Xander Atalay
DS 3002 Final Project
05/12/2022

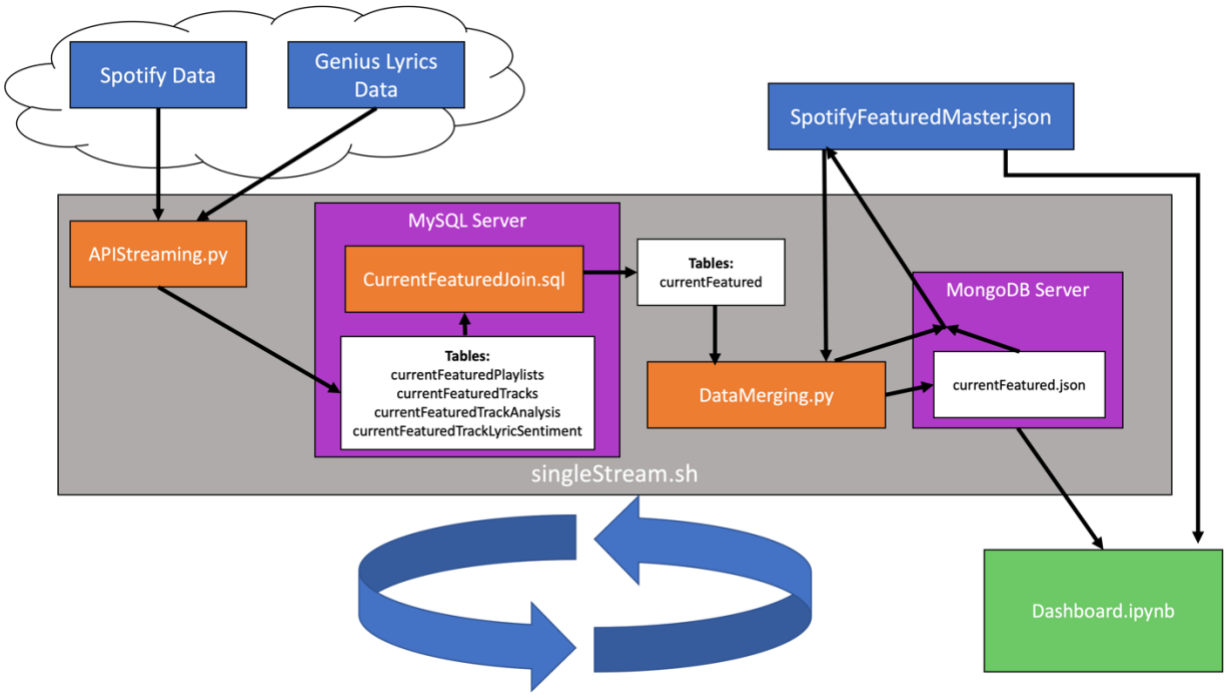
Project Overview:

Throughout the day, Spotify has a changing set of playlists that are recommended for all users. For this project, I am interested in streaming featured playlist data from the Spotify API, acquiring song data for every playlist, and obtaining track lyrics from the Genius API. Next, I'll perform sentiment analysis on the song data, and combine that with other song metrics to create a simple dashboard that shows the differences between the playlists that Spotify recommends at different timepoints. To do this and simultaneously demonstrate my knowledge of SQL and NoSQL technologies, I'll offload API data to a SQL and Mongo database, performing light modifications to the data in each.

Driving Question:

How do the characteristics of featured Spotify playlists compare to one another, and how do the playlist recommendations change at different points in the day?

Final Pipeline:



Fitting Project Requirements:

- **"Your solution must include a Date dimension to enable the analysis of the business process over various intervals of time"**

    - The date dimension that I will be using is the *featured_dt* column, which shows when the data was streamed in and thus when the track was in a featured playlist.

- **"Your solution must include at least 3 additional dimension tables (e.g., buyers, sellers, products)"**
    - This solution includes 4 dimensional tables before merging:
        - Featured Playlists
        - Featured Track Fact Table
        - Featured Track Analysis
        - Featured Track Lyric Sentiments

- **"Your solution must include at least 1 fact table that models the business process"**
    - The fact table is created with a SQL script that joins all of the dimension tables.

- **"Your solution must populate its dimensions using data originating from multiple sources:"**
    - **A relational database like MySQL, Oracle or SQL Server**
        - I offload my data to a SQL database to create a fact table.
    - **A NoSQL database like MongoDB, Redis, Cassandra or HBase**
        - I move my fact table from a SQL database a Mongo database where I join the active streamed data with the all-time acquired data.
    - **An API that returns a message payload (e.g., JSON, CSV, text)**
        - I retrieve my original data from two APIs

- **"Your solution must integrate datum of differing granularity (static and near real-time)"**
    - The final dashboard includes visualizations of currently featured playlists as well as a combination of all previously streamed data

- **"Your solution must include one or more visualizations that demonstrate the business value of your solution."**
    - My simple data visualization is completed in Jupyter notebook using plotnine.

- **"Your solution must demonstrate accumulating data that originates from a real-time (streaming) data source for a predetermined interval (mini-batch), integrating it with reference data, and then using the product as a source for populating some aspect of your dimensional data mart."**
    - I obtain real-time data from the Spotify API by running the acquisition script on a set interval and use Mongo DB to integrate it with reference data.
- ***If you opt to use any cloud-hosted services then please identify them so we may faithfully replicate your project. –*** I am not opting to use any cloud-hosted services.

"Note: You may utilize any combination of on-premises and/or Cloud service technologies."

UPDATE - There is now also a databricks .dbc file that shows ingestion of this data for the potential creation of a dashboard