**Ankit Yadav**

College of engineering  Northeastern University

**Final Project**

**INFO7250- Engg Of Big Data**

**CONTENTS**

**1. Analysis of Flight Data using MAP REDUCE on Hadoop**

**1- Getting total count of all the data:**

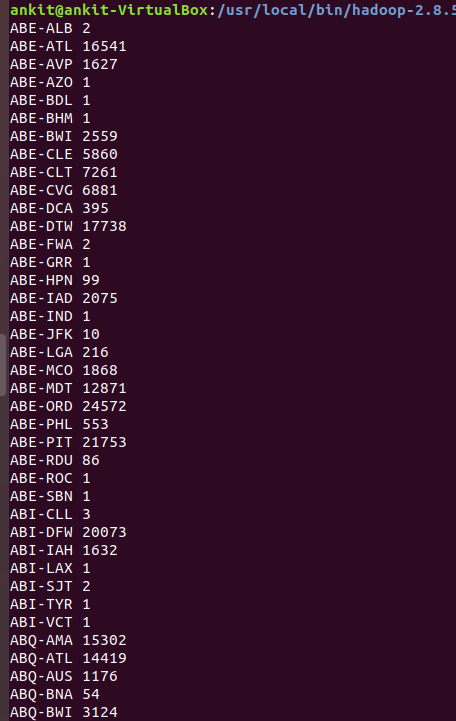
This is a very basic map reduce use case in which we count the whole data to get a sense of how many total records are there:

hadoop jar ~/Downloads/ProjectJars/count.jar hadoop.project.total\_count.MRCount /flight-data /FinalProjectMROutput/2-Total-Data-Count

The final count is: **123534970**

**2- Getting the total flights from all source destinations pairs in from 1987 to 2008:**

This was a huge data and MapReduce made this analysis quite simple and fast:



**3: Top 30 source destination pairs**

Sorting the above data to get top 30 most busy Source Destination pair:



**4: Delay in flight percentage**

We considered the delay greater than or equal to 15 minutes as delay . Now we need to count those flights which had delay greater than or equal to 15 minutes.

Delayed flight Count:

Percentage of departure delayed flights: Total Flight Count/ Departure Delayed Flight count

**= (19690422/123534970) \* 100 = 15.94 %**

So, this shows that the actual delay greater that 15 minutes is very less and generally flights depart on time.

Let’s check the same for arrival delay

Percentage of departure delayed flights: Total Flight Count/ Delayed Flight count

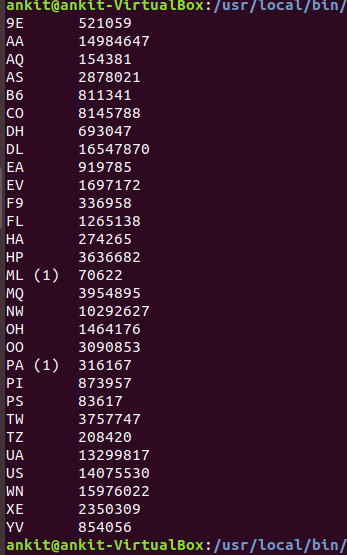
**= (24627925/123534970) \* 100 = 19.9 %**

**So, the delay in departure and arrival is between 15 to 20 % range.**

**So, it shows that overall flights are mostly on time from/to all source destination**

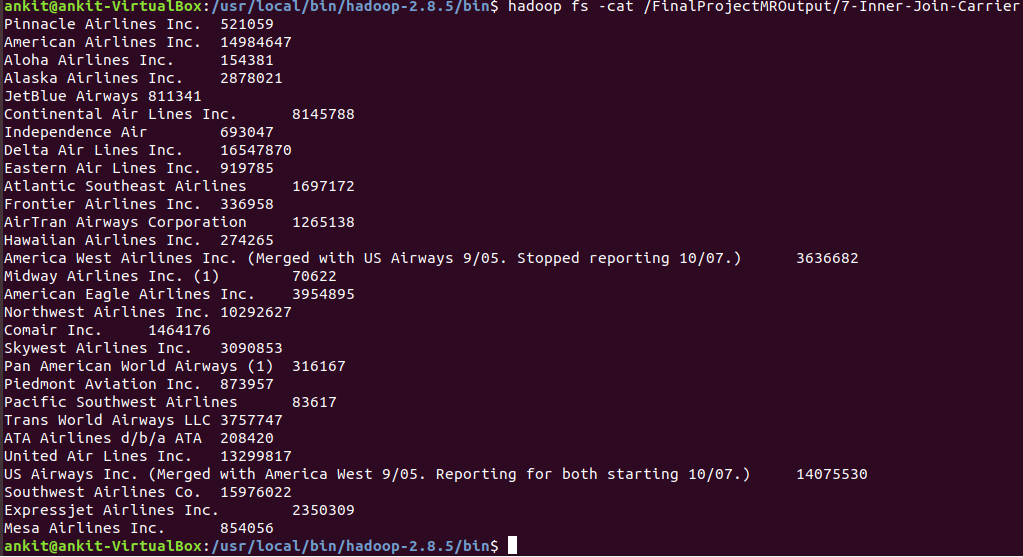
**5- Count of unique carrier’s flights**

The data for unique carriers are as follows:

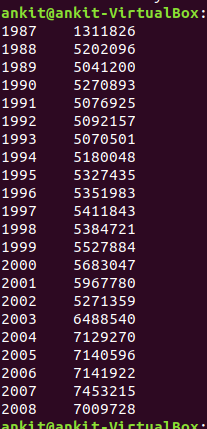


**6- Inner Join to get the full name for unique carriers**

We did inner join with between two files to get carrier names instead of carrier codes

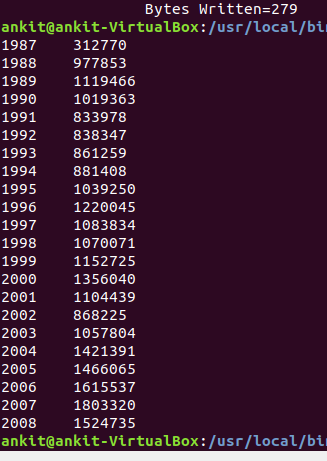


**7- Getting Flight data by year**

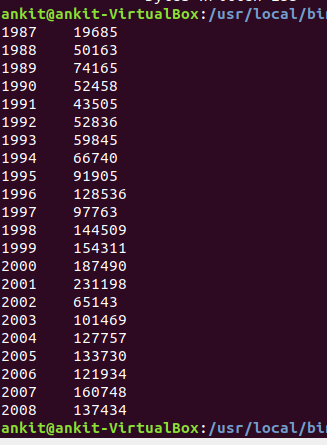


**8- Delayed flights per year**

In this we will check delayed flights per year( we will count flights as delayed only is the delay time if greater than equal to 15 minutes)



**9- Cancelled flights by year**



**10- Ratio of delayed flight and cancelled flights per year to total flights**

**11- Total flights by day of week and ratio to delayed and cancelled**

**12- Total flights by months of year and ratio to delayed and cancelled**