phData is a customer-focused consulting company, this means our employees must be able to deliver production-ready applications solving business needs for our customers. This interview project aims to replicate a client project work experience.

The following problem is designed to challenge candidates applying to either a Data Engineer or Solutions Architect role. You will be provided all the necessary credentials and access for completing the project, if you have any questions or uncertainties please reach out to the Recruiter.

Applicants should be prepared to present their solution as you would to a client. This should include a presentation to explain the problem & solution, a code review, and ultimately demo a running example of the code.

Overview

Northwoods Airlines is a client engaging you in a proof of value (POV) project using Snowflake and Databricks. They are working with phData to investigate the benefits of these cloud-native platforms. Their Data Owners have provided us datasets containing airports, airlines, and flights in a shared location and they have asked phData to load these datasets into the cloud environments (Databricks and Snowflake) so that they can develop reports and gather insights against their competition. As a Data Engineer or Solutions Architect, you have been tasked with creating the necessary tables in Snowflake using tooling provided by either Databricks or Snowflake, load the data from the shared location, and finally, be able to create views or run reports to gather insights based on the datasets. It is recommended to materialize the reports into a table within Snowflake. Any ETL needs to happen within Spark using the Snowflake connector.

Non-Requirements

- Completing in a specific amount of time. Life is busy and chaotic. We understand you will not be able to work full time on the project.
- Machine learning. Machine learning **could** be used to solve this problem, but it is **not** required.
- There is no need to run this code at scale. Everything should be done either a single node pseudo cluster or a small cluster.
- An exact end result. Two candidates given this assignment will find different solutions. Feel free to choose your own adventure as long as the base requirements are met.

Requirements

Use both of the cloud environments to build the solutions. You may create and use your own Snowflake and Databricks environment.

Snowflake

***Note**: Snowflake Free edition comes with 30days and $400 credit. Register for free and build this solution. If you have difficulty creating and registering an environment, contact the phData Recruiter.*

- Create & use the INTERVIEW_WH warehouse
- Use your USER_<name> database
- Create tables based on the defined datasets above
- Load data from the external stage into corresponding tables

Databricks

***Note**: Databricks community edition is free with small cluster size. Register for free and build this solution. If you have difficulty creating and registering an environment, contact the phData Recruiter. You may use the community edition (https://community.cloud.databricks.com)*

- Use the community cluster
- Create a Databricks application using a notebook using Scala or Python, or compiled jar
- The application should read in the provided CSV as data frames
- All data needs to be persisted within Snowflake.

Reports

Once the data has been loaded into the respective platform(s), Clients will request various insights or KPI reports derived from the provided data. The Executive team at Northwoods Airline's have requested reports or dashboards to determine how their competition is performing.

- Total number of flights by airline and airport on a monthly basis
- On time percentage of each airline for the year 2015
- Airlines with the largest number of delays
- Cancellation reasons by airport
- Delay reasons by airport
- Airline with the most unique routes

These reports should be in the form of views created in your user database in Snowflake and visualized using Databricks notebooks (preferred) with the graphing/display capabilities to show the details.

Datasets (link)

Airlines

Data Entity	Airlines
File Name	airlines.csv
Description	Airline codes and names

Data Structure

Column Name	Data Type	Description
IATA_CODE	String	Airline Identifier
AIRLINE	String	Airline name

Airports

Data Entity	Airports
Location	airports.csv
Description	Airport codes, names, and locations

Data Structure

Column Name	Data Type	Description
IATA_CODE	String	Airline identifier
AIRPORT	String	Airport name
CITY	String	City where the airport is located
STATE	String	State where the airport is located
COUNTRY	String	The country where the airport is located
LATITUDE	Number	Latitude of the airport
LONGITUDE	Number	Longitude of the airport

Flights

Data Entity	Flights
Location	flights/partition-x.csv
Description	Flight Records

Data Structure

Column Name	**Data Type**	**Description**
YEAR	Number	Year of the flight
MONTH	Number	The month of the flight
DAY	Number	The day of the flight
DAY_OF_WEEK	Number	Day of the week of the flight
AIRLINE	String	Airline identifier
FLIGHT_NUMBER	String	Flight identifier
TAIL_NUMBER	String	Aircraft identifier
ORIGIN_AIRPORT	String	Starting airport
DESTINATION_AIRPORT	String	Destination airport
SCHEDULED_DEPARTURE	String	Planned departure time
DEPARTURE_TIME	String	Wheels up time
DEPARTURE_DELAY	Number	Total delay on departure
TAXI_OUT	Number	The time duration elapsed between departure from the origin airport gate and wheels off
WHEELS_OFF	String	The time point that the aircraft's wheels leave the ground
SCHEDULED_TIME	Number	Planned time amount needed for the flight trip
ELAPSED_TIME	Number	Total trip time
AIR_TIME	Number	The time duration between wheels_off and wheels_on time
DISTANCE	Number	The distance between two airports

WHEELS_ON	Number	The time point that the aircraft's wheels touch on the ground
TAXI_IN	Number	The time duration elapsed between wheels-on and gate arrival at the destination airport
SCHEDULED_ARRIVAL	Number	Planned arrival time
ARRIVAL_TIME	String	Time of arrival
ARRIVAL_DELAY	String	Arrival time - Scheduled Arrival
DIVERTED	Number	Aircraft was diverted
CANCELLED	Number	Aircraft was canceled
CANCELLATION_REASON	String	Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D - Security
AIR_SYSTEM_DELAY	Number	Delay caused by the air system
SECURITY_DELAY	Number	Delay caused by security
AIRLINE_DELAY	Number	Delay caused by the airline
LATE_AIRCRAFT_DELAY	Number	Delay caused by the aircraft
WEATHER_DELAY	Number	Delay caused by weather

Platform Documentation

Both Databricks and Snowflake have excellent documentation that will be useful when completing this project

- Databricks Documentation
- Snowflake Documentation

Feedback and Assistance

When you have questions or require assistance throughout the project please reach out to the phData Recruiter who will point you in the right direction to help you through any possible setup or configuration questions. We do not want you to spend a lot of time working through connection or configuration issues, instead, you should be focused on delivering the business value for Northwoods Airlines. Overall, this project should take ~10 hours to complete. Please reach out if you have any questions or challenges with the community versions of these platforms.