Kierany Lewis
DS 2002

DOCUMENTATION

For my midterm, I created data marts that model sales, customers, products, and specific dates that allows for queries that analzye performance, trends, and customer behavior over time. To do this, I used the adventureworks data base to create fact_sales, dim_product, dim_customers, and dim_date. I used INSERT and JOIN statements to populate them with data from the adventureworks database, and wrote a stored procedure to generate and insert date records. To verify the schema was built correctly, I ran SELECT COUNT (*) queries to conform row counts, checking for missing kets, and performing test joins between the fact and dimension tables. I also ran summary queries to confirm the relationships were correct and the data mart would support queries.

In Python, I built an ETL pipelines using pandas and SQLAlchemey to connect MySQL to transform and enrich my data. I used MongoDB and a JSON file I created that represented additional product and consumer information, then used helper functions to read those files and make dataframes. Then, I converted the text fields for consistent capitalization and transformed list values into strings to load the data into staging tables in MySQL. From there, I used SQL UPDATE statements through Python to merge the data into my existing dimension tables. Lastly, I validated the ETL process by running select statements to confirm the data appeared correctly.