Diagram

Description automatically generated

Figure 1. Data Architecture diagram

Figure 1 illustrates the process of extracting, transforming, and loading data. This diagram depicts four primary data sources. These data sources originate from cocktaildb.com, which offers five distinct datasets; we incorporated four of these datasets into our project. The initial dataset, Category, is presented in CSV format and comprises two attributes: strCategory and strDescription.

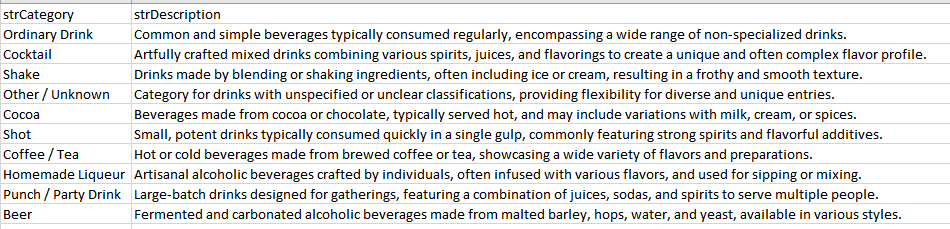


Figure 2. Category data extraction

The second dataset, comprised of ingredients, is structured as a CSV file and includes two attributes: strIngredient1 and strCountryofOrigin

Table

Description automatically generated with medium confidence

Figure 3. Ingredient Data extraction

he third and fourth datasets are formatted as JSON files. The third dataset, referred to as the glass table, comprises one attribute as it can be seen at figure 4.

Graphical user interface, text, application, email

Description automatically generated

Figure 4. Glass JSON file

The fourth dataset, known as the drink table, contains 51 attributes and is characterized by semi-structured dataGraphical user interface, text, application, email

Description automatically generated

Figure 5. Drink JSON file

The extraction of the Drink JSON file was accomplished by grouping entries based on the first letter of the drink name. Using Python code, this extraction process was automated through a loop to systematically handle each complete first letter of the drink name.

Text

Description automatically generated

Figure 6. Drink JSON extraction

After extracting these datasets, they were converted into data frames for review. The Category, Ingredients, and Glass datasets contain structured data and can be loaded in batches. However, the Drink dataset contains semi-structured data, with some attributes containing images or file texts.

No transformations were conducted on the structured data. However, since the Drink dataset has 51 attributes, many of which are unnecessary for the project's goals, dropping these attributes will expedite the loading process. Additionally, a simple transformation was applied to the Drink dataset by adding an additional column to identify whether it's an adult drink or not.

Text

Description automatically generated

Figure 7. Drink Data clean-up

Once the clean-up was completed, the data were loaded into our Azure Test Database. Before loading the data, a series of syntaxes to create tables were executed.

Text

Description automatically generated

Figure 8. Table creation

Upon the tables’ creation, a reverse engineering process was conducted in SQL Workbench to generate the final data schema, as illustrated in Figure 9. The schema depicts one-to-many relationships between category and drink, one-to-many relationships between ingredient and drink, and one-to-many relationships between glass and drink. Each of these relationships with the Drink table is connected to the primary keys of Category (strCategory), Glass (strGlass), and Ingredient (strIngredient1).

Diagram

Description automatically generated

Figure 9. Final Database schema