Overview

Northwoods Airlines is a client engaging you in 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.

Important Note

● 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.

● There is no need to run this code at scale. Design your approach considering the community

edition of databricks and snowflake data warehouse platform

● 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.

● Time management (We cannot stress these two enough.)

○ Build the simplest possible solution first, utilizing your favorite programming language

(Spark/Scala or PySpark)

○ Don’t get stuck on one aspect of the project.

● Prepare a presentation to go along with your code demo or document your approach within the

databricks notebook using the markdown feature. Act as a consultant selling your understanding

of the problem and your solution.

Non Functional Requirements

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

Snowflake and Databricks environment.

● Load the given data set to the databricks file system before loading it to the snowflake.

● Keep the snowflake database DDLs for tables & views as SQL file in a private git repository

● Make sure you push the code on regular basis (daily or within a day when you finish one logical

unit of work) to avoid the loss of work.

Databricks Instance

***Note*** *: Databricks community edition is free with one node two cores cluster. 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 ) using your personal*

*email address*

● Create and use the community edition instance and databricks cluster

● Create a databricks application using a notebook (Scala or Python), or compiled jar

● Load all the data set given for this project challenge in the databricks file system

● The application should read in the provided CSV as data frames

● All data needs also to be persisted within Snowflake.

Snowflake Instance

***Note*** *: Snowflake Free edition comes with 30days and $400 credit. Register for a free edition 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

*○ Hint: you can use a different schema like raw & curated as per the nature of the data.*

● Load data from the external stage into corresponding tables

Github Instance (Version Control System)

Github also provides a free repo for personal use. Register for the free edition and have your source

code available there.

***Note*** : *Make sure you create a private repository and use the repo to manage your code.*

Reports (Functional Requirements)

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 has 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

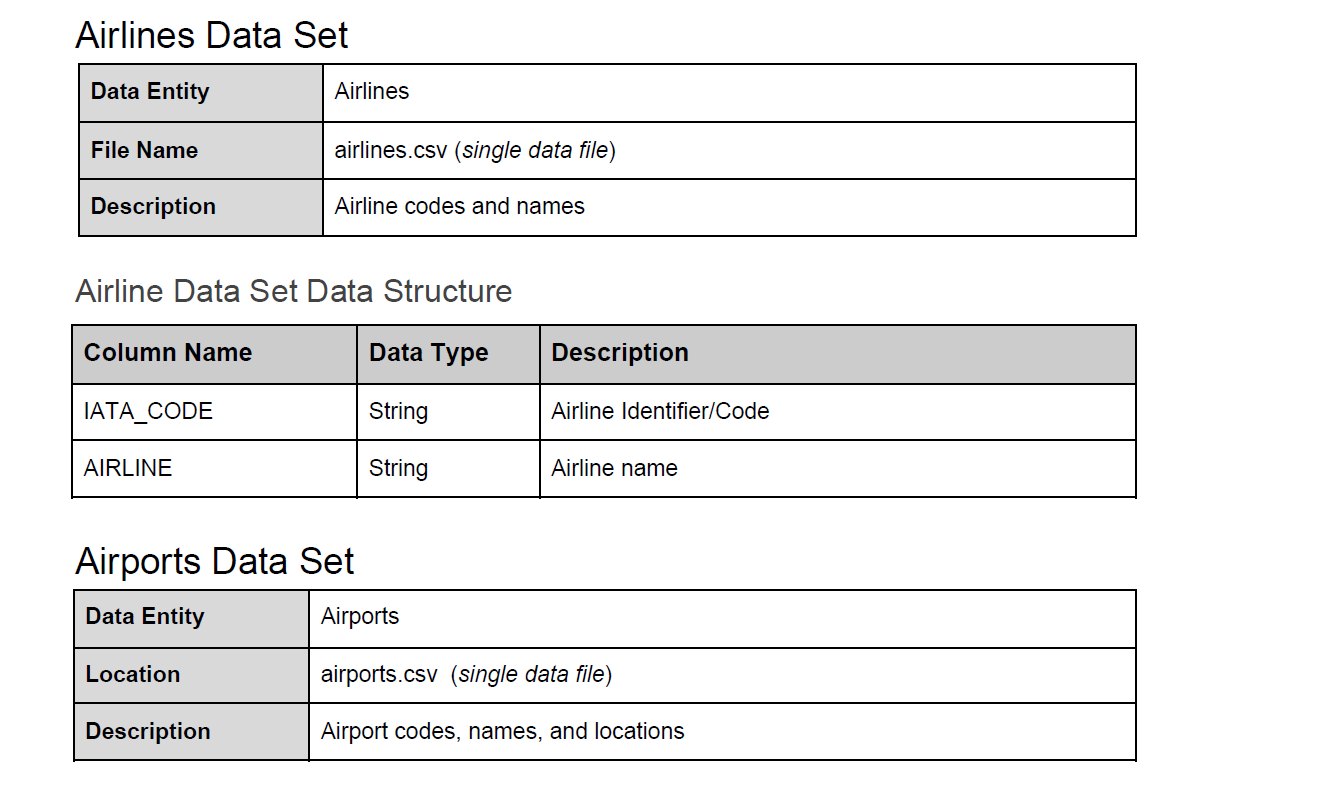
● Airline with the most unique routes

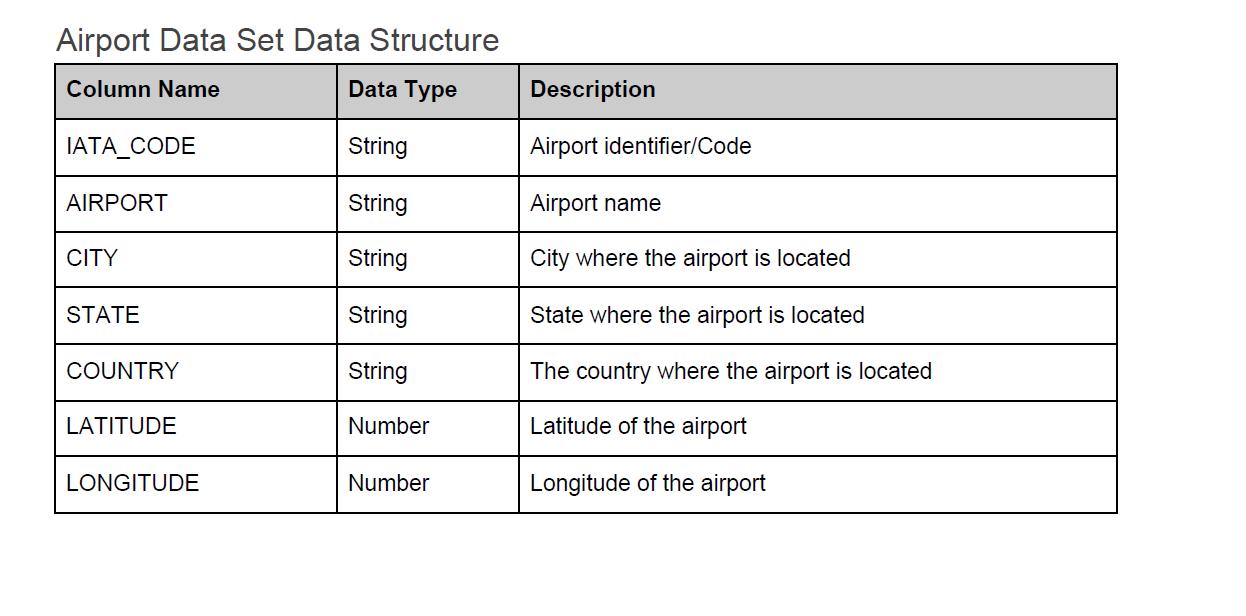
These reports should be in the form of views created in your user database in Snowflake and visualized

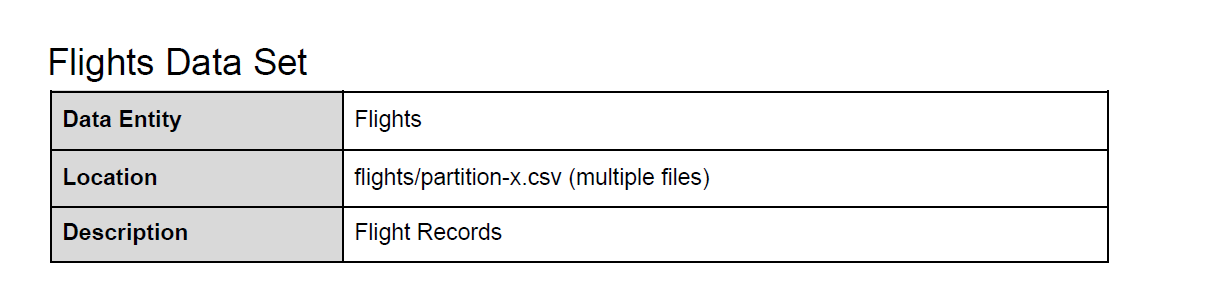
using Databricks notebooks (preferred) with the graphing/display capabilities to show the details.

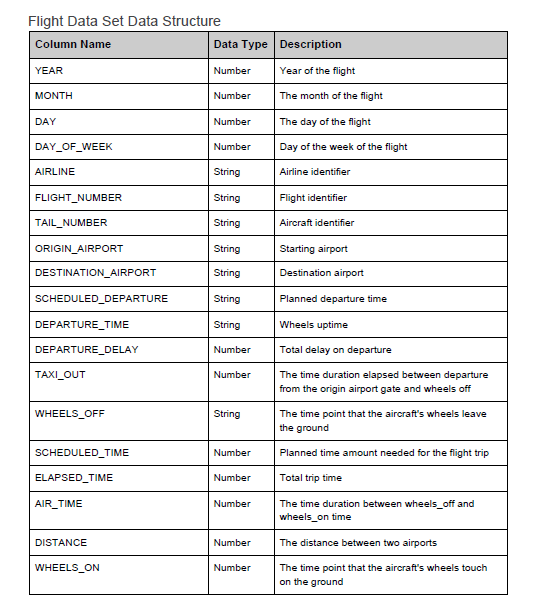
Datasets & Entity Description

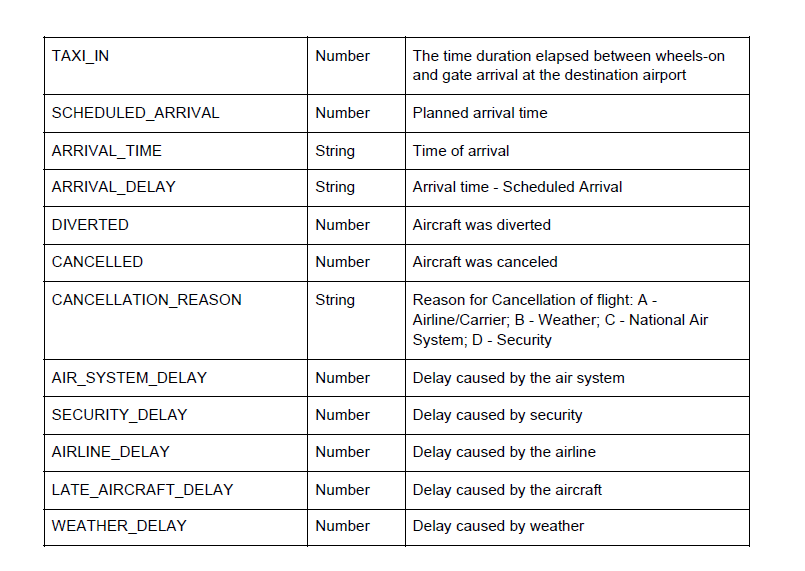
[*https://drive.google.com/drive/folders/18Mkt2Ku3gIxenT-zjYi68kcufpcvNwbv*](https://drive.google.com/drive/folders/18Mkt2Ku3gIxenT-zjYi68kcufpcvNwbv)











Platform Documentation

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

project

● Databricks Documentation

● Snowflake Documentation