**ETL Project**

By: Matt Egan & Mark Munanura

**Preprocessing:**

For this project we set to aggregate data for future analysis on the potential impact weather has on an airline's stock performance. We accessed Delta’s historical stock performance using Python’s yfinance module which we converted into a CSV before transforming the data. We also used Kaggle to grab a data set detailing weather events across the United States from 2016-2020. This data was also formatted as a CSV.

**Sources:**

- Python’s yfiance library

- https://www.kaggle.com/sobhanmoosavi/us-weather-events

**Transformation:**

To prepare both datasets to get for further analysis we performed the following:

- We split the two DateTime columns from the original weather data CSV into separate date/ time columns to facilitate easier joins once the data was loaded to a database.

- Changed column names on both datasets to comply with SQL best practices.

- Excluded unnecessary data columns from the final data frames.

**Loading:**

To store our data we chose to use a PostgreSQL server. Given that our data was tabular, it was easy to setup up the tables before we loaded the data into the database. PostgresSQL will provide our team a much easier way to join data based on date/time making it a superior choice to MongoDB for our use case.

To load the data into our database, we leverage SQLAlchemy to make a connection with our Jupyter notebook and PostgreSQL server.