

# Data Engineering Case Study

Nirosh Kumar R• AA.SC.P2MCA2107478

Project link <https://github.com/NiroshKumarR/etlretailsales>

# Introduction

- This project focuses on performing Extract, Transform, and Load (ETL) operations on a Retail Sales dataset.
- The objective is to process the raw data, transform it into a suitable format for analysis, and store it in a PostgreSQL database.

# Data Transformation

- The raw data was analyzed and transformed using Python and the Pandas library.
- Data transformation steps included cleaning, filtering, aggregating, and structuring the data.
- The transformed data provides valuable insights into sales trends, popular products, and category performance.

# Data Storage and Retrieval:

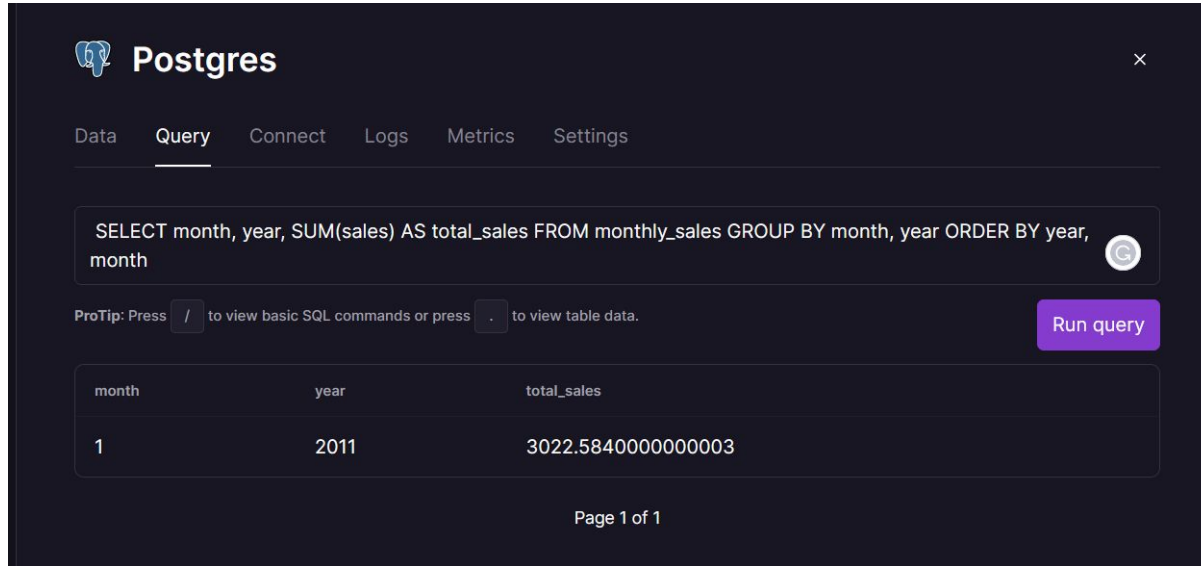
- The transformed data is stored in a PostgreSQL database for efficient storage and retrieval.
- The psycopg2 library is used to establish a connection to the database and execute SQL queries.
- The database schema is designed to accommodate the transformed data, ensuring easy access for analysis.

# Data Analysis and Insights:

- The transformed data enables comprehensive analysis of retail sales performance.
- Key insights include monthly sales trends, top-selling products, and revenue by region.
- Visualizations and charts are used to present the findings and aid in understanding the data.

# Business Questions:

- Monthly sales trends over the years



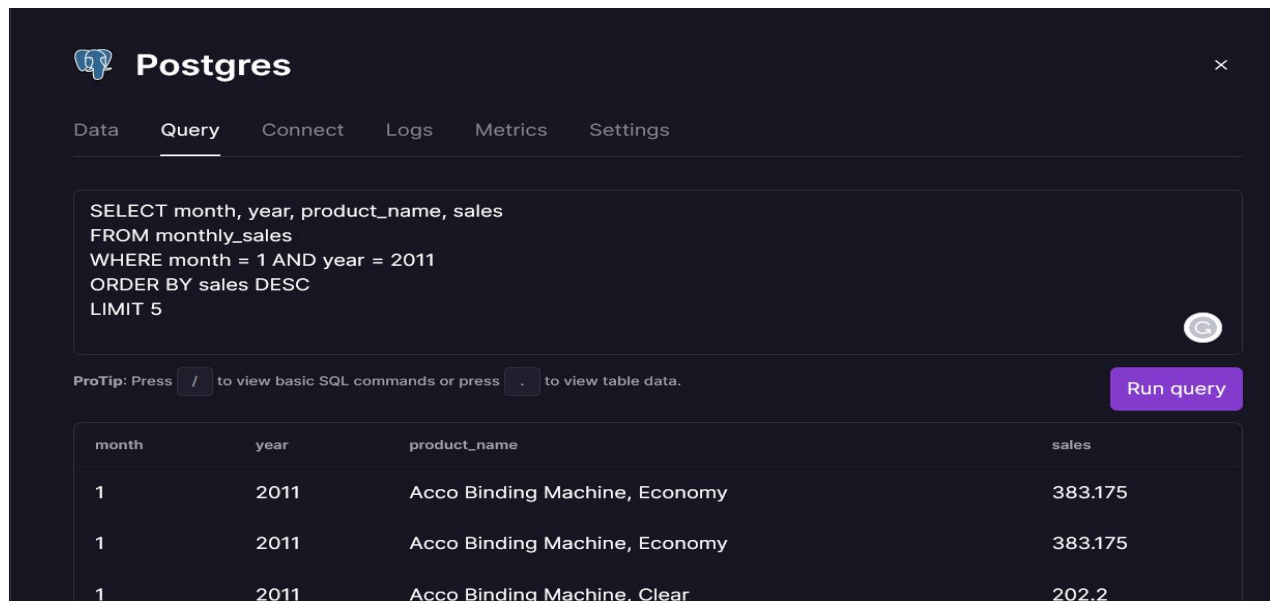
The screenshot shows the PostgreSQL web interface. At the top, there's a header with the PostgreSQL logo and the word "Postgres". Below the header is a navigation bar with tabs: "Data", "Query" (which is active), "Connect", "Logs", "Metrics", and "Settings". The main area contains a text input field with the following SQL query: `SELECT month, year, SUM(sales) AS total_sales FROM monthly_sales GROUP BY month, year ORDER BY year, month`. To the right of the query is a circular refresh icon. Below the query field is a "ProTip" message: "Press / to view basic SQL commands or press . to view table data." To the right of this message is a purple button labeled "Run query". Below the query field is a table with the following data:

month	year	total_sales
1	2011	3022.5840000000003

At the bottom of the interface, it says "Page 1 of 1".

# Business Questions:

- Top-selling products in a given month and year



The screenshot shows the PostgreSQL web interface. At the top, there's a navigation bar with tabs: Data, Query (selected), Connect, Logs, Metrics, and Settings. Below the tabs, a query editor contains the following SQL code:

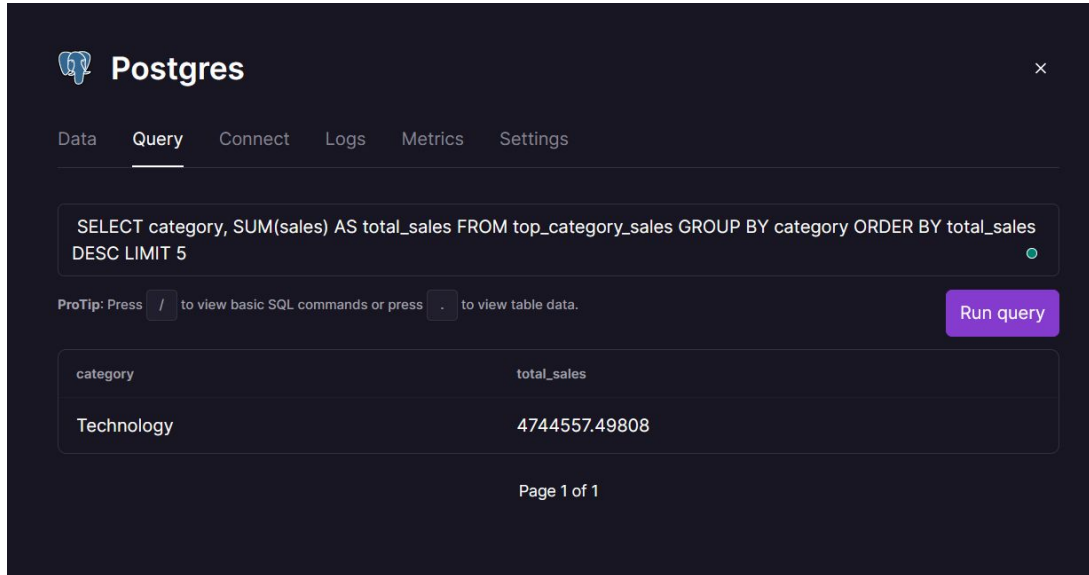
```
SELECT month, year, product_name, sales
FROM monthly_sales
WHERE month = 1 AND year = 2011
ORDER BY sales DESC
LIMIT 5
```

Below the query editor, a "ProTip" message states: "Press / to view basic SQL commands or press . to view table data." To the right of this message is a purple "Run query" button. Below the query editor, a table displays the results of the query:

month	year	product_name	sales
1	2011	Acco Binding Machine, Economy	383.175
1	2011	Acco Binding Machine, Economy	383.175
1	2011	Acco Binding Machine, Clear	202.2

# Business Questions:

- What are the top-selling product categories overall?



The screenshot shows the PostgreSQL query editor interface. At the top, there's a header with the PostgreSQL logo and the word "Postgres". Below the header, there's a navigation bar with tabs: "Data", "Query", "Connect", "Logs", "Metrics", and "Settings". The "Query" tab is currently selected. In the center, there's a text area containing the following SQL query:

```
SELECT category, SUM(sales) AS total_sales FROM top_category_sales GROUP BY category ORDER BY total_sales DESC LIMIT 5
```

Below the query text area, there's a "ProTip" message: "ProTip: Press / to view basic SQL commands or press . to view table data." To the right of this message is a purple button labeled "Run query". Below the query text area, there's a table displaying the results of the query:

category	total_sales
Technology	4744557.49808

At the bottom of the interface, it says "Page 1 of 1".



## Conclusion:

- The ETL Retail Sales Data Engineering Project demonstrates the importance of data preparation for meaningful analysis.
- The project highlights the use of Python, Pandas, and PostgreSQL for efficient data processing and storage.
- The insights gained from the transformed data can drive informed business decisions and identify areas for improvement.

Thank You