Capstone Project

End-to-End Data Engineering workflow

Abdullah Alsalloum, 5/5/2025

Project Overview

- Module 1: MySQL.
- Module 2: MongoDB.
- Module 3: PostgreSQL.
- Module 4: Looker Studio.
- Module 5: ETL pipeline.
- Module 6: PySpark.

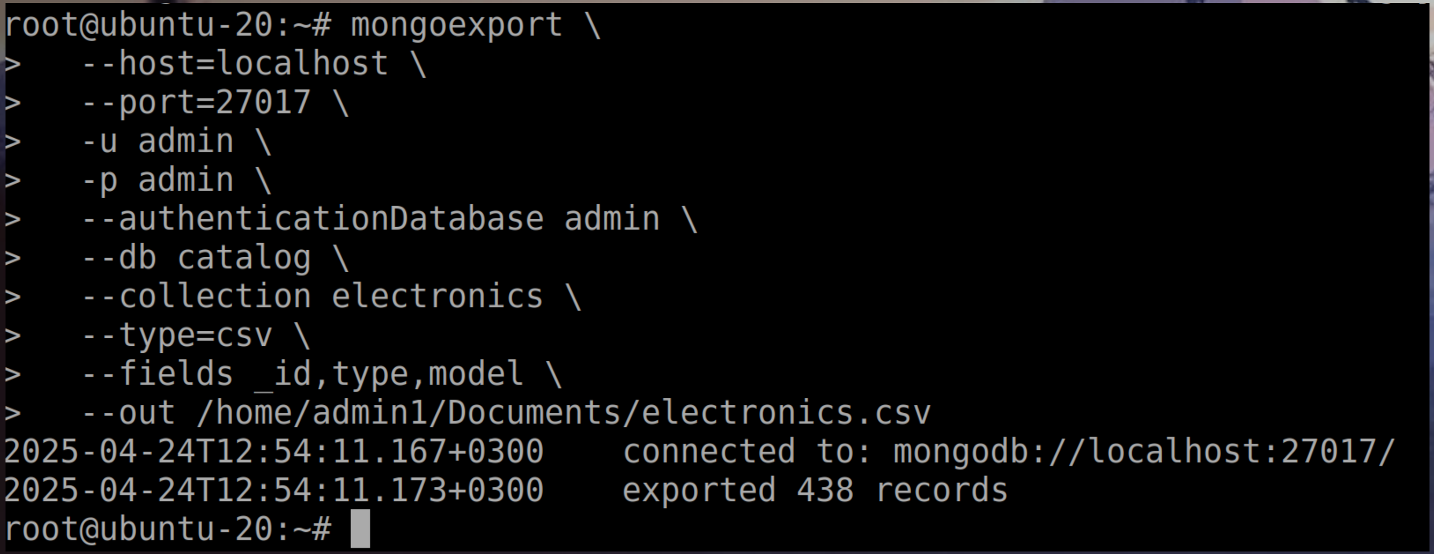
MySQL

- Task 1 & 2: Create Sales DB, sales_data table.
- Task 3: Import Oltp.csv.
- Task 4: Create index on timestamp.

MongoDB

- Task 1: Import catalog.json.
- Task 2: Create index on type field.
- Task 3: Export selected fields to csv.

Exported Fields

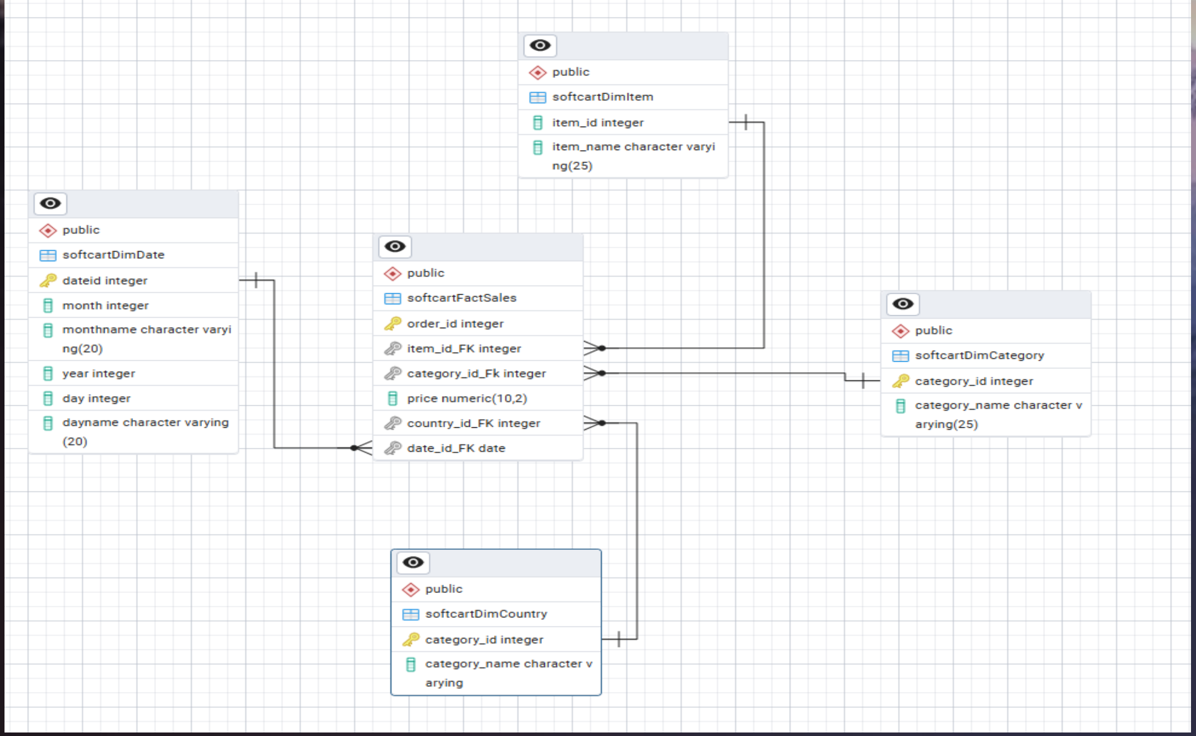


```
root@ubuntu-20:~# mongoexport \
>   --host=localhost \
>   --port=27017 \
>   -u admin \
>   -p admin \
>   --authenticationDatabase admin \
>   --db catalog \
>   --collection electronics \
>   --type=csv \
>   --fields _id,type,model \
>   --out /home/admin1/Documents/electronics.csv
2025-04-24T12:54:11.167+0300    connected to: mongodb://localhost:27017/
2025-04-24T12:54:11.173+0300    exported 438 records
root@ubuntu-20:~# 
```

PostgreSQL

- Task 1: Create Dim and Fact tables.
- Task 2: Set foreign keys.
- Task 3: Create relations between tables.

UML Class Diagram



public
softcartDimItem
item_id integer
item_name character varying(25)

public
softcartDimDate
dateid integer
month integer
monthname character varying(20)
year integer
day integer
dayname character varying (20)

public
softcartFactSales
order_id integer
item_id_FK integer
category_id_Fk integer
price numeric(10,2)
country_id_FK integer
date_id_FK date

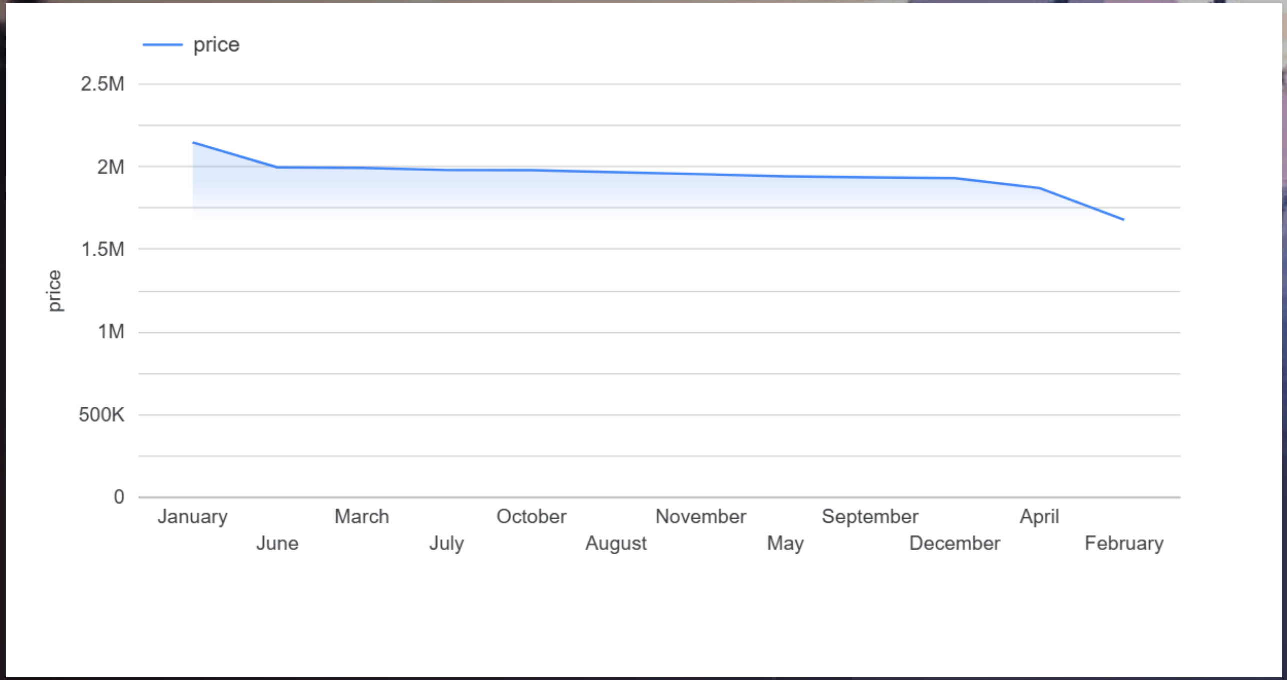
public
softcartDimCategory
category_id integer
category_name character varying(25)

public
softcartDimCountry
category_id integer
category_name character varying

Looker Studio

- Task 1: Prepare data for visualizations.
- Task 2: Import pie chart, line chart and bar chart.

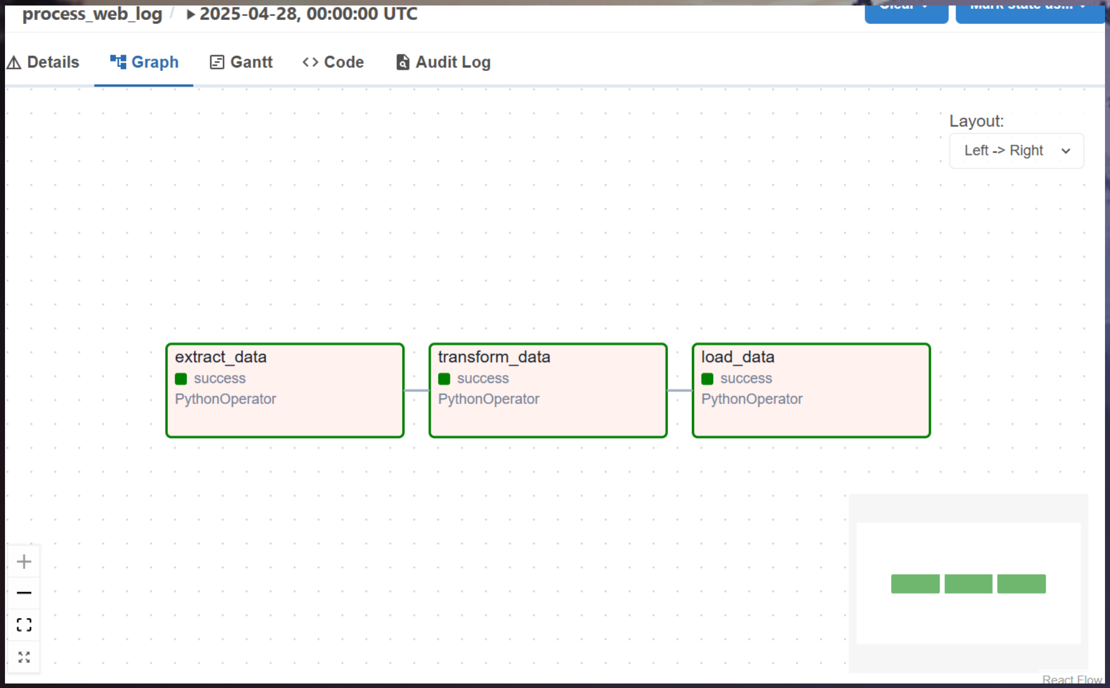
Looker Studio



ETL Pipeline

- Task 1: Extract max(id) from postgres DB.
- Task 2: Extract ids greater than max(id) from MySQL DB.
- Task 3: Load new extracted data into postgres DB.

Apache Airflow ETL



process_web_log / ▸ 2025-04-28, 00:00:00 UTC

Layout: Left -> Right

extract_data
🟩 success
PythonOperator

transform_data
🟩 success
PythonOperator

load_data
🟩 success
PythonOperator

PySpark

- Task 1: Importing Linear regression model.
- Task 2: Split data into training set and test set.
- Task 3: Make predictions for target column.

Summary

- Tools used: (MySQL, MongoDB, Looker Studio)
- Skills applied: (ETL, data modeling, indexing, querying)
- What I learned ?

Questions ?