**GitHub: https://github.com/shailja-somani-0/ADS-507-Team-4**

**System Architecture San Diego, California Weather Data and Alerts Pipeline**

Shailja Somani, Tara Dehdari, and John Vincent Deniega

Shiley-Marcos School of Engineering, University of San Diego

**\*\*\*REMOVE:** One **PDF** document outlining your system architecture. This should describe the source data, the specifics of your pipeline ETL or ELT, and the output. It will also include “next steps” that outline any shortcomings of your current system and what could be done to improve it. **Provide the link to your GitHub repository at the top of the Design Document.\*\*\***

**Introduction**

Meteorology and weather – more broadly – can be observed in real-time or over specific periods of interest. Thus, this domain allows for data scientists to utilize either streaming data recorded via remote terminal units such as thermometers and barometers or point observations at known intervals such as hourly or daily weather data. For the purposes of utilizing well-established source reporting via the National Weather Service [NWS], the University of California San Diego [UCSD], and OpenWeather, a casual user can utilize the services afforded by these organizations to retrieve basic weather information, while a data scientist may utilize these data sources to pipeline into an analytical engine for applications that may range from historical analysis to meteorological forecasts and projections. However, with various sources accessible to accomplish these scientific objectives, the data must be treated or transformed in such a way such that the various sources are compatible with each other and standardized into a common schema to help ensure data integrity and validity. By implementing best practices, understanding the source data schema, and the target data schema, data scientists may extract from these three sources to transform into a common and practical schema, and then load it for higher-level analytics or other business intelligence applications.

**Method**

First, the ‘pymysql’ Python package allowed for an object to be created to take in network socket information as arguments to be passed to a MySQL server. This object acts as the interface in which embedded MySQL commands may be passed over Python such as selecting a database, passing MySQL syntax operations, or triggering execution of specific commands to the server. The server is directed to select a database titled “Weather\_Database” where three tables are created if they do not exist: DailyWeatherConditions, DailyWeather Alerts, and HistoricalWeatherData. The day, month, and year are integer values that cannot contain null values as these will serve as the primary key for each table. Generally, any lists of areas are stored as strings of VARCHAR of length 5000 as these lengths may vary depending on the quantity of alerts upon API call. Headline information is stored as a string of VARCHAR of length 150 as a reasonable length to capture the topic of the descriptive information. Description of the weather alert is stored as strings of TEXT as this narrative information contains verbose and extended data pertaining to the weather alert. Values such as humidity are stored as integer values as decimal values are not observed.

With the MySQL server configured, data was extracted from three different sources. The ‘nwsapy’ Python package allowed interfacing with the official NWS API via Python language and associated methods and attributes. A user agent method established a connection to the NWS server by passing API and identification information as arguments to the user agent object. Once the connection is established, the API is called to extract weather alerts for a given location. The “get\_alert\_by\_zone” method was called on the user agent object and was passed a six-character argument. This argument consisted of three A through Z characters followed by three zero through nine characters as a string value corresponding with the NWS’s identified geographic zones. An “AlertByZone” object is returned, which may be printed to standard output as either a dictionary or a Pandas DataFrame using the returned object’s respective methods.

However, not all geographic areas may have an active alert at the time the API is called, so the object that is returned to the user may contain a null value. In order to avoid inserting malformed null data into the MySQL database, this object is checked if the value is null, which if it is will set values in accordance with the target schema containing the primary keys used across all three sources: Date, Month, and Year. For this and the remaining values in accordance with the schema, null values and a headline of “No Alert Today” are staged in a string variable, which will be passed to the MySQL connection object for execution as an “INSERT INTO” statement.

The pipeline takes in a second data source, the UCSD historical weather dataset, in the form of a static comma-separated value [CSV] file. Data from this source contains daily observations of weather data from the period year 1939 to 2019, which was originally collected by the National Oceanic and Atmospheric Administration. Temperature data was transformed into the same units per the schema and dates were parsed into their respective day, month, and year values for primary key purposes. Columns that were valid, but did not have significant overlap or contextual importance with the other data sources were dropped, primarily consisting of temperature and wind speed flag information, of which the data was sparse in density. The transformed data was parsed into their respective schema variables and staged into a string variable containing the MySQL statement to be executed as an “INSERT INTO” statement.

The pipeline takes in the final data source, the OpenWeather API, which utilizes the ‘requests’ Python package to extract data via a uniform resource locator with an embedded latitude and longitude for San Diego, California as well as an assigned application identifier. Similarly to the two sources above, data only from the current day is extracted, which is returned in a dictionary object. Data is identified in accordance with the target schema and staged into a string variable that is to be executed as an “INSERT INTO” statement much like the previous sources.

Lastly, a daily “CRON” operation is executed via a separate Python file to automate daily extraction from the data sources and into the MySQL database, which is set for 10:00 am daily. As such, the pipeline extracts, by calling the respective API procedures, executes the transformation in the provided schema, and loads the data into the MySQL database using the embedded INSERT INTO statement.

**Results**

**Discussion**

**References**

*API Web Service.* (n.d.) National Weather Service (NWS) National Oceanic and Atmospheric Administration (NOAA). Retrieved February 1, 2024, from https://www.weather.gov/documentation/services-web-api.

*San Diego Weather - GHCN Daily.* (n.d.) The Scripps Institution of Oceanography at University of California, San Diego (UCSD). Retrieved February 1, 2024, from https://oceaninformatics.ucsd.edu/datazoo/catalogs/ccelter/datasets/13.

*Weather API.* (n.d.) OpenWeather. Retrieved February 1, 2024, from https://openweathermap.org/api.