**ASSIGNMENT – 22**

**Dataset:** Flights\_Delay.csv

1. **Create external table in the Database Airline\_DelayDB**

create database if not exists Airlines\_DelayDB;

USE Airlines\_DelayDB;

drop table if exists Flights;

create table if not exists Flights (

ID INT,

YEAR INT,

MONTH STRING,

DAY STRING,

DAY\_OF\_WEEK STRING,

AIRLINE STRING,

FLIGHT\_NUMBER STRING,

TAIL\_NUMBER STRING,

ORIGIN\_AIRPORT STRING,

DESTINATION\_AIRPORT STRING,

SCHEDULED\_DEPARTURE STRING,

DEPARTURE\_TIME STRING,

DEPARTURE\_DELAY STRING,

TAXI\_OUT STRING,

WHEELS\_OFF STRING,

SCHEDULED\_TIME STRING,

ELAPSED\_TIME STRING,

AIR\_TIME STRING,

DISTANCE STRING,

WHEELS\_ON STRING,

TAXI\_IN STRING,

SCHEDULED\_ARRIVAL STRING,

ARRIVAL\_TIME STRING,

ARRIVAL\_DELAY INT,

DIVERTED INT,

CANCELLED INT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE

tblproperties ("skip.header.line.count"="1");

1. **Describe the table schema & show 10 rows of Dataset**

DESCRIBE Flights;

load data local inpath '/home/hadoop/Downloads/Flights\_Delay.csv' into table Flights;

SELECT \* FROM Flights LIMIT 10;

1. **Average arrival delay caused by airlines**

Select AIRLINE, INT(AVG(ARRIVAL\_DELAY)) AS T FROM Flights where ARRIVAL\_DELAY > 0 GROUP BY AIRLINE ORDER BY T DESC;

1. **Days of months with respected to arrival delays**

SELECT DAY, INT(AVG(ARRIVAL\_DELAY)) AS T FROM Flights WHERE ARRIVAL\_DELAY>0 GROUP BY DAY ORDER BY T DESC;

1. **Arrange weekdays with respect to the average arrival delays caused**

SELECT DAY\_OF\_WEEK, COUNT(FLIGHT\_NUMBER) AS T FROM Flights GROUP BY DAY\_OF\_WEEK ORDER BY T DESC;

1. **Arrange Days of month as per cancellations done in Descending**

SELECT DAY, COUNT(CANCELLED)AS T FROM Flights WHERE CANCELLED=1 GROUP BY DAY ORDER BY T DESC;

1. **Finding busiest airports with respect to day of week**

SELECT DESTINATION\_AIRPORT, COUNT(DESTINATION\_AIRPORT) AS X FROM Fly GROUP BY DESTINATION\_AIRPORT ORDER BY X DESC;

1. **Finding airlines that make the maximum number of cancellations**

SELECT AIRLINE, COUNT(CANCELLED) AS T FROM Fly WHERE CANCELLED=1 GROUP BY AIRLINE ORDER BY T DESC;

1. **Finding and ordering airlines in descending that make the most number of diversions**

SELECT AIRLINE, COUNT(DIVERTED) AS T FROM Flights WHERE DIVERTED=1 GROUP BY AIRLINE ORDER BY T DESC;

1. **Finding days of month that see the most number of diversion**

SELECT DAY,COUNT(DIVERTED)AS T FROM Fly WHERE DIVERTED=1 GROUP BY DAY ORDER BY T DESC;

1. **Calculating mean and standard deviation of departure delay for all flights in minutes**

SELECT AVG(DEPARTURE\_DELAY) as stddev\_pop(DEPARTURE\_DELAY) from Flights;

Responses: 11.329091145205275 39.620548511500886

1. **Calculating mean and standard deviation of arrival delay for all flights in minutes**

SELECT AVG(ARRIVAL\_DELAY) as stddev\_pop(ARRIVAL\_DELAY) from Flights;

1. **Create Bucketing Flights\_Bucket using MONTH into12 Buckets**

CREATE TABLE IF NOT exists Flights\_Bucket (

ID INT,

YEAR INT,

DAY STRING,

DAY\_OF\_WEEK STRING,

AIRLINE STRING,

FLIGHT\_NUMBER STRING,

TAIL\_NUMBER STRING,

ORIGIN\_AIRPORT STRING,

DESTINATION\_AIRPORT STRING,

SCHEDULED\_DEPARTURE STRING,

DEPARTURE\_TIME STRING,

DEPARTURE\_DELAY STRING,

TAXI\_OUT STRING,

WHEELS\_OFF STRING,

SCHEDULED\_TIME STRING,

ELAPSED\_TIME STRING,

AIR\_TIME STRING,

DISTANCE STRING,

WHEELS\_ON STRING,

TAXI\_IN STRING,

SCHEDULED\_ARRIVAL STRING,

ARRIVAL\_TIME STRING,

ARRIVAL\_DELAY INT,

DIVERTED INT,

CANCELLED INT)

clustered by (MONTH) into 3 buckets

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE

tblproperties ('serialization.null.format'='', "skip.header.line.count"="1");

FROM Flights

INSERT overwrite TABLE Flights\_Bucket

SELECT \*;

1. **Get count of data of each bucket**

SELECT count(\*) from Flights\_Bucket tablesample(bucket 1 out of 3 on MONTH);

SELECT count(\*) from Flights\_Bucket tablesample(bucket 2 out of 3 on MONTH);

SELECT count(\*) from Flights\_Bucket tablesample(bucket 3 out of 3 on MONTH);

1. **Finding all diverted Route from a source to destination Airport & which route is the most diverted**

SELECT ORIGIN\_AIRPORT, DESTINATION\_AIRPORT, COUNT(DIVERTED)AS T FROM Flights WHERE DIVERTED=1 GROUP BY ORIGIN\_AIRPORT, DESTINATION\_AIRPORT ORDER BY T DESC;

1. **Finding AIRLINES with its total flight count, total number of flights arrival delayed by more than 30 Minutes, % of such flights delayed by more than 30 minutes when it is not Weekends with minimum count of flights from Airlines by more than 10. Also Exclude some of Airlines 'AK', 'HI', 'PR', 'VI' and arrange output in descending order by % of such count of flights.**

Select AIRLINE, count (\*) as cnt,

sum (if (ARRIVAL\_DELAY >30, 1, 0)) as flights\_delayed,

round (sum (if (ARRIVAL\_DELAY >30, 1, 0))/count (\*),2) as rate

FROM Flights

WHERE DAY\_OF\_WEEK not in (6,7) and ORIGIN\_AIRPORT not in ('AK', 'HI', 'PR', 'VI') and DESTINATION\_AIRPORT not in ('AK', 'HI', 'PR', 'VI')

GROUP by AIRLINE HAVING cnt > 10 ORDER by rate DESC LIMIT 30;

1. **Finding AIRLINES with its total flight count with total number of flights departure delayed by less than 30 Minutes, % of such flights delayed by less than 30 minutes when it is Weekends with minimum count of flights from Airlines by more than 10. Also Exclude some of Airlines 'AK', 'HI', 'PR', 'VI' and arrange output in descending order by % of such count of flights.**

Select AIRLINE, count(\*) as cnt,

sum(if(DEPARTURE\_DELAY < 30, 1, 0)) as flights\_delayed,

round(sum(if(DEPARTURE\_DELAY < 30, 1, 0))/count(\*),2) as rate

FROM Flights

WHERE DAY\_OF\_WEEK in (6,7) and ORIGIN\_AIRPORT not in ('AK', 'HI', 'PR', 'VI') and DESTINATION\_AIRPORT not in ('AK', 'HI', 'PR', 'VI')

GROUP by AIRLINE HAVING cnt > 10 ORDER by rate DESC LIMIT 30 ;

1. **When is the best time of day/day of week/time of a year to fly to minimise delays.**

DO IT YOURSELF