Flights Case Study

Scenario

A client has provided you with a flat file export containing data about domestic flights in the United States. The analytics consultant that you are working with needs the file loaded into a Snowflake database and appropriately modeled for use with business intelligence tools. In this case, that tool is Tableau.

Goal

The goal of this case is to assess your ability to:

1. Work with unfamiliar data and quickly gain familiarity with that data.
2. Load the data from a file-based source into a target database. You may use whatever combination of tools you wish, or the Snowflake UI directly to complete the case study, but the final results **must** be in the candidate_#### schema in the recruitment_db database located in the Snowflake instance:
   InterWorks Data Team Recruitment - Snowflake DB
3. Transform the data into a normalized, dimensional data model with appropriate data types
4. Think critically about preparing data for business users and analytics consultants
5. Present issues with data quality that do not have obvious solutions and solve issues that do have more straightforward solutions

Data

The data is provided as a compressed flat file called flights.gz. The file is encoded in UTF-8 and is delimited by the pipe character. Column headers are provided on the first line. You can access the file by using the named stage that has been setup in Snowflake for you and located here RECRUITMENT_DB.PUBLIC.S3_FOLDER.

You can see the file and interact stage by using the following command:

```
LIST @RECRUITMENT_DB.PUBLIC.S3_FOLDER;
```

To assist in automated grading:

- **DO NOT CHANGE THE NAMES OF ANY COLUMNS IN THE DATA**
- **ENSURE YOU USE QUOTES WHEN STIPULATING TABLE, VIEW AND FIELD NAMES TO ENSURE THEY ARE CREATED IN THE REQUESTED CASE.**
- **CREATE TABLES AND VIEWS ACCORDING TO THE NAMING CONVENTIONS:**
  - **DIM_* for dimension tables (e.g. DIM_DATE or DIM_AIRPORT)**
  - **FACT_FLIGHTS**
  - **VW_FLIGHTS**

Failure to follow these instructions will cause issues with grading and may cause your test to be rejected.

For example, two columns in the source data are named TRANSACTIONID and CANCELLED. They should be named TRANSACTIONID and CANCELLED in your view, VW_FLIGHTS. In the example below, we ask you to create a column DISTANCEGROUP and it should be named DISTANCEGROUP in all tables and all views. There should be no underscores or spaces added.

Case Requirements

1. Load the provided data in the file using the provided credentials & named stage to your candidate schema. *Hint – use a COPY INTO command in Snowflake with the stage name mentioned above to load this file.*
2. Create and load one fact table named FACT_FLIGHTS to contain data about the flights
3. Create and load appropriate dimension table(s) named DIM_*.
4. Create a view named VW_FLIGHTS that joins your fact table to your dimension tables and returns columns useful for analysis. Please see the "View" section for more information.
5. Prepare a 10 to 15-minute presentation to show your work, discuss your approach and any data quality issues that were encountered and your resolution to these issues. More information on the presentation content is detailed below.

Additional Details and Instructions

Fact Table

1. Create an additional column named DISTANCEGROUP that bins the distance values into groups in 100-mile increments. Example: 94 miles is 0-100 miles. 274 miles is 201-300 miles. Please make special note of the bins: 0-100, 20**1**-300, 30**1**-400, etc. Please make sure that the format of the bins is "201-300 miles". Not "201-300".
2. Create an additional column named DEPDELAYGT15 that indicates (0/1) if the departure delay in minutes (DEPDELAY) is greater than 15.
3. Create an additional column named NEXTDAYARR that indicates (0/1) if the flight arrival time (ARRTIME) is the next day after the departure time (DEPTIME).
4. Choose appropriate data types and perform conversions to load the data from the source into these types.
5. Fix obviously bad data when encountered, if possible. Note these instances.

Dimension Table(s)

1. Create at least one dimension table and load it from the source data.
2. Use your judgment about what columns from the source data should end up in the dimension tables. Be prepared to explain your decisions.
3. Clean up the AIRLINENAME column by removing the airline code from it.
4. Clean up the ORIGAIRPORTNAME and DESTAIRPORTNAME columns by removing the concatenated city and state.
5. Fix obviously bad data when encountered, if possible. Note these instances.

View

Your final view (VW_FLIGHTS) should contain all columns you deem useful for analysis.

Please also make sure your view (VW_FLIGHTS) includes at least each of these columns and remember, as stated above, to not change any of the column names:

- TRANSACTIONID
- DISTANCEGROUP
- DEPDELAYGT15
- NEXTDAYARR
- AIRLINENAME
- ORIGAIRPORTNAME
- DESTAIRPORTNAME

Additionally, if you find and repair any other data issues, please make sure to include those columns in the final view.

Presentation

If you successfully complete the case study, you will be asked to present your findings. This presentation should not be a code review. If you make it to the presentation round it means we've already reviewed your code. InterWorks is a consultancy. We are interested to see how you would present your results and the value you have provided to a non-technical or semi-technical audience.

You will be judged on your ability to convey technical information in an easy-to-understand format.

Good Luck!