

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.

```
File Edit View Terminal Tabs Help
Terminal - ec2-user@ip-172-31-14-36 ~
root@blackbox:~# cd /dev/shm
2015-04-03 22:53:50.132 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.map.class is deprecated. Instead, use mapreduce.job.map.class
2015-04-03 22:53:50.132 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.job.map is deprecated. Instead, use mapreduce.job.map
2015-04-03 22:53:50.132 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.output.value.groupfn.class is deprecated. Instead, use mapreduce.job.output.value.groupfn.class
2015-04-03 22:53:50.132 [JobControl] WARN org.apache.hadoop.conf.Configuration - dfs.safemode.extension is deprecated. Instead, use dfs.namenode.safemode.extension
2015-04-03 22:53:50.132 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.reduce.class is deprecated. Instead, use mapreduce.job.reduce.class
2015-04-03 22:53:50.132 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.inputformat.class is deprecated. Instead, use mapreduce.job.inputformat.class
2015-04-03 22:53:50.132 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
2015-04-03 22:53:50.132 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.outputformat.class is deprecated. Instead, use mapreduce.job.outputformat.class
2015-04-03 22:53:50.132 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.outputkey.class is deprecated. Instead, use mapreduce.job.outputkey.class
2015-04-03 22:53:50.132 [JobControl] WARN org.apache.hadoop.conf.Configuration - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2015-04-03 22:53:50.405 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.HadoopUtil - Total input paths to process : 1
2015-04-03 22:53:50.405 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.HadoopUtil - Total input paths (combined) to process : 2
2015-04-03 22:53:50.449 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Number of splits: 2
2015-04-03 22:53:50.458 [JobControl] WARN org.apache.hadoop.conf.Configuration - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-04-03 22:53:50.458 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max
2015-04-03 22:53:50.460 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2015-04-03 22:53:50.460 [JobControl] WARN org.apache.hadoop.conf.Configuration - dfs.safemode.extension is deprecated. Instead, use dfs.namenode.safemode.extension
2015-04-03 22:53:50.460 [JobControl] WARN org.apache.hadoop.conf.Configuration - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2015-04-03 22:53:50.460 [JobControl] WARN org.apache.hadoop.conf.Configuration - mapreduce.working.dir is deprecated. Instead, use mapreduce.job.working.dir
2015-04-03 22:53:50.626 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - HadoopUtil: job_1428112418487_0001
2015-04-03 22:53:50.626 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - Processing alias: wllage_recc_records,tot_miles
2015-04-03 22:53:50.626 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - Detailed locations: M: records[1..10], records[1..1], tot_miles[9..12]
2015-04-03 22:53:50.626 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - More information at: http://localhost:50030/jobdetails.jsp?jobid=job_1428112418487_0001
2015-04-03 22:53:51.095 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1428112418487_0001
2015-04-03 22:53:51.319 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-172-31-14-36.us-west-2.compute.internal:8088/proxy/application_1428112418487_0001
2015-04-03 22:53:51.363 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - 0% complete
2015-04-03 22:53:51.363 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - 20% complete
2015-04-03 22:53:51.363 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - 40% complete
2015-04-03 22:53:52.248 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - 60% complete
2015-04-03 22:54:01.859 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - 80% complete
2015-04-03 22:54:01.859 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - 100% complete
2015-04-03 22:54:01.859 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
3.0.6-alpha 0.11.1 ec2-user 2015-04-03 22:53:28 2015-04-03 22:54:01 GROUP_BY
Success!
Job Stats (time in seconds):
JobID Prog Reduce BytesWritten HdfsMapTime AvgMapTime HdfsReduceTime AvgReduceTime MedianReduceTime Alias Feature Outputs
job_1428112418487_0001 2 1 13 13 13 9 9 9 wllage_recc_records,tot_miles GROUP_BY,COMBINEER /user/root/totallmiles,
Input(s):
Successfully read 1381827 records (12747756 bytes) from: "/user/root/1987.csv"
Output(s):
Successfully stored 1 records (10 bytes) in: "/user/root/totallmiles"
Counters:
Total records written: 1
Total bytes written: 10
Total Hadoop Heavy Hanger spill count: 0
Total bags proactively spilled: 2
Total records proactively spilled: 564717
Job DAG:
job_1428112418487_0001
2015-04-03 22:54:01.994 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopUtil - Encountered Warning: FIELD_DISCARDED_TYPE_CONVERSION_FAILED 1016 tia
[ec2-user@ip-172-31-14-36 ~]$ hdfs dfs -cat /user/root/totallmiles/part-r-00000
-cat: Unknown command
[ec2-user@ip-172-31-14-36 ~]$ hdfs dfs -cat /user/root/totallmiles/part-r-00000
775009272
[ec2-user@ip-172-31-14-36 ~]$
```

```
2015-04-03 22:54:01,934 [main] WARN org.apache.pig.backend.hadoop.executionengin
e(s).
2015-04-03 22:54:01,934 [main] INFO org.apache.pig.backend.hadoop.executionengin
[ec2-user@ip-172-31-14-36 ~]$ hdfs dfs -car /user/root/totallmiles/part-r-00000
-car: Unknown command
[ec2-user@ip-172-31-14-36 ~]$ hdfs dfs -cat /user/root/totallmiles/part-r-00000
775009272
[ec2-user@ip-172-31-14-36 ~]$
```

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,

Assignment #7

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.