Data Engineer Coding Exercise

Context

You’re working as a Data Engineer for New York X Airlines (NYXA). The company has an internal system that manages all elements related to bookings, passengers' information, schedules, etc. The company is developing analytical models to improve flight punctuality, reduces costs, detect bad weather conditions on time, and optimize processes.

Challenge 1: NYC Flights ETL

There’s an automatic process that scrapes data from other airlines, standardizes them with the company's data, and saves the result as **CSV** files in a storage service. The company wants to make the data easily accessible to other teams.

The data is stored as the **CSV** files **‘nyc_airlines.csv**‘, **‘nyc_airports.csv**‘, **‘nyc_flights.csv**‘, **‘nyc_planes.csv**‘, and **‘nyc_weather.csv**‘, inside the directory ‘challenge1/dataset’.

1. You should create an ETL to load this data into the Data Warehouse (PostgreSQL in our example), so they can make queries and manipulate the data easily using SQL.

2. After completing the last topic, you should create an ETL for extracting and transforming the data from the Data Warehouse into the model described below and store it into a storage service.

The model requires a single CSV file with its column names at the first line and with the following fields:

- **airline**: Airline name, ex. United Air Lines Inc.
- **flight**: Composed by the concatenation of Airline code and flight number, ex. UA-1545.
- **flight_date**: Date of the flight following the format YYYY-MM-DD, ex. 2013-01-01.
- **tailnum**: Tail number of the plane, ex. N14228.
- **sched_dep_time**: Scheduled departure time following the format HH:MM, ex. 05:15.
- **actual_dep_time**: Actual departure time following the format HH:MM, ex. 05:17.
- **dep_delay**: Departure delay in minutes, ex. 2.
- **sched_arr_time**: Scheduled arrival time following the format HH:MM, ex. 08:19.
- **actual_arr_time**: Actual arrival time following the format HH:MM, ex. 08:30
- **arr_delay**: Arrival delay in minutes, ex. 11.
- **origin_faa**: Origin airport FAA code, ex. EWR.
- **origin_name**: Origin airport name, ex. Newark Liberty Intl.
- **origin_latitude**: Origin airport latitude, ex. 40.6925.
- **origin_longitude**: Origin airport longitude, ex. -74.168667.

- **origin_altitude**: Origin airport altitude, ex. 18.
- **origin_temp**: Temperature at origin airport, ex. 39.02.
- **origin_dewp**: Dewp at origin airport, ex. 28.04.
- **origin_humid**: Humidity at origin airport , ex. 64.43.
- **origin_precip**: Precipitation at origin airport, ex. 0.
- **origin_pressure**: Pressure at origin airport, ex. 1011.9.
- **origin_visib**: Visibility at origin airport, ex. 10.
- **dest_faa**: Destination airport FAA code, ex. IAH.
- **dest_name**: Destination airport name, ex. George Bush Intercontinental.
- **dest_latitude**: Destination airport latitude, ex. 29.984433.
- **dest_longitude**: Destination airport longitude, ex. 95.341442.
- **dest_altitude**: Destination airport altitude, ex. 97.
- **air_time**: Flight air time in minutes, ex. 227.
- **distance**: Distance between origin and destination airports, ex, 1400.
- **airplane_model**: Composed by the concatenation of the airplane manufacturer and the airplane model, ex. BOEING 737-824.
- **airplane_seats**: Number of airplane seats, ex. 149.