10Alytics Data Engineer Capstone Project

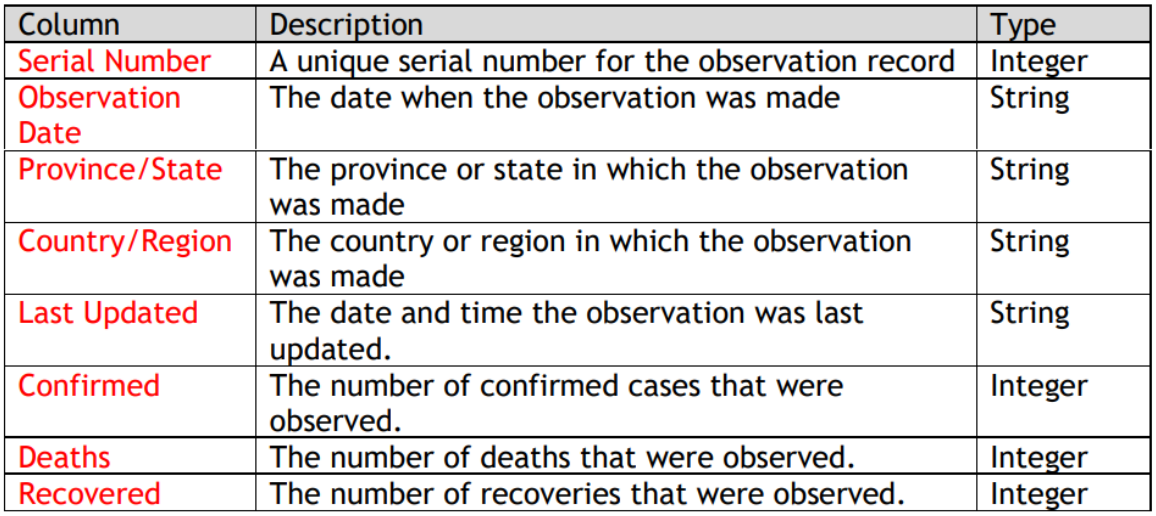
Part 1 (SQL)

Introduction:

The Covid 19 pandemic has wreaked havoc and led to the dramatic loss of lives and livelihoods. Its impact continues to affect the way in we live and interact. In this project, you will analyze sample data related to COVID-19 cases as recorded from January 2019 to December, 2020.

The data is provided as a CSV file.

Data Summary:



Column	Description	Type
Serial Number	A unique serial number for the observation record	Integer
Observation Date	The date when the observation was made	String
Province/State	The province or state in which the observation was made	String
Country/Region	The country or region in which the observation was made	String
Last Updated	The date and time the observation was last updated.	String
Confirmed	The number of confirmed cases that were observed.	Integer
Deaths	The number of deaths that were observed.	Integer
Recovered	The number of recoveries that were observed.	Integer

Instructions:

NB: The you're expected to use PostgreSQL as your database tool.

Follow this link to download, install and setup a Postgresql database If you don't currently have Postgresql installed: https://www.postgresqltutorial.com/install-postgresql/

1. Create a database and a table called **covid_19_data** to hold the data in postgresql.

2. You’re to define the data type of ObservationDate as DATE type in the database instead of String as shown on the data summary image above.
3. Use a python script to download the Covid_19_data.csv file and load into a Postgresql database.
4. Use the Posgresql PG4 Admin for writing and running your SQL Queries.
5. You will submit only one SQL file with all the queries.

Your task:

1. Write a simple Python script to download the Covid_19_data.csv file using this link .
2. The script should also be used to load the data into the table you created in a PostgreSQL Database.
3. Write suitable sql queries to analyze and generate insight from the data.

Below are the questions you are to write SQL queries to find answers to:

1. Retrieve the total confirmed, death, and recovered cases.
2. Retrieve the total confirmed, deaths and recovered cases for the first quarter of each year of observation.
3. Retrieve a summary of all the records. This should include the following information for each country:
    - The total number of confirmed cases
    - The total number of deaths
    - The total number of recoveries
4. Retrieve the percentage increase in the number of death cases from 2019 to 2020.
5. Retrieve information for the top 5 countries with the highest confirmed cases.
6. Compute the total number of drop (decrease) or increase in the confirmed cases from month to month in the 2 years of observation.

Submission:

You’re required to put all the queries in a single .sql file. Then create a repository on Github to host your project. Your repo should contain:

1. A single SQL file containing all the queries.
2. A folder in the repository called “outputs”. This folder should contain screenshots of the outputs from running the queries.
3. A single python file for downloading and loading the data into PostgreSQL.

Finally, you will submit a link to the Github repository.

Part 2 (Python & Cloud)

10Alytics Job Board Service:

10Alytics is a Leading Data and Strategy Firm that provides technology training service to help its clients acquire industry relevant tech skills to propel their tech career.

As part of expansion, we want to provide job placement services to our trainees. We have therefore hired you as our Data Engineer to help us build a data infrastructure solution that provides daily job update on our job board.

Project requirement:

You're expected to build a data pipeline to extract daily job posting data from a public API available here: https://rapidapi.com/letscrape-6bRBa3QguO5/api/jsearch/. You need to first create an account on Rapid API website in other to use this API. Details on how to query the API is available on the page.

You're expected to follow the following steps:

1. Design a pipeline architecture diagram to show the flow of data from the source to the destination.
2. Your pipeline should extract raw data of the jobs posted for the current day and specifically **Data Engineer** and **Data Analyst** jobs posted in either UK, Cannada or the US.
3. Your pipeline should stage the raw data from step 1 above in an Amazon S3 bucket called "raw_jobs_data". The raw data should be saved in a JSON format.
4. Your pipeline should transform the raw data and load into another S3 bucket called "transformed_jobs_data". The transformed data should be saved in a CSV format. Your transformed data should contain only these columns: **employer_website, job_id, job_employment_type, job_title, job_apply_link, job_description, job_city, job_country, job_posted_at_timestamp, employer_company_type.**
5. Your pipeline should finally pull data from the transformed_jobs_data bucket and load into an Amazon Redshift data warehouse.
6. Bonus. Schedule your pipeline to run once in a day (choose anytime convenient for you) using Apache Airflow.
7. Your code should be hosted into a Github repository. Note that your commit history will be used to determine how you went about solving the problem. So, make sure you commit your code intermittently to show your working.
8. Your repository should contain a README file which has detailed information about the project and how to run the code.
9. Your project should contain (1) etl.py file (which does the extraction, transformation and loading (2) A util.py file (which contains utility functions

e.g database connection) (3) A main.py file (where the entire program can be run from) and any other python file you dim fit for the purpose of organization.

Submission:

You're expected to submit a single link to the Github repository hosting the project.

Part 3 (Python & SQL):

The World Port Index contains the location and physical characteristics of, and the facilities and services offered by major ports and terminals world-wide. You're hired as a Data Engineer by **GoFrieghts**, a leading logistics company to migrate the World Port Index data from an old Access database to its modern relational database management system (PostgreSQL) and create data marts from analysis of the data.

Your task:

1. Build and Extract Load (EL) pipeline in Python to migrate the World Port Index data from Access database available here: on this link to PostgreSQL.

For each of the following questions, create a separate Python script that creates a table in your PostgreSQL database.

Questions:

1. What are the 5 nearest ports to Singapore's JURONG ISLAND port? (country = 'SG', port_name = 'JURONG ISLAND').Your answer should include the columns port_name and distance_in_meters only.

2. Which country has the largest number of ports with a cargo_wharf? Your answer should include the columns country and port_count only.

3. You receive a distress call from the middle of the North Atlantic Ocean. The person on the line gave you a coordinates of lat: 32.610982, long: -38.706256 and asked for the nearest port with provisions, water, fuel_oil and diesel. Your answer should include the columns country, port_name, port_latitude and port_longitude only.

Submission:

You're expected to submit a single link to the Github repository hosting the project.

NB: You're expected to submit 3 links to the different project repositories. Each repository should contain a README file that explain your code implementation and how to run the code.

Email to submit project to will be shared later on the group chat.