Emma Hickey
DS 2002
11/11/24

Midterm Project

Overview, my process, and deployment strategy:

For my DS 2002 Midterm project, I will be designing a dimensional data mart, called Music_Data_Mart, using the Chinook Database to represent the business process of music sales, specifically through invoices and inventory information. I then will develop an ETL pipeline that extracts, transforms, and loads this data into my data mart using SQL database tables, a CSV table file, and JSON table file (using MongoDB). Lastly I will be demonstrating proper functionality through the SQL query that selects and then aggregates data.

After I set up my various database sources and ran the Chinook and date dimension scripts in MySQL, I then exported the necessary artist and album tables. I established the necessary connections for pandas, pymongo, and sqlalchemy libraries and defined the functions to then be able to read the album table CSV file into the pandas dataframe, make transformations and then return it to MySQL as a dimension table. I used the JSON artist table file to read data from the MongoDB collection and repeated a similar process. Next, I made my fact table through a series of lookup operations and transformations and then returned it back to MySQL. I then was able to perform the final query that provided summary information for customer orders.