Assignment – Hive

**Task 1:**

Describe the dataset in your words. What is your basic understanding about this dataset?

**(25 marks)**

#Write your detailed answer here

Answer: We have the data of on-time performance of domestic flights operated by large carriers, The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the data on yearly basis and we have a subset of that data from the year 1987 to 2008. There data are saved in amazon s3 as CSV files and each year have separate CSV files, so we have 22 csv files which we have to merge in one table and provide some insights.

**Task 2:**

Create an external table in Hive using the given metadata where the suggested partition is using Year. (Hint the files with data are available inside **s3://ughdfsdemo/airlinedata/** bucket. You need to create your external table under **/tmp/<your\_DDA\_number>/**)

**(100 marks)**

#Write your command and specifically explain your view on selecting the datatypes for various fields.

Answer:

CREATE

EXTERNAL TABLE

DDA1610218.airline\_external (

YEAR INT

,MONTH INT

,DayofMonth INT

,DayofWeek INT

,DepTime string

,CRSDepTime string

,ArrTime string

,CRSArrTime string

,UniqueCarrier string

,FlightNum string

,TailNum string

,ActualElapsedTime string

,CRSElapsedTime string

,AirTime string

,ArrDelay INT

,DepDelay INT

,Origin String

,Dest String

,Distance INT

,TaxiIn string

,TaxiOut string

,Cancelled string

,CancellationCode string

,Diverted string

,CarrierDelay string

,WeatherDelay string

,NASDelay string

,SecurityDelay string

,LateAircraftDelay string

) row format delimited fields terminated BY ',' stored AS textfile location "s3://ughdfsdemo/tmp/DDA1610218/Airline/" TBLPROPERTIES (

"skip.header.line.count" = "1"

)

;

**I have use data type as integer for year, month, dateofmonths, ext.. for variable which are minutes we have use string as we are not sure about its format. We have also check the data sets and we have found that most of the variable which are looking as number are not reading as integer so we have use string for those as well.**

**Task 3:**

Find out the number of unique carriers from this data.

**(15 marks)**

#Write your command and comment for your command.

Answer: 30

select count(distinct uniquecarrier) as uniquecarriercount from DDA1610218.airline\_external;

#To count the unique carrier from this data we first use distinct on uniquecarrier variable so that it will give only unique carrier code and then we have use count to count the number of those unique carriers.

**Task 4:**

Find out the number of cancelled flights of carrier “9E”.

**(10 marks)**

#Write your command and comment for your command.

Answer: 10060.0

select sum(cancelled) as TotalCancelledFlight from DDA1610218.airline\_external where uniquecarrier = "9E";

#Since the cancelled variable is int type I have use sum to aggregate the cancelled flight where unique carrier is “=9E”.

**Task 5:**

Find out yearly air time for each flight provided by every flight provider.

**(25 marks)**

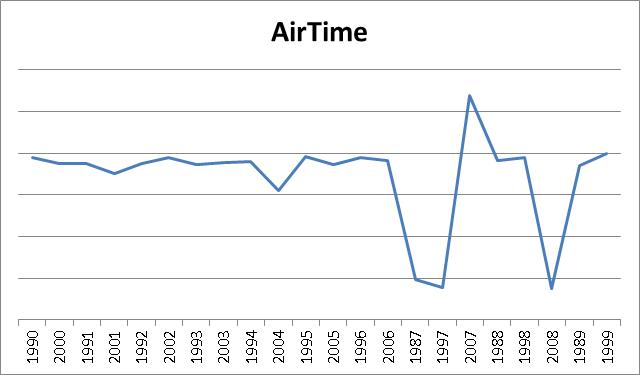
#Write your command and comment for your command.

Answer: select year,uniquecarrier,flightnum,sum(arrtime) as total\_airtime\_in\_min from DDA1610218.airline\_external group by year,uniquecarrier,flightnum;

Or

select year,sum(arrtime) as total\_airtime\_in\_min from DDA1610218.airline\_external group by year;

#Since we need to find out yearly airtime for every flight we have use sum to add the airtime of each flight by year, flight provider which is carrier and each flight which can be represented by flight number. From second query we see 2007 records the highest airtime.



**Task 6:**

How many distinct flights travel between IND to MCO and who are their carriers.

**(25 marks)**

#Write your command and comment for your command.

Answer: select uniquecarrier,count(distinct flightnum) as No\_Of\_Distinct\_Flight from DDA1610218.airline\_external where Origin = "IND" and Dest = "MCO" group by uniquecarrier;

#to count the distinct flight we have count the distinct flightnumber and we have use group by statement to list the carriers with their distinct count;

**Task 7:**

Find out the average delay caused due to Weather for each year.

**(25 marks)**

#Write your command and comment for your command.

Answer: 43.743969697550504

SELECT AVG (weatherdelay) AS Avg\_Delay\_Due\_to\_Weather

FROM

DDA1610218.airline\_external

WHERE

weatherdelay <> 0

;

#So I have use avg function to get the average of weather delay, but since the column has 0 and NA’a values we have to get avg without considering those values. So we have specified where command to ignore 0 values and since SQL automatically ignores NA or null values we don’t need to specify anything else.

**Task 8:**

Find out number of flights cancelled every year where Cancellation code is “NAS”.

**(25 marks)**

#Write your command and comment for your command.

Answer: 0

select sum(cancelled) as Total\_No\_of\_Flights\_Cancelled\_With\_Cancellation\_Code\_NAS from DDA1610218.airline\_external where cancellationcode = "NAS";

#we have use sum to know the total number of cancelled flights with cancellation code NAS, but we see none.

**Task 9:**

Find out the month of year when average Taxi -in time is highest.

**(25 marks)**

#Write your command and comment for your command.

Answer: 7

SELECT

MONTH

,AVG (taxiin) AS MAX\_AVG\_TAXI\_IN\_TIME

FROM

DDA1610218.airline\_external

WHERE

taxiin <> 0

GROUP BY

MONTH

ORDER BY

MAX\_AVG\_TAXI\_IN\_TIME DESC LIMIT 1

;

#I have use the average function on taxiin and since nested max function was not working I have use order by function to list only the maximum avg time with respect to month.

**Task 10:**

Find out the total number of flights from each source to each destination.

**(25 marks)**

#Write your command and comment for your command.

Answer: select origin,dest,count(flightnum) as total\_no\_of\_flights from dda1610218.airline\_external group by origin,dest;

#since we need to find the number of flights from each source to each destination , we have taken origin and dest variable and group it with count of flight number.

**Task 11:**

Find out the source destination pair with highest number of flights scheduled for it.

**(25 marks)**

#Write your command and comment for your command.

Answer: LAX – SFO total flights scheduled 298915.

SELECT

origin

,dest

,COUNT (flightnum) AS total\_no\_of\_flights

FROM

dda1610218.airline\_external

GROUP BY

origin

,dest

ORDER BY

total\_no\_of\_flights DESC LIMIT 1

;

#Since we have already list out the source and destination pair in the previous task, we have just use the order by descending on total flight with limit one.

**Task 12:**

Find out number of flights diverted due to weather conditions with their source and destination airport.

**(25 marks)**

#Write your command and comment for your command.

Answer: select origin,dest,count(diverted) as Flight\_Diverted\_Due\_to\_Bad\_Weather from dda1610218.airline\_external where diverted = "1" and CancellationCode = "B" group by origin,dest;

#Since we have to find out the number of diverted flights with weather condition, we have use count on diverted and put a condition where diverted =1 and cancellation code =B, and use group by statement to list origin and destination airport.

**Task 13:**

Create a managed table called cancelled\_flights which will have year, day of the month, date, flight number, carrier code, cancellation code, source, destination.

(**Hint:** You need to first find the cancelled flights. You need to also read about creating table with dynamic schema**.**)

**(100 marks)**

#Write your command and comment for your command, specifically how this table is being created?

Answer:

CREATE

TABLE

DDA1610218.cancelled\_flights row format delimited fields terminated BY ',' stored AS RCFile AS SELECT

YEAR

,DayofMonth

,FlightNum

,UniqueCarrier

,CancellationCode

,Origin

,dest

FROM

DDA1610218.airline\_external

WHERE

Cancelled = "1"

;

#To create a managed table we have use the create table statement and specify the table name and format, since the file needs to be managed by hive we have use RCFile (Record Columnar File) format then we have use as select statement to write our query about the data. So the format and data are copied from the original table airline\_external.

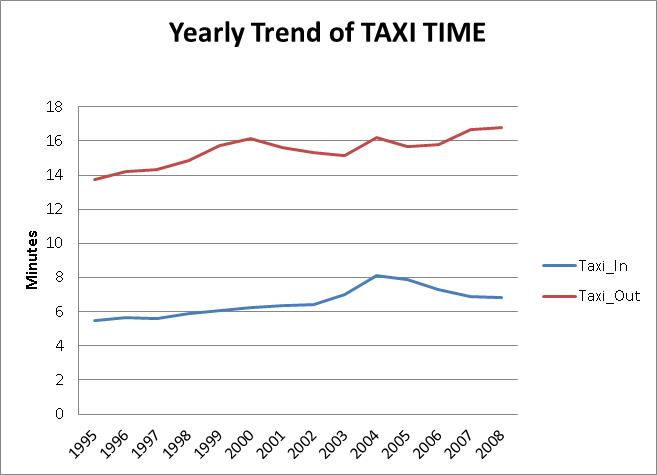
**Task 14:**

Comment on the trend that you find for taxi-in and taxi-out average time for the given data set. (**Hint:** Yearly average time to be considered)

**(100 marks)**

#Write your command and comment for your command.

Answer: We see that Taxi\_Out Time which is average time spent on the ground by these flights after leaving the gate increased from 13.8 minutes in 1995 to 16.8 minutes in 2008, an increase of 21 percent. And Taxi\_In Time average time spent getting to the gate after touching down has also increased over the years, from an average of 5.5 minutes in 1995 to 6.8 minutes in 2008—an increase of 25 percent.



SELECT

YEAR

,AVG (taxiin) AS Yearly\_Avg\_Taxi\_In

,AVG (taxiout) AS Yearly\_Avg\_Taxi\_Out

FROM

DDA1610218.airline\_external

WHERE

taxiin <> 0

AND taxiout <> 0

GROUP BY

YEAR

;

#So, I have use avg on both taxiin and taxiout and group by year to get the above yearly trend.

**Task 15:**

Comment on the time zones when maximum arrival delays and departure delays reported every year.

(**Hint**: Find out the CRSDepTime and CRSArrTime, divide it into primary time zones like early morning, morning, afternoon, evening, night. Then prepare your answer for this.)

**(125 marks)**

#Write your command and detailed comment for your command.

Answer: To calculate the time zone we have created a separate managed table with different time zone, and use that table to see the trend of delay in flights. So we see that in the 1997 there is a sudden drop in both departure and arrival delay and in the year 2007 we see that there is a sudden spike in both departure and arrival delay. We also see that afternoon and evening are the busiest time zone were high number of delay have been recorded. Below are the codes and some graphs.

#1. Create managed table: -

CREATE

TABLE

DDA1610218.Yearly\_Flights\_Delay row format delimited fields terminated BY ',' stored AS RCFile AS SELECT

YEAR

,CASE

WHEN crsdeptime BETWEEN 600 AND 859 THEN "Early\_Morning"

WHEN crsdeptime BETWEEN 900 AND 1159 THEN "Morning"

WHEN crsdeptime BETWEEN 1200 AND 1659 THEN "Afternoon"

WHEN crsdeptime BETWEEN 1700 AND 2159 THEN "Evening"

ELSE "Night"

END AS Departure\_Time\_Zone

,CASE

WHEN CRSArrTime BETWEEN 600 AND 859 THEN "Early\_Morning"

WHEN CRSArrTime BETWEEN 900 AND 1159 THEN "Morning"

WHEN CRSArrTime BETWEEN 1200 AND 1659 THEN "Afternoon"

WHEN CRSArrTime BETWEEN 1700 AND 2159 THEN "Evening"

ELSE "Night"

END AS Arrival\_Time\_Zone

,DepDelay

,ArrDelay

FROM

dda1610218.airline\_external

;

#2. Pulling query with sum of total arrival and departure delay:-

SELECT

YEAR

,Departure\_Time\_Zone

,Arrival\_Time\_Zone

,SUM (DepDelay) AS Total\_Delay\_In\_Departure\_In\_Min

,SUM (ArrDelay) AS Total\_Delay\_In\_Arrival\_In\_Min

FROM

DDA1610218.Yearly\_Flights\_Delay

GROUP BY

YEAR

,Departure\_Time\_Zone

,Arrival\_Time\_Zone

;

#using the max Departure\_Time\_Zone, we got Evening in the year 2007 where maximum delay were reported, and using max Arrival\_Time\_Zone we got same Evening in the year 2007.

