PRD: Capstone

Product Engineering Project 3

By Justin Quinn, Finn Mikkola, Alexander Daughters

Change History

- Doc created Feb 19, 2024
- Finished project Feb 22, 2024

Overview

- Our team plans to architect and implement a data pipeline reflective of industry practices using semi-structured data.

Objectives

- extract data from API into S3 Bucket
- Facilitate AWS glue to structure a ETL pipeline
- Clean and transform the data using pandas or some equivalent service
- Load clean data into AWS redshift for data warehousing
- Use Redshift Analytic tools to extract actionable insights from the data

Checkpoints

- Use cloud storage services (e.g., AWS S3) to host the data lake, with automated or semi-automated data ingestion methods.
- Leverage cloud-based ETL services (e.g., AWS Glue) to automate the transformation process.
- Utilize a cloud-based data warehouse (e.g., Amazon Redshift) for data storage, benefiting from the scalability and managed services.
- Conduct analytical queries on the structured data to derive insights. Include analysis of both the dataset itself (via descriptive statistics) and correlations among the data (via inferential statistics).

Deliverables:

- **Data Extraction:**
    - Documentation on data source, extraction logic, and justification for chosen methods.
        - Source: API
        - Logic: timing of extraction/api call/ file update
    - Extraction code and a data sample
        - Screenshot of extraction code
        - Screenshot of data entries
- **Data Transformation:**
    - A report detailing transformation rules, challenges, and solutions.

        - Amazon glue: deleting doubles
    - Transformation code along with data samples pre and post-transformation.
        - Screenshot of null entries
        - Screenshot after dropping county column
        - ?Fetching only county column data? (special cases)
- **Data Loading:**
    - A description of the data warehouse architecture, schema design, and loading strategy.
        - Draw.io diagram
    - Code or configurations for the loading process, along with proof of successful execution.
        - Screenshot of data in Redshift
- **Data Analysis:**
    - An analysis report with insights, including the queries or scripts used, visualizations, and interpretations of the data.
        - the dataset itself
            - queries or scripts used
            - Visualizations
            - Interpretations
        - correlations among the data
            - queries or scripts used
            - Visualizations
            - Interpretations

Timeline and Release Planning:

- Release Date 2/22/24

Scenario:

- Global Corp Ltd. employs a diverse workforce in IT services around the world. They are starting a new initiative to help promote multicultural awareness among its staff. S3 Musketeer Solutions has been tasked with creating a cloud based service that will identify local holidays from around the world using the Public Holiday API.

User Stories:

- As a Global Corp Ltd. employee, I want to be able to easily identify holidays from around the world.
- As a Human Resources manager at Global Corp Ltd., I want the cloud-based service to be user-friendly and intuitive so that all employees, regardless of their technical expertise, can easily use it to identify local holidays.

- As a software developer at S3 Musketeer Solutions, I want to integrate the Public Holiday API into the cloud-based service effectively so that it accurately retrieves and displays local holidays from various countries.
- As a team leader at Global Corp Ltd., I want the cloud-based service to have customizable settings for different regions or departments within the company so that employees can filter and view relevant local holidays based on their specific needs.

Open Issues:

- Connecting to Redshift from local machine