**MAPREDUCE/PIG/HIVE:**

**GROUP MEMBERS:**

HUSSAN MUSAFER 917307

ARUNA VEERAMREDDY 1004293

SANDEEP GUNTUPALLI 1002318

**EXECUTION PROCEDURE:**

**MAPREDUCE:**

MapReduce contains three programs the reducer, the Mapper and the Driver.

The is uploaded using the command “-put”

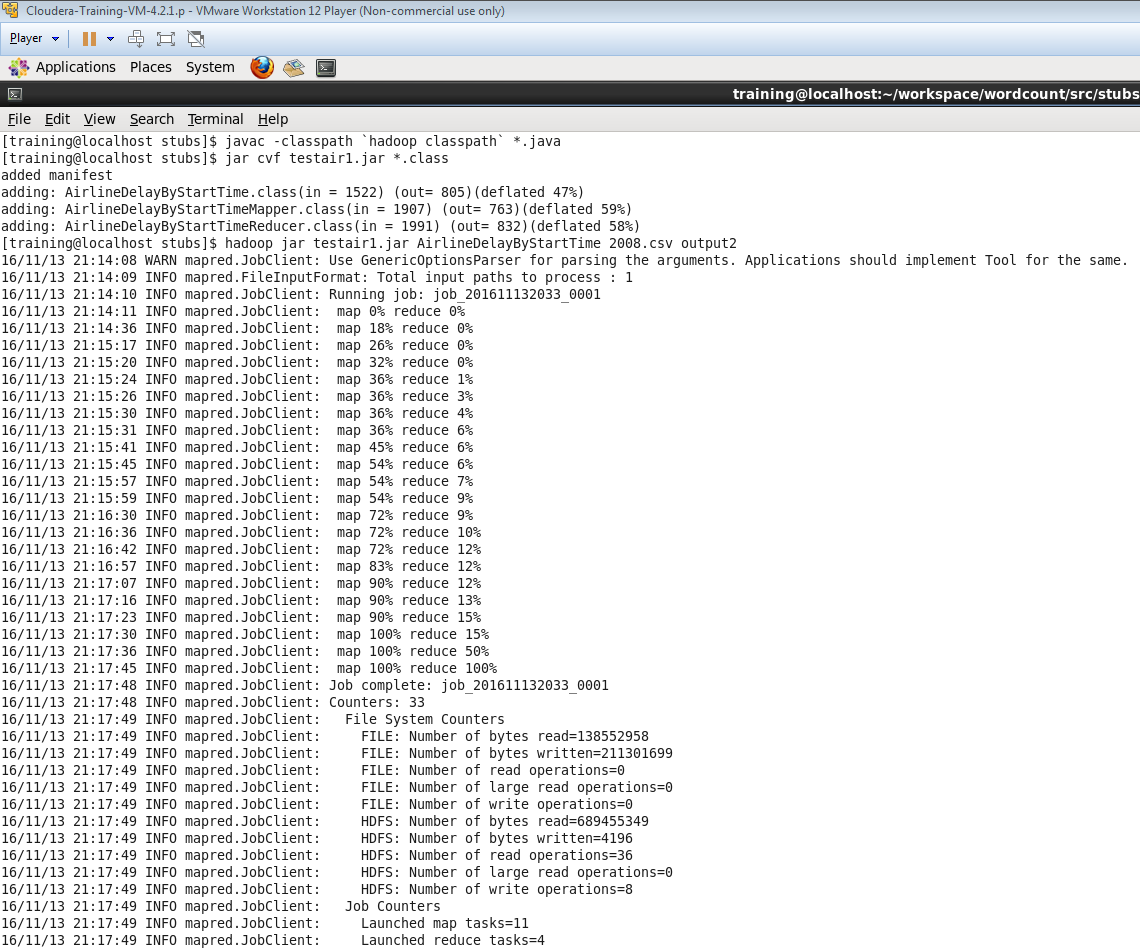
The jar file is created.

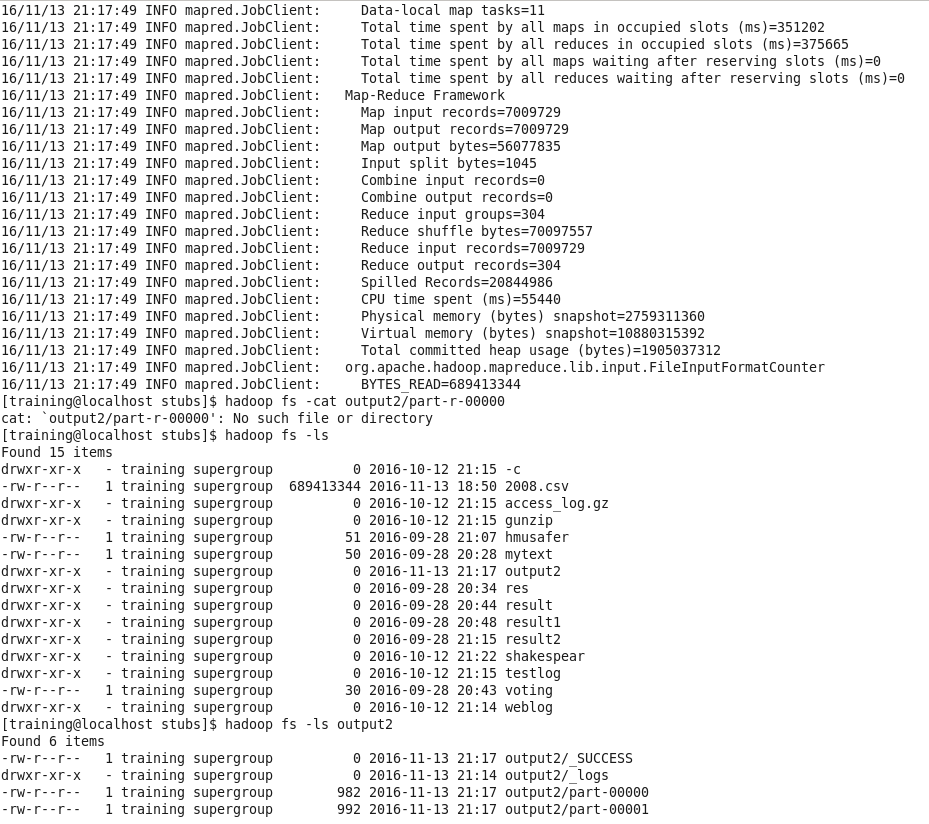
The jar file creates the executed class files for each program. The jar is placed in the program sorce location and completed.

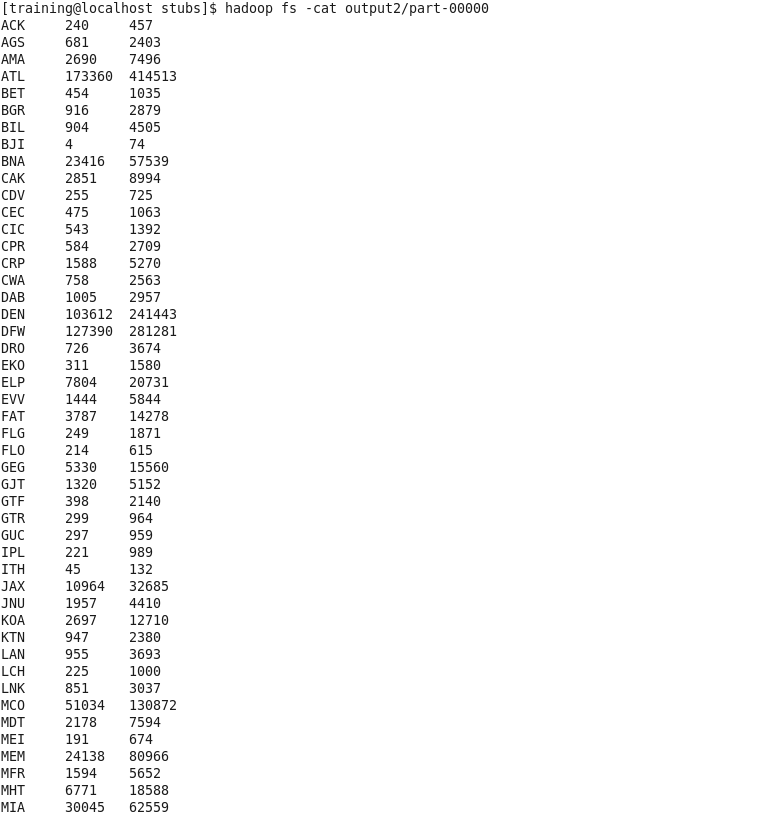
**Query**:

To show the origin flight,total number of flight and delay flights.

The screenshot of the execution is as follows:







**Hive:**

**To Create a external table in HIVE and to ingest the data(.CSV file) into the table that we hacve created:**

**As we have created a seperate database for our data, we use the following command to use that database.**

**use group3;**

**To create a external table, flightdata we used the following create table statement. As we do not need the header, we use the skip.header in this statement.**

**create external table group3.flightdata**

**(Year int, Month int, Dayofmonth int, Dayofweek int,**

**DepTime int, CRSDepTime int, ArrTime int, CRSArrTime int,**

**Uniquecarrier string, Flightnum int, Tailnum string, ActualElapsedTime int,**

**CRSElapsedTime int, AirTime int, ArrDelay int, DepDelay int, Origin string,**

**Dest string, Distance int, TaxiIn int, TaxiOut int,**

**Cancelled int, CancellationCode string,**

**Diverted string, CarrierDelay string, WeatherDelay string,**

**NASDelay string, SecurityDelay string,**

**LateAircraftDelay string) row format delimited fields**

**terminated by ',' lines terminated by '\n' location**

**'/user/group3/' tblproperties ("skip.header.line.count"="1");**

**The path where we saved the .CSV file is**

**hdfs dfs -ls /user/group3/new\_2008.csv**

**Inorder to load the data into the table that we have created in the Hive, we use the following statement.**

**load data inpath '/user/training/new\_2008.csv' into table flightdata;**

**By the above mentioned steps, we are able to ingest the data in to the table that we have created in the Hive.**

**we have run the following sample queries :**

**select AVG(Airtime) from flightdata;**

**The Result is as follows:**

**MapReduce Total cumulative CPU time: 26 seconds 580 msec**

**Ended Job = job\_201611101759\_0012**

**MapReduce Jobs Launched:**

**Job 0: Map: 3 Reduce: 1 Cumulative CPU: 26.58 sec HDFS Read: 689454877 HDFS Write: 18 SUCCESS**

**Total MapReduce CPU Time Spent: 26 seconds 580 msec**

**OK**

**104.0185891263188**

**Time taken: 83.943 seconds**

**PIG LATIN:**

**To Ingest the data into pig, we use the following Code:**

**a = LOAD '/user/group3/new\_2008.csv' using PigStorage(',') as (Year: int, Month: int, Dayofmonth: int, Dayofweek: int,**

**DepTime: int, CRSDepTime: int, ArrTime: int, CRSArrTime: int,**

**Uniquecarrier: chararray, Flightnum: int, Tailnum: chararray, ActualElapsedTime: int,**

**CRSElapsedTime: int, AirTime: int, ArrDelay: int, DepDelay: int, Origin: chararray,**

**Dest: chararray, Distance: int, TaxiIn: int, TaxiOut: int,**

**Cancelled: int, CancellationCode: chararray,**

**Diverted: chararray, CarrierDelay: chararray, WeatherDelay: chararray,**

**NASDelay: chararray, SecurityDelay: chararray,**

**LateAircraftDelay: chararray);**

**Now the CSV file is uploaded into PIG**

**We have run the following Sample Queries :**

**b = foreach a generate Origin as O, Dest as D;**

**c = group b by (O,D);**

**COUNT = foreach c generate group, count(b);**

**dump COUNT;**

**We have got the following sample output:**

**((ABE,ATL),853)**

**((ABE,BHM),1)**

**((ABE,CLE),805)**

**((ABE,CLT),465)**

**((ABE,CVG),247)**

**((ABE,DTW),997)**

**((ABE,JFK),3)**

**((ABE,LGA),9)**

**((ABE,ORD),1425)**

**((ABE,PHL),2)**

**((ABI,DFW),2660)**

**((ABQ,AMA),368)**

**((ABQ,ATL),1067)**

**((ABQ,AUS),433)**

**((ABQ,BWI),546)**