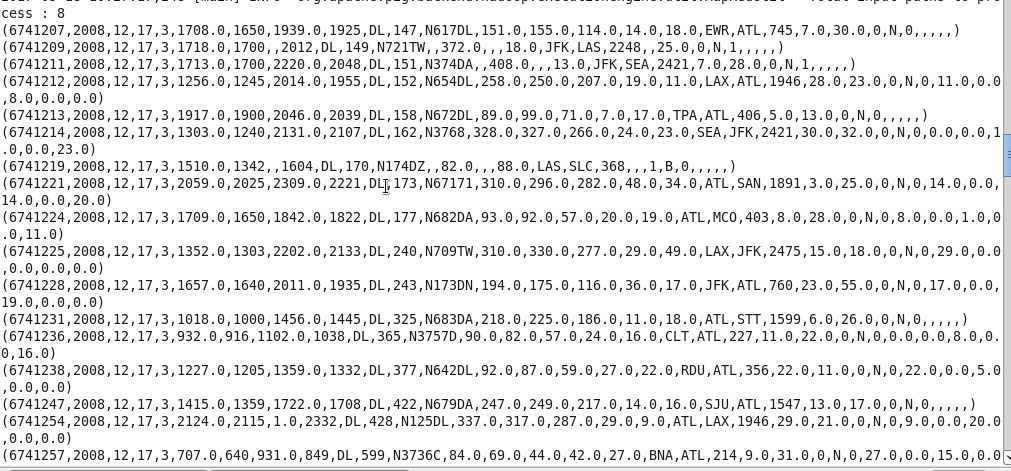
**Problem Statement 1**

# Find out the top 5 most visited destination

REGISTER ‘/path/piggybank.jar’;

1. sourceFile = LOAD 'DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

--First we will load our DelayedFlights.csv file using load function CSVExcelStorage , This



1. fetchDest = FOREACH sourceFile GENERATE $18 as(destination :bytearray) , 1 as (val : int);

--Generate the details about dest from relation sourceFile which is at index $18 , we will generate int 1 for each records so as to group with it

1. groupDest = GROUP fetchDest by $0 ;

-- --groupDest: {group: chararray,fetchDest: {(destination: chararray)}}

Group the relation fetchDest by dest

1. countDest = FOREACH groupDest GENERATE group as (dest : bytearray) , COUNT(fetchDest) as (destCount : long);

-- Get the destination and the count the number of destinations

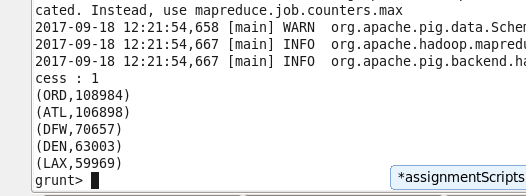
5) orderDest = ORDER countDest by destCount DESC;

-- Sort the relation countDest by descreasing order



1. Top5Dest = LIMIT 5 orderDest;

--Get the top 5 destinations using limit operator



1. airportFile = LOAD 'airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

--Load airport data and skip the header

8) cityInfo = FOREACH airportFile GENERATE $0 as (dest : bytearray) , $2 as (city : chararray ), $3 as (state :chararray) ,$4 as (country : chararray);

-- Fetch the required details from airportFile relation

**Join Top5Dest with cityInfo:**

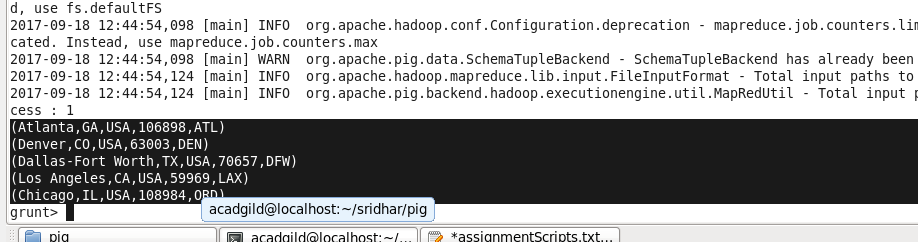
9) joinDest = JOIN cityInfo by $0 RIGHT OUTER, Top5Dest by dest ;

--joinDest: {cityInfo::dest: bytearray,cityInfo::city: chararray,cityInfo::state: chararray,cityInfo::country: chararray,Top5Dest::dest: bytearray,Top5Dest::destCount: long}

-- Perform right outer join between cityInfo and Top5Dest to get the location details

10) Top5Destinations = FOREACH joinDest GENERATE cityInfo::city , cityInfo::state , cityInfo::country , Top5Dest::destCount, Top5Dest::dest ; **(Final output)**

-- Fetch the Location details



**Problem Statement 2**

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

1. sourceFile = LOAD 'DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

--Load the source file

1. filterByCancellation = FILTER sourceFile by $22 == 1 AND $23 == 'B';

--Filter the records where cancellationcode is 1(TRUE) which is at index 22 and CANCELLATION is done due to bad weather (B) which is at index location 23

1. removeJunkRecords = FILTER filterByCancellation by $2 is not null

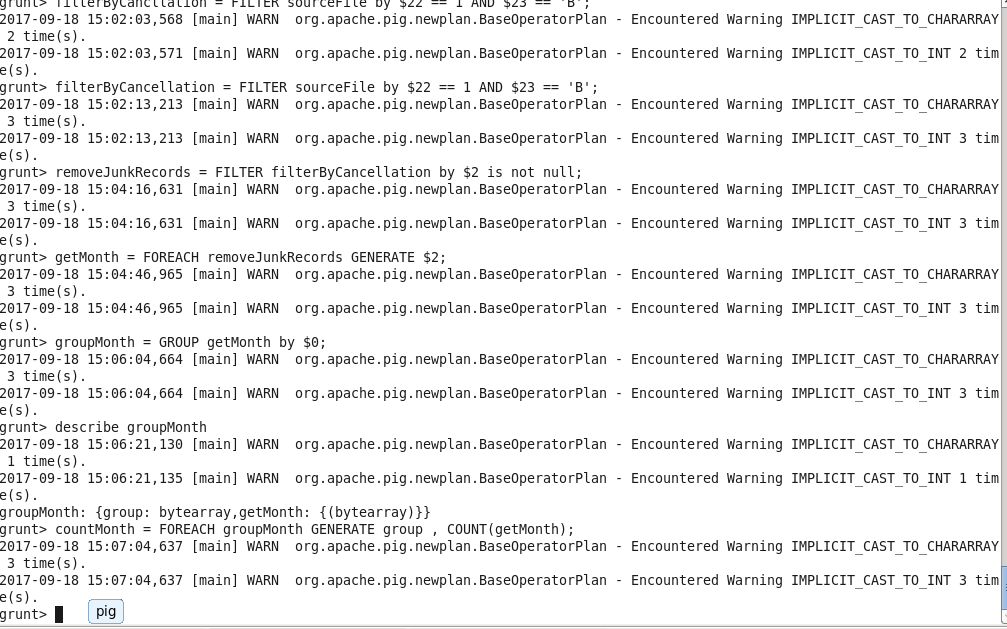
--Remove any junk records where null exists

1. getMonth = FOREACH removeJunkRecords GENERATE $2;

--Fetch the months from the relation

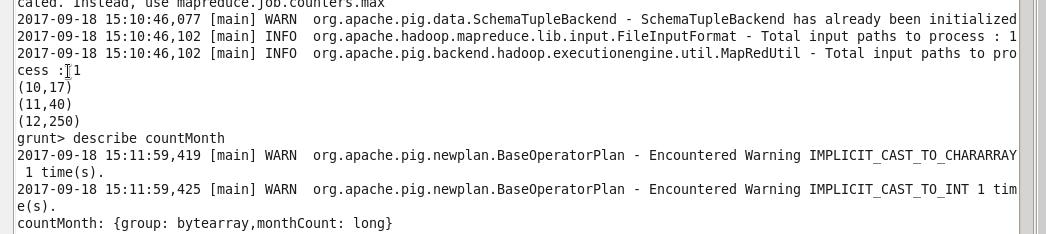
1. groupMonth = GROUP getMonth by $0;

--group getMonth by months fetched



1. countMonth = FOREACH groupMonth GENERATE group , COUNT(getMonth) as (monthCount : LONG);

--Count the number of months for each month

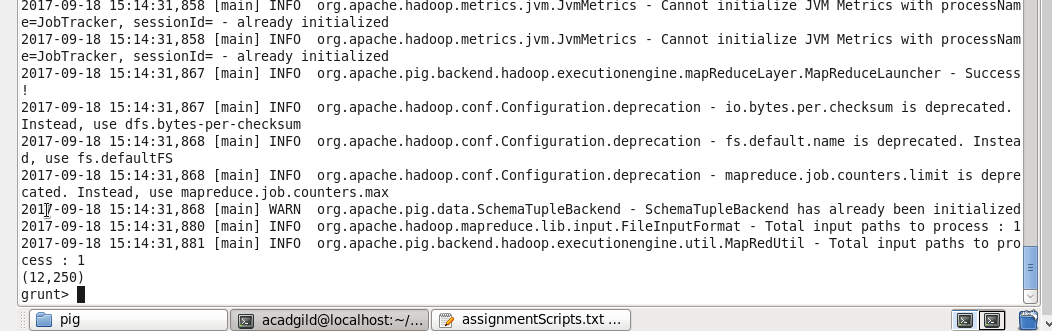


1. orderMonth = ORDER countMonth by monthCount desc;

--sort the relation in descending order of monthCount to get the hishest month where cancellation is done

1. MaxMonth = LIMIT orderMonth 1; **(Final output)**

--Get the month where highest cancellation is done due to bad weather



**Problem Statement 3**

# Top ten origins with the highest AVG departure delay

1. sourceFile = LOAD 'DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

--First load the file using CSVExcelStorage

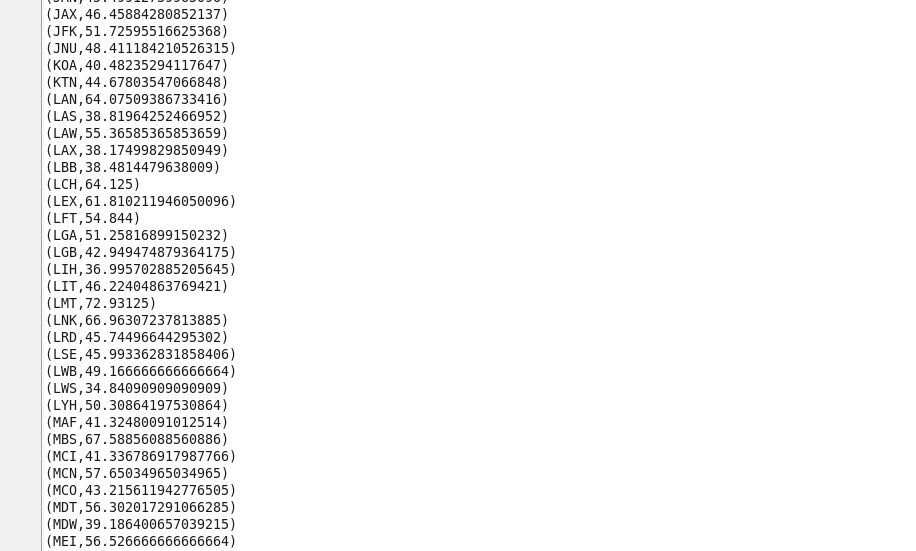
1. getOrigin = FOREACH sourceFile GENERATE $16 as (debDelay : bytearray ), $17 as (origin : bytearray);

--Fetch the Origin which is at index $17 and dep delay which is at index $16 from the relation

1. groupOrigin = GROUP getOrigin by $1;

--groupOrigin: {group: bytearray,getOrigin: {(debDelay: bytearray,origin: bytearray)}}

Group the relation getOrigin by origin so as to calucalte the avg on depDelay for each origin

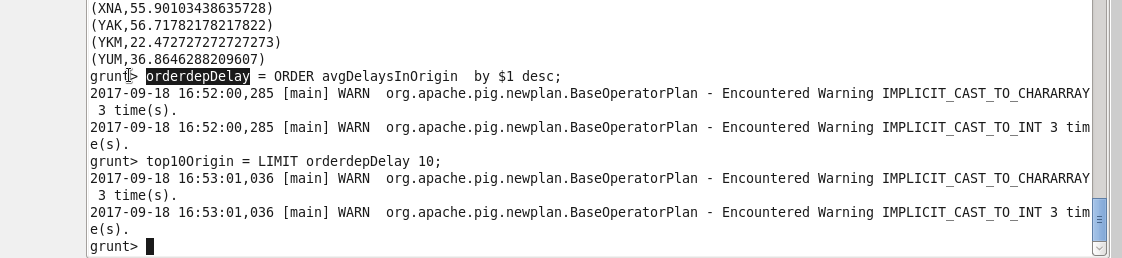


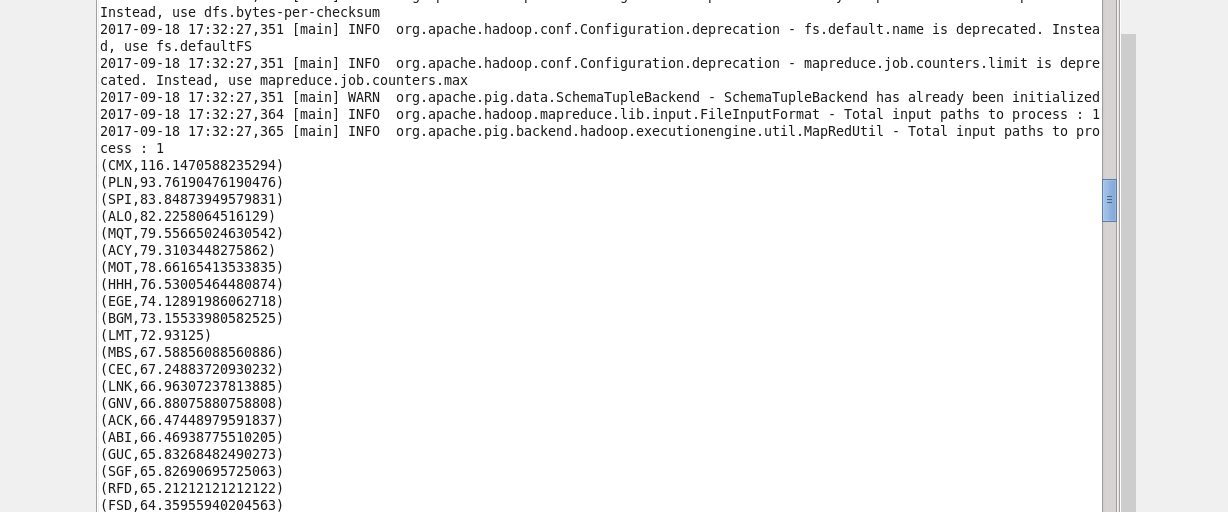
1. avgDelaysInOrigin = FOREACH groupOrigin GENERATE (chararray)group , AVG(getOrigin.debDelay) as (depDElay : double) ;

--count the average of depDelay for each origin

1. orderdepDelay = ORDER avgDelaysInOrigin by $1 desc;

--sort the relation avgDelaysInOrigin by descending of avg

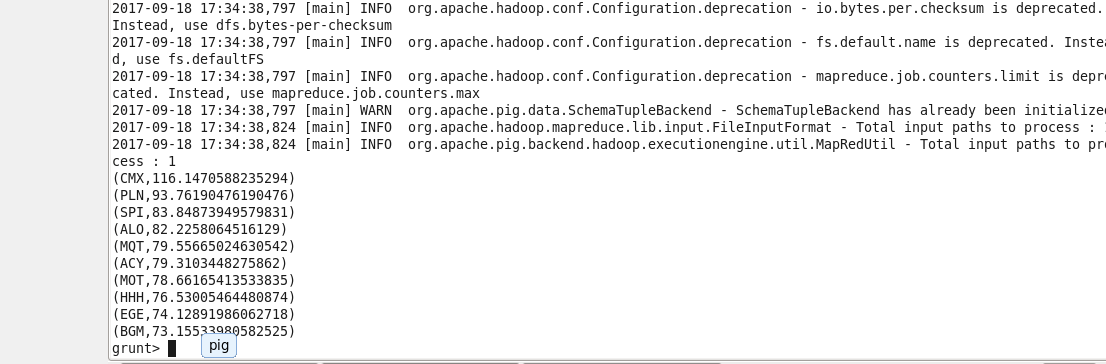




1. top10Origin = LIMIT orderdepDelay 10; **(Final output)**

--Fetch the 10 ten origin by using the limit operator

**top10Origin: {group: chararray,depDElay: double}**



**Problem Statement 4**

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

1. sourceFile = LOAD 'DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

--Load the source file using CSVExcelStorage

2) FetchRoutes = FOREACH sourceFile GENERATE $17 ,$18, $24 ;

--Fetch the origin which is at index $17 . destination which is at $17 and diversion which is at index $24

1. filterDiverted = FILTER FetchRoutes by ( $0 is not null )AND ($1 is not null) AND ($2 == 1 ) ;

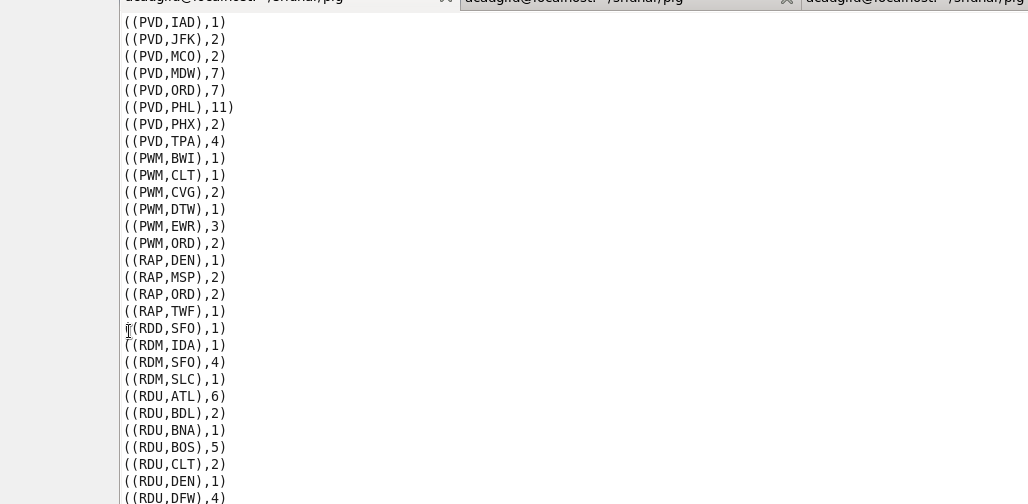
--filter the junk records where origin and dest is null and where diversion is true

1. groupRoutes = GROUP filterDiverted by($0,$1);

--group relation with origin and destination

1. routeCount = FOREACH groupRoutes GENERATE group , COUNT(filterDiverted);

--Generate the number of diversions taken with repsect to each route



1. orderRouteCount = ORDER routeCount by $1 DESC;

--order the relation in descending order of route count to get the highest diverions

1. topDiverted = LIMIT orderRouteCount 1; **(Final output)**

--fetch the top diverted route

