num\_Cancelled\_Flights;



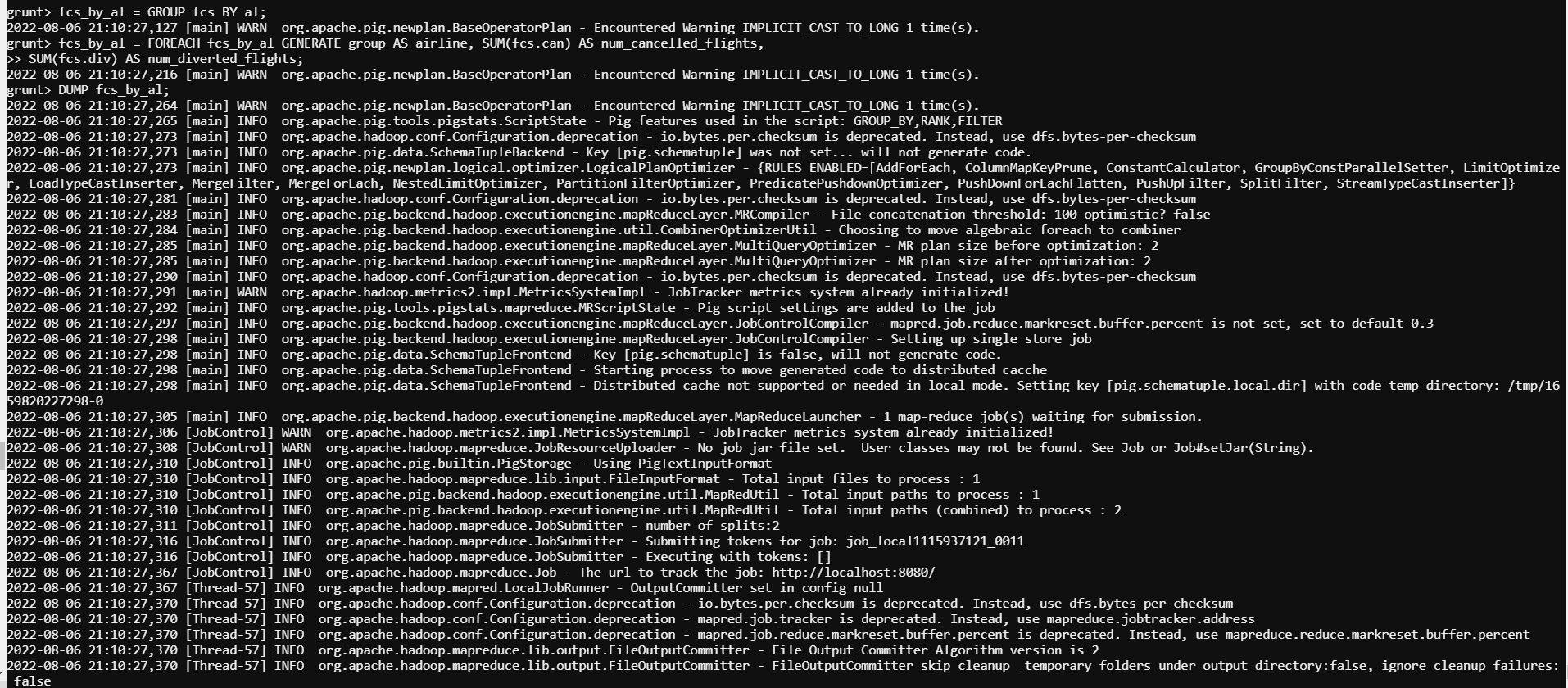


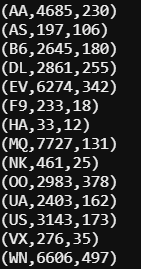
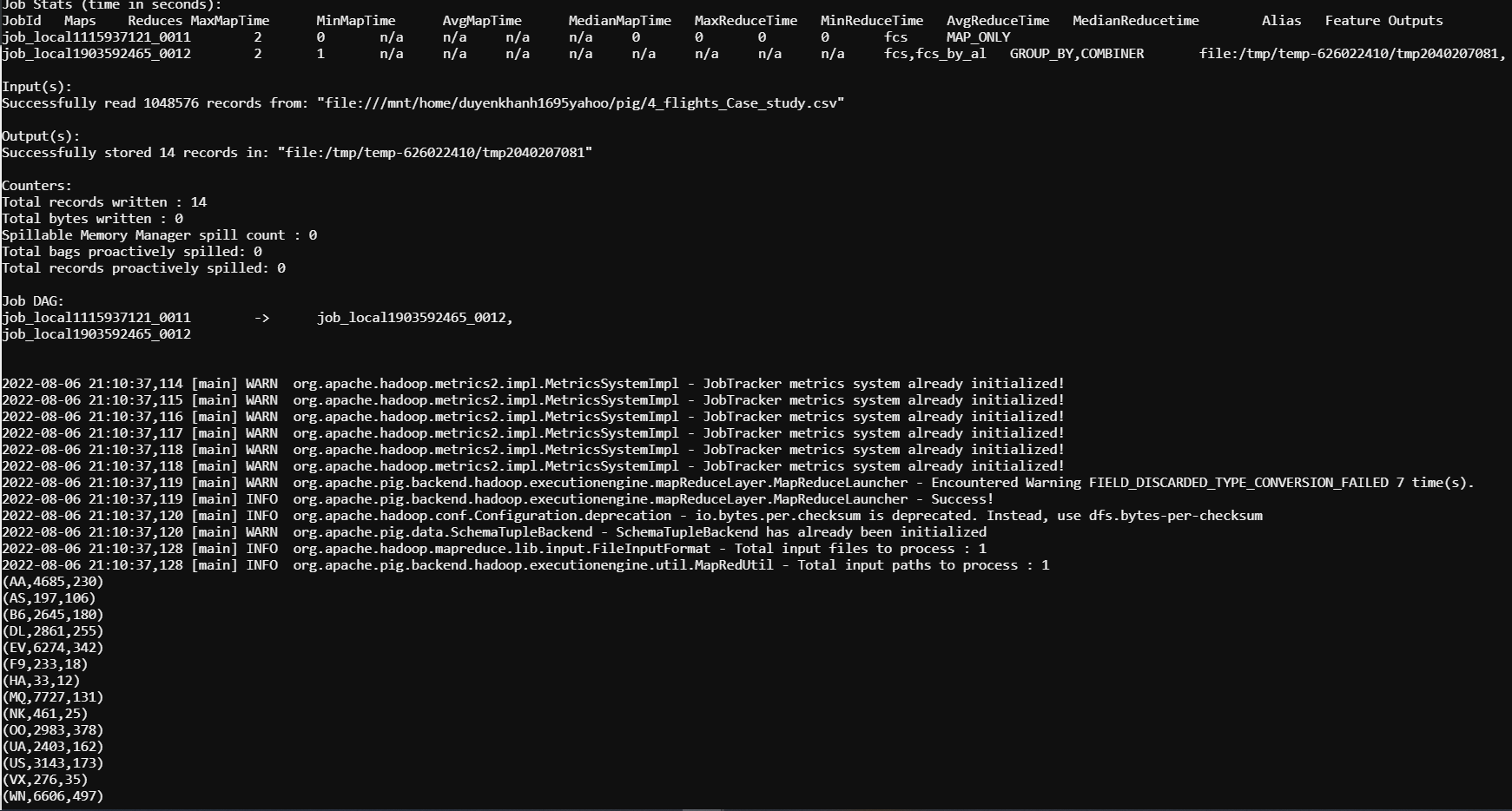
**2. The airline names and the number of times their flights were canceled and diverted**

fcs\_by\_al = GROUP fcs BY al;

fcs\_by\_al = FOREACH fcs\_by\_al GENERATE group AS airline, SUM(fcs.can) AS num\_cancelled\_flights, SUM(fcs.div) AS num\_diverted\_flights;

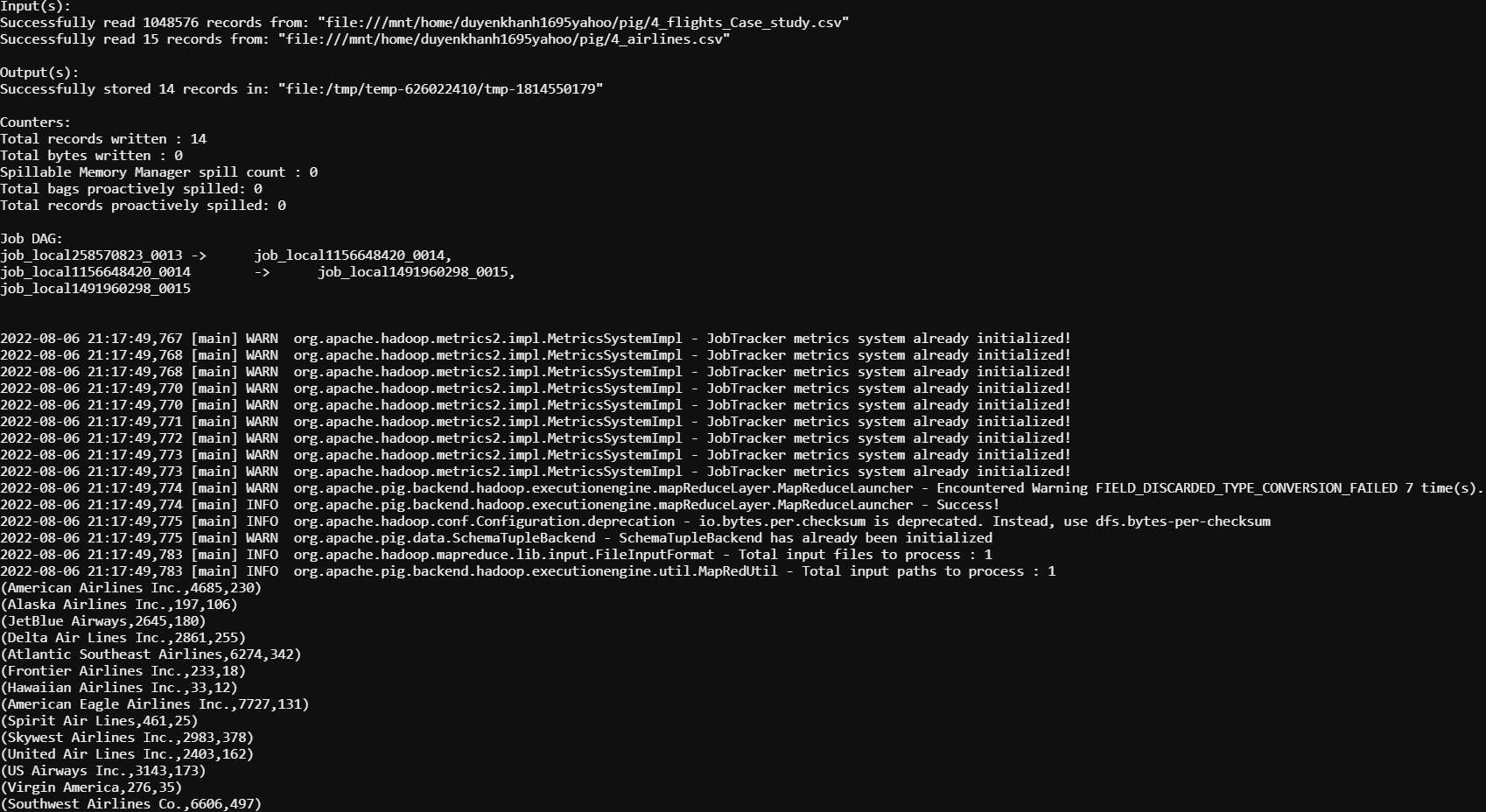
DUMP fcs\_by\_al;





fcs\_by\_al\_names = JOIN fcs\_by\_al BY airline, airl BY ic;

query2 = FOREACH fcs\_by\_al\_names GENERATE al AS airline\_names, num\_cancelled\_flights, num\_diverted\_flights;





**3. The number of times flights were delayed for each airline**

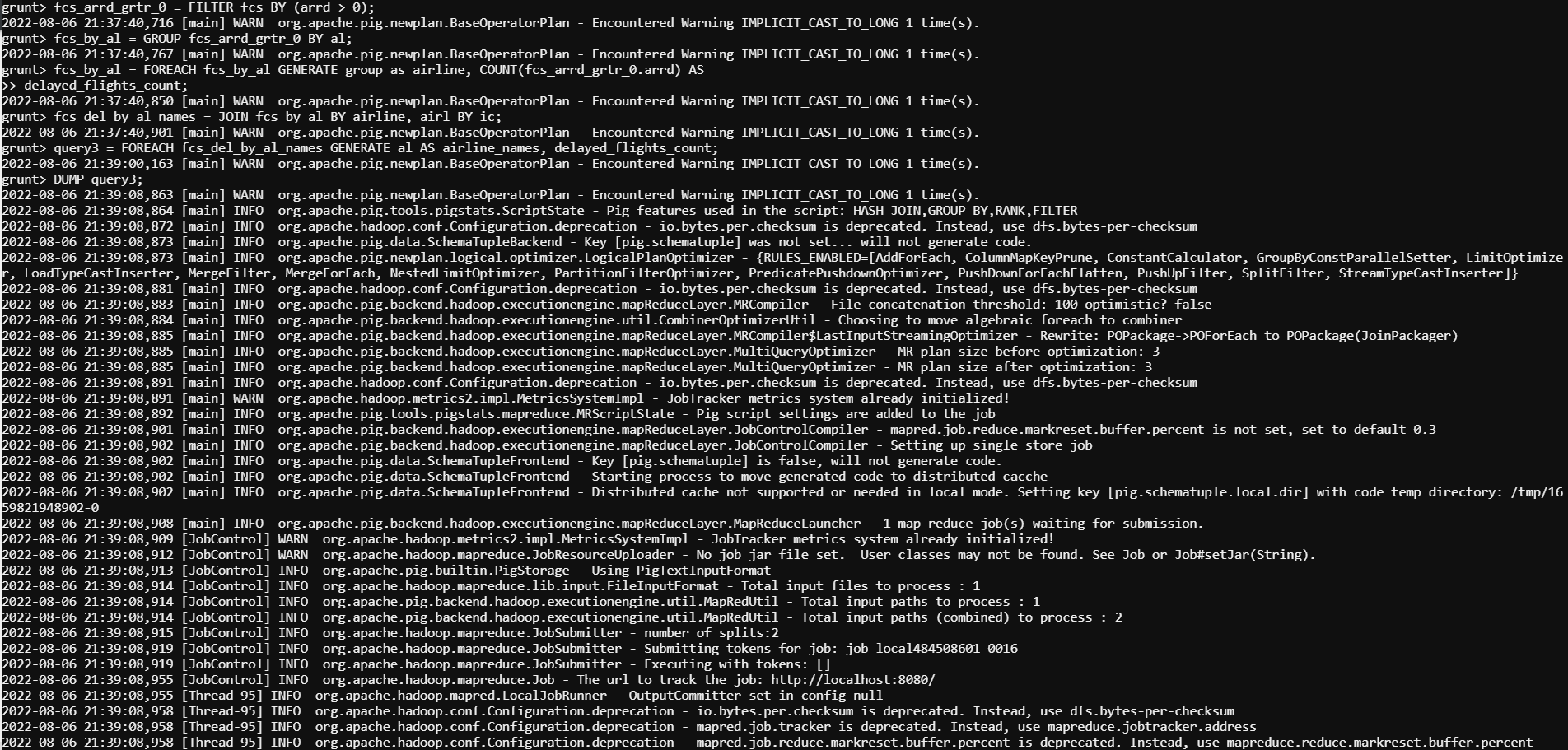
fcs\_arrd\_grtr\_0 = FILTER fcs BY (arrd > 0);

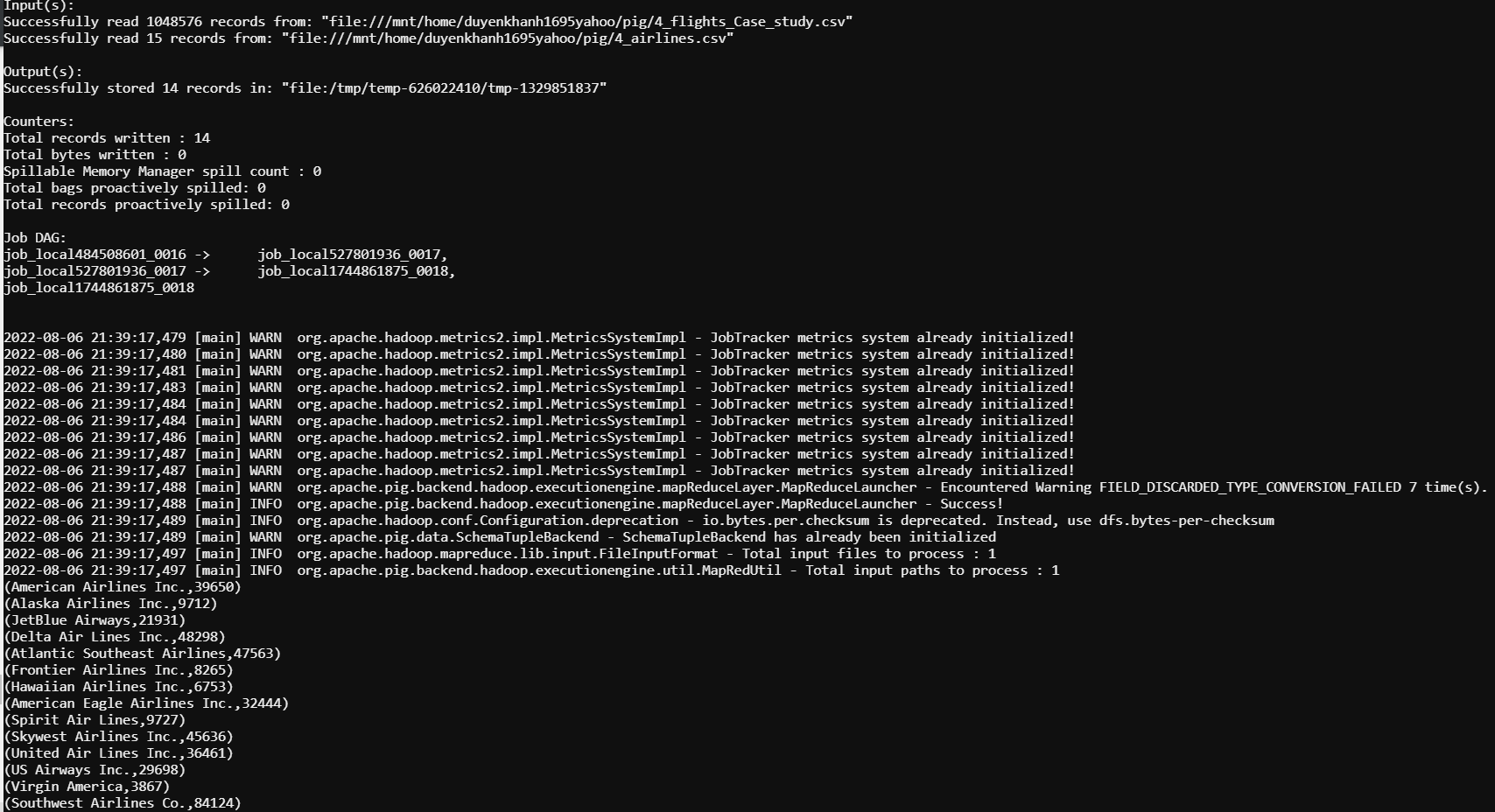
fcs\_by\_al = GROUP fcs\_arrd\_grtr\_0 BY al;

fcs\_by\_al = FOREACH fcs\_by\_al GENERATE group as airline, COUNT(fcs\_arrd\_grtr\_0.arrd) AS delayed\_flights\_count;

fcs\_del\_by\_al\_names = JOIN fcs\_by\_al BY airline, airl BY ic;

query3 = FOREACH fcs\_del\_by\_al\_names GENERATE al AS airline\_names, delayed\_flights\_count;







**Store Data:**

Store query1 INTO ‘/mnt/home/duyenkhanh1695yahoo/pig/pig\_Output\_2’ USING PigStorage(‘,’);

