Part 1 – Practical Assignment

Context

The Analytics team have a requirement to generate some insights and business reporting on bird strikes occurring across the U.S. Currently, a daily csv file from is received from a supplier which contains the new and updated records for the previous day. A consolidated weekly file is also received which contains the full data set. We would like to make this data available to the Analytics team to allow them to deliver on their known business reporting requirements as well as allow them to run analytic models to discover further (currently unknown) insights.

The Analytics Team would like to persist all history as they are not yet sure how they may use the data.

Details about the files

- Bird Strikes - Base.csv: file contains the initial data set and should be loaded first. Assume that this reflects the initial state of the data.
- Bird Strikes - Day 1.csv, Bird Strikes - Day 2.csv, Bird Strikes - Day 3.csv: contain incremental inserts and updates for 3 days and should be loaded sequentially.
- Bird Strikes - Final.csv: is the final snapshot received at the end of the week containing a full snapshot of the data. This can include additional inserts and updates as well as any deletes.

Your assignment is to design a process to ingest and store the data, implement data quality checks to ensure the data is valid, apply business rules to enrich the data, and then perform some analysis over the data. This should include:

1. An overall architecture to load and store the data as well as make the data available for reporting and analytics.
2. Design a data model optimised for reporting in order to present this data through a BI tool.
3. Define an approach to address potential data quality issues in the source data. How would you identify these and provide updates on the state of the data? An example is the Values in the Feet above ground field not falling within the corresponding Altitude Bin field.
    a. Review the data and propose 1 of your own Data Quality checks. Why do you think it's an important check?
4. How would you implement business logic, derived fields to make using the data for reporting easier. For example, how could you model the data in the "Conditions: Precipitation" field to make it easier to use for analysis and reporting?

Data Augmentation

1. Assuming you could make suggestions to the supplier to change how they provide the data:
    a. What changes would you propose and why?
    b. How would this influence your design?
2. What additional third-party data sets would you attempt to source to augment the bird strikes data to provide better insights into this data?

Data Analysis

1. Use the data to answer the following questions:
    a. Which departure airport has the highest number of bird strikes?
    b. What is the overall trend in reported bird strikes over time?
        i. Why do you think the trend exists?
2. Propose 2 of your own questions – and answers. Why would these be important for the business.

Communication

1. Prepare a demo and presentation that to us through at your next interview. This should ideally include a demo of working code. Please bring in your own laptop to perform the demo and presentation.
2. If you have any questions please email me (jimmov@bizcover.com.au) as necessary. While you will need to make some assumptions because this is an assignment, part of your evaluation will include reaching out to ask some questions to validate some of your assumptions. I will be available to answer questions during working hours.

Good luck and enjoy!

Part 2 - Additional technical questions (bonus questions)

Completing Part 1 of the assignment is #1 priority. If you have time and are interested, you can take a look at the following questions. Acknowledging you may be unfamiliar with Snowflake and DBT, Googling is OK, but see if you can:

- Relate your answers to the scenario the Part 1 of the assignment.
- Provide an opinion on why these features are particularly useful.
- Identify any limitations which you might need to work around.

Questions:

1. Snowflake:
    a. What is Snowpipe?
    b. What is the variant data type?
    c. What is a Snowflake Dynamic Data Masking?
    d. What is UNLOAD in Snowflake?

2. DBT:
    a. What is a DBT source?
    b. What is a DBT Snapshot? What are some of the drawbacks of using snapshots?
    c. Why doesn’t DBT support ‘identity’ columns for things like surrogate keys