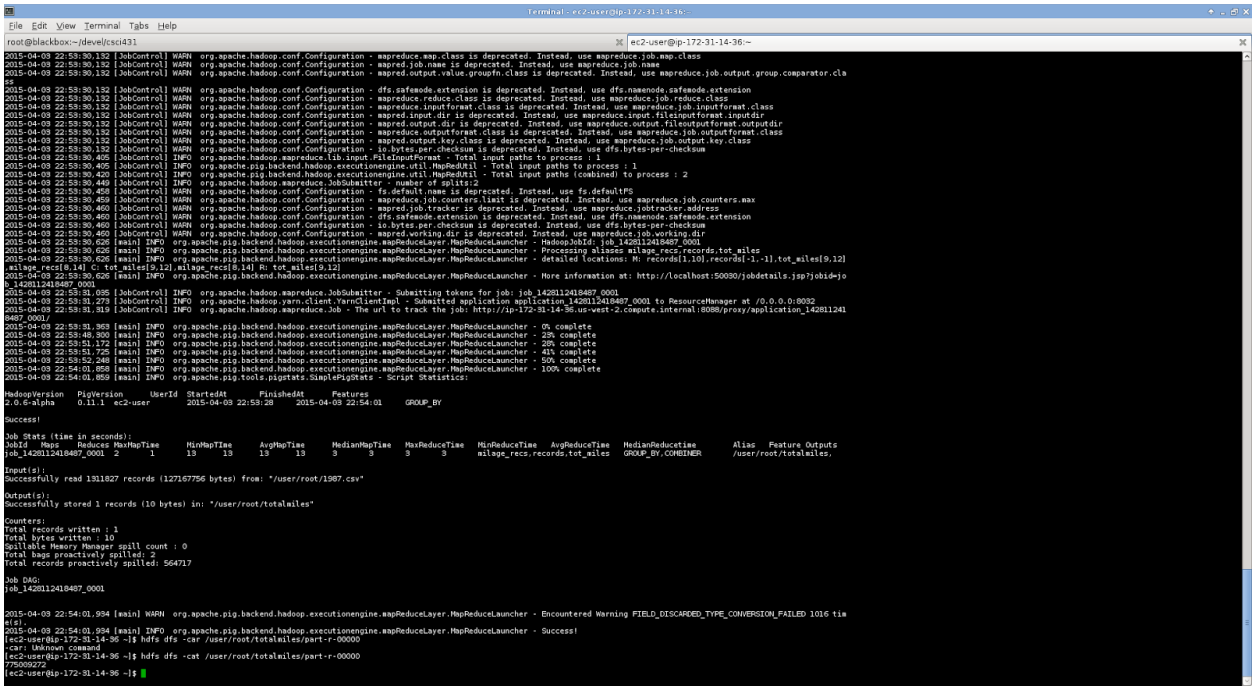
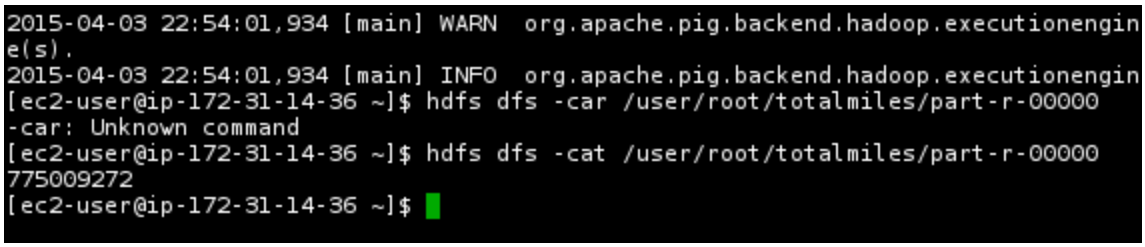
Jason Boutte
CINS 370
4/3/2015

Assignment #7

1. The number generated by the command in step 16 represents the total number of miles traveled, or the sum of the distance field of all rows in the relation. The command achieves this by creating a relation using the LOAD command then iterating through each row of the relation summing the value of the Distance field, finally writing the resulting relation to a file.





2. records = LOAD '/user/root/1987.csv' USING PigStorage(',') AS
(Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,
CRSArrTime,UniqueCarrier:chararray,FlightNum,TailNum,ActualElapsedTime:in,CRSElapsedTime,AirTime,ArrDelay,
DepDelay,Origin,Dest, Distance,TaxiIn,TaxiOut,Cancelled,CancellationCode,
Diverted,CarrierDelay,WeatherDelay,NASDelay,SecurityDelay, LateAircraftDelay);
carriers = GROUP records BY UniqueCarrier;
no_header = FILTER carriers BY group != 'UniqueCarrier';
carrier_count = FOREACH no_header GENERATE group, COUNT(records.UniqueCarrier) AS count;
carrier_sum_time = FOREACH carriers GENERATE group, SUM(records.ActualElapsedTime) AS time;
joined = JOIN carrier_count BY group, carrier_sum_time BY group;
result = FOREACH joined GENERATE carrier_count::group, time/count;
STORE result INTO '/user/root/avg_time';

   This command generates a final relation showing the UniqueCarrier and the Average Time for each carrier. It creates a relation of all the data, groups by the value for UniqueCarrier, filters out the header row, generates a count for each UniqueCarrier,

generates a sum for ActualElapsedTime, joins the two relations by the group field (UniqueCarrier), and finally a relation is created containing the UniqueCarrier and the time divided by the number of flights giving the average time per unique carrier.