Quantitative Analysis of Airline Dataset

Apurv Deshpande
School of Computing
Masters in Cloud Computing
National College Of Ireland
Email: Apurv.Deshpande@student.ncirl.ie

***Abstract*—Airline industry is fast growing as number of people are turning towards Air Transport as it fast, reliable mode of transport. With the ever increasing growth in Airline Industry there comes a question how this system can be more efficient to facilitate the needs of people and how Airline industry can further be enhanced . The following research is a motivation of that. The following analysis research covers the following three main tasks: First task is the analysis to find the top ten busiest airport in USA. The second task is to find the number of flights grouped by month and year filter. The third task is to calculate minimum airtime between two airports in USA.**

***Keywords*—Airline,Flights,Airports,Data Analysis**

I. Introduction

The Airline continues to grow at a phenomenal rate so it is the time to take a modern look in this industry. Hence data analytics approach could be taken in this industry to understand the semantics and help grow the Airline Industry. The tasks that I am going to carry out in this research include the number of flights that originates from a particular Airport, the number of flights analysis filtered by month, year and minimum airtime between source and destination airports.

A. Research Question

How data analysis will help the growth of Airline Industry and increase the scalability Of Airline Industry. How this analysis will help in solving the common problems like Flight Delays, overcrowded flights, empty seats, and the needs of some airports which are unable to handle the passenger capacity at Airports and to faciliate the increasing number of flights at some airports.

II. Related Work

The project work was inspired from the competition on Airline on-time performance which was carried out in 2009 by Data expo. The participants in this competition were given data sets of commercial airlines in USA and they have to carry out a series of analysis on this dataset like what is the best month or week in a year to fly which will cause minimum delays or is a weather is an important factor for delays. The most interesting task according to me was to find out how delays at one particular airport were creating delays at another tasks. The dataset provided for this competition was from October 1987 to April 2008. The records total was nearly 120 million. Also some supplement data was provided like airport records which consist of list of airports in USA with their city and state location and their name. The competitors in this competition provided a variety of results out of which the standout result that one competitor predicted was that the best time in the day to fly was early in the morning and to avoid evening flights from 5pm to 7pm. Some of the results found out by the competitors were very obvious like air traffic increased form 1987 to 2007. Another impressive thing about this competition was that the competitors were provided with the basic knowledge for example, how to create table and import data in database like Sqlite and R. Also basic command line tools were also provided. Due to this the gap between the competitors with high programming skills and with the one who has less programming skills were substantially reduced. This gave all the competitors equal opportunity at starting level.

III. Methodology

The dataset used for this programming activity and research consist of 30 columns. The first column is a auto generated incremental RowNum primary key .The description of rest of variables used in dataset is as follows: Name Description

1 Year 2007-2008
2 Month 1-12
3 DayofMonth 1-31
4 DayOfWeek 1 (Monday) - 7 (Sunday)
5 DepTime actual departure time (local, hhmm)
6 CRSDepTime scheduled departure time (local, hhmm)
7 ArrTime actual arrival time (local, hhmm)
8 CRSArrTime scheduled arrival time (local, hhmm)
9 UniqueCarrier unique carrier code
10 FlightNum flight NUMBER
11 TailNum plane tail number
12 ActualElapsedTime in MINUTES
13 CRSElapsedTime in minutes
14 AirTime in MINUTES
15 ArrDelay arrival delay, in minutes
16 DepDelay departure delay, in minutes
17 Origin origin IATA airport code
18 Dest destination IATA airport code
19 Distance in miles
20 TaxiIn taxi in time, in minutes
21 TaxiOut taxi out time in minutes
22 Cancelled was the flight cancelled?
23 CancellationCode reason for cancellation (A = carrier, B = weather, C = NAS, D = SECURITY)
24 Diverted 1 = yes, 0 = no
25 CarrierDelay in minutes
26 WeatherDelay in minutes
27 NASDelay in minutes

28 SecurityDelay in minutes
29 LateAircraftDelay in minutes

This dataset was chosen because it has all the information about the Flights. This dataset will help to analyse the common problems associated with the airline like how many Flights are delayed in a year and what is the most common factor of delay of flights for example,Carrier Delay, Security Delay. It will help to analyse the effectiveness of current state of Airline System in USA and what factors need to be solved or improve to further strengthen the Airline System.

A. Preparation of Data to Analyse

The first problem with this dataset is that it has some values defined as not Applicable(NA). So I used SQL commands to delete those data. For example the following command was used to delete the 'NA' value of field14 of table Flight.
DELETE FROM Flight WHERE Field14='NA';
This command deleted the Not Applicable 'NA' fields from table.The first header row was deleted from the database using the following command
DELETE FROM Flight LIMIT 1;
The data didnt have a primary key so an auto incremental primary key was created called rownum. The data was thus ready to import into HDFS. Sqoop was used to import data from MYSQL database to HDFS.

B. Activities Carried

I carried out three programming tasks on this dataset. Two Map/Reduce tasks were carried out in JAVA and one in Hive. The first task carried out in Java was to count the number of flights originated from Origin Airports. The Algorithm used is as follows:
1)A mapper was defined with key as airport origin code and a counter was set which counts the re-occurrence of the particular airport which is our key.
2)Then the output of this mapper was passed to reducer.
3)Then the reducer made the count of number of flights on each airport code which is output of this analysis.
Further visualisation was done on this output to find out the top ten airports where most number of flights were originated.The result of this visualisation is shown in the graph below. This analysis was carried out to analyse and find the busiest airports in USA. This analysis will help to improve the facilities and flight management at this airports. It will help to drive attentions to airports which are busy and further help to increase the size of passenger carrying capacity of airports. It will also help to know if you have to built a new runway to manage the ever increasing number of flights. The second task that I carried out in Java was to count the number of flights which is grouped by two fields month and year. The Algorithm used is as follows:
1) In mapper a concatenation operator was used to concatenate the fields month and year. 2)The new string generated by concatenation was key and a counter was set which counts the re-occurrence of the combination which is our key.
3)Then the output of this mapper was passed to reducer.
4)Then the reducer made the count of number of flights on the combination string which is output of this analysis.
This task will help to analyse the busiest time in a year where there is more number of flights.This analysis will help to manage this peak periods of months. As the number of flights increased in particular month means there are more passengers travelling. This analysis will help to drive measures to increase flights when it is busy time of year and thus will solve the problems of overbooked flights and also empty seats in flight when the period of passenger travelling is less. The third task is carried out in Hive which is to calculate the minimum air time between the combination of Origin and Destination. The task was carried out in following way.
1)The field 14 is the AirTime field. Field 17 and Field 18 is origin and destination respectively.
2)A minimum operator was called on AirTime and it was grouped by Origin and Destination which gave the output as minimum airtime between various combinations of Source and Destination.
This task is carried out because it gives information about minimum time of travelling between source and destinations. This will help to analyse which airplanes are taking more time to travel between the same source and destination. hence by this analysis it will be clear whether older planes are taking more time to travel and it will also help passengers to choose from the airlines which are fast.

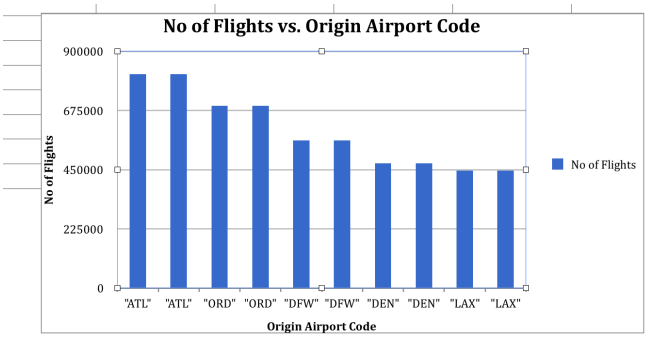
C. Technologies used

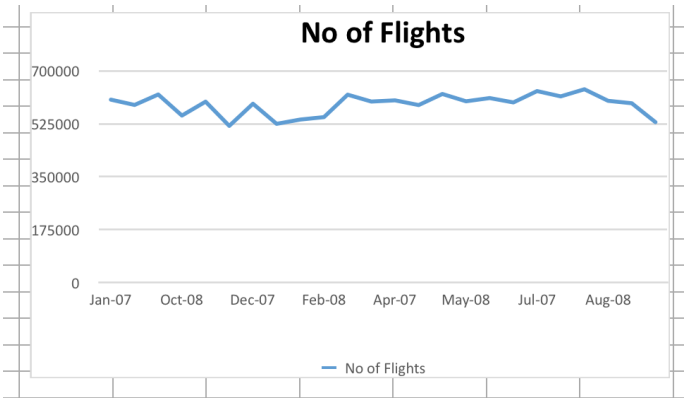
For storing the data at the start MYSQl is used. MYSQl was chosen because this database is easy to use and reliable. The commands like create, alter and delete in MYSQL is like talking of creation, change and deletion in everyday life. To make changes in data at prequisite position and making it ready for map reduce activity is easy in MYSQL. In my project Deletion of rows that were of no use like having 'NA' values was easy in MYSQL. Importing data from mysql is easy and it is one of the recommended databased to be used with HDFS. hence I used MySQL database.
Java is used because the mapreduce programs written in Java is fast compared to Pig and Hive. Another advantage of Java is that it is an open source universally accepted language. It is the first basic language in which Hadoop Map/Reduce were developed. Compared to other languages like Pig and Hive, Java give users the flexibility of knowing syntAx errors before the execution actual code. Hence I used Java Language.
I used hive because writing complex queries is easy in Hive. It is a language that is close to MYSQl queries. Hive makes the complex code look simple. It was developed and used in big organisation like Facebook. I was finding it difficult to write the code of finding minimum and then grouping it with other fields difficult in Java and so I used Hive.

IV. Results

The result of first task is as follows:

The figure shows the comparison of number of flights with origin airport code. It is found from the figure that the busiest airport by number of flights is ATL. And ATL,ORD,DFW,DEN,LAX are the top ten busiest airport in the country. But the number of flights the ATL carry is almost double of flight traffic Of LAX.
The result of second task is as follows:





The figure shows the number of flights grouped by Month and Year. The figure shows the not a huge spike between the number of flights. But still there is a gap of around 15,000 flights. The results shows that busiest time is between Feb2008 and April2008 while Sept2008 is least busiest.
The result of third task is that it shows the minimum airtime between two airports. But by looking at the output of the data I am finding some discrepancies in the output. In some of the results the minimum airtime between two airports is just 1 min which is not really possible.

V. Conclusion and Future Work

The following data analysis shows that as the year go pass by the number of passengers is increasing and so is the the number of flights. The ATL airport is the busiest airport in USA. The limitation of this analysis is some of the tasks weird results are coming as explained in Task 3 result.
The future work can be done in the section of delayed flights. The data provided has five different delays for example Weather Delay, Security Delay and Carrier Delay. An Analysis can be done on which factor including above mentioned is responsible for more number of delays. Also analysis can be done on the basis of airports like which airport is contributing for more number of delays and it can be filtered by delay type.