|  |
| --- |
| **Problem Statement: Analyse flight delay information-based input datasets**  **1. Datasets -**                  1. Int\_Flights\_Aircraft\_Weather.csv                  2. airports.json  **2. Data Cleansing** - Write a python function/functions for data cleansing process                  1. In airports.json file contains different type attribute key & values                  2. Make Sure if any integer attribute values are defined to None/NULL/Empty, change those to 0                  3. Make Sure if any string attribute values are defined to None/NULL/Empty, change those to BLANK                  4. Make Sure if any float attribute values are defined to None/NULL/Empty, change those to 0.0  **3. Data validations: -**   1. Validate all column level data had same datatype (model datatypes) before data ingestion into db tables. 2. Create table DDL statements                   1. Flight\_Weather                  2. airports datasets  4**. Data Load: -**   1. Load CSV & JSON files into database tables in multiple chunks (due to volume of csv records) – Write functions to load the data 2. Make sure error handling conditions are defined properly (try/catch & rollback/commit).   **5. Data Analysis** - SQL queries   1. Find the most frequent flight path (orign and desitnation combination) 2. Find Top 5 destinations airports with most delays 3. Find the Origin and Destination path with most delays 4. Find the Avg delay by country and carrier |

**Note :- Code from airport data and origin/dest are key columns to join.**