**Case Study: Analysis of Airline Delay using Spark**

Airline delay is a critical issue affecting both airlines and passengers. In this assignment, you will use PySpark to analyze a dataset containing information about airline flights and predict flight delays.

**Technology: Spark, SQL**

**Dataset:** Flights\_Delay.csv

Dataset Description:

* ID: Rows ID
* YEAR: 2015
* MONTH: 1-12
* DAY: 1-31
* DAY\_OF\_WEEK: 1 (Monday) - 7 (Sunday)
* AIRLINE: Airline CODE
* FLIGHT\_NUMBER: Flight Number
* TAIL\_NUMBER: Flight’s tail number
* ORIGIN\_AIRPORT: Origin IATA airport code
* DESTINATION\_AIRPORT: Destination IATA airport code
* SCHEDULED\_DEPARTURE: Actual departure time (local, hhmm)
* DEPARTURE\_TIME: Scheduled departure time (local, hhmm)
* DEPARTURE\_DELAY: Departure delay, in minutes
* TAXI\_OUT: Taxi out time in minutes
* WHEELS\_OFF:
* SCHEDULED\_TIME: Scheduled arrival time (local, hhmm)
* ELAPSED\_TIME: in Minutes
* AIR\_TIME: in Minutes
* DISTANCE: in Miles
* WHEELS\_ON:
* TAXI\_IN: Taxi in time, in minutes
* SCHEDULED\_ARRIVAL: Scheduled arrival time (local, hhmm)
* ARRIVAL\_TIME: Actual arrival time (local, hhmm)
* ARRIVAL\_DELAY: Arrival delay, in minutes
* DIVERTED: 1 = yes, 0 = no
* CANCELLED: 1 = yes, 0 = no

Questions:

1. Create a new Spark Session with new SparkConfig
2. Create new instance of Spark SQL session and define new DataFrame using Flights\_Delay.csv dataset.
3. Create table Spark HIVE table flights\_table
4. Describe the table schema & show top 10 rows of Dataset
5. Apply Query performance optimization techniques like – creating Partitioning DataFrame by a specific column, parquet data, caching, predicate pushdown methods etc.

Write Spark SQL queries to show following analysis with Visualization on Databricks Community Edition.

1. Average arrival delay caused by airlines
2. Days of months with respected to average of arrival delays
3. Arrange weekdays with respect to the average arrival delays caused
4. Arrange Days of month as per cancellations done in Descending
5. Find Top 10 busiest airports with respect to day of week
6. Finding airlines that make the maximum number of cancellations
7. Find and order airlines in descending that make the most number of diversions
8. Finding days of month that see the most number of diversion
9. Calculating mean and standard deviation of departure delay for all flights in minutes
10. Calculating mean and standard deviation of arrival delay for all flights in minutes
11. **Finding all diverted Route from a source to destination Airport & which route is the most diverted**
12. 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.
13. 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.
14. **When is the best time of day/day of week/time of a year to fly with minimum delays?**
15. **Which airlines are best airline to travel considering number of cancellations, arrival, departure delays and all reasons affecting performance of airline industry.**