Data Science Portofolio

Building Automated ETL Workflows with Airflow, MySQL, and PostgreSQL

Hello Everyone!

Passionate about turning raw data into actionable insights. With a background in management and skills in data science and machine learning, I bridge the gap between business needs and technological solutions.

Let's harness data to unlock business potential and drive innovation together.

Data Science

Dibimbing.id
Jul 2023 - Dec 2024

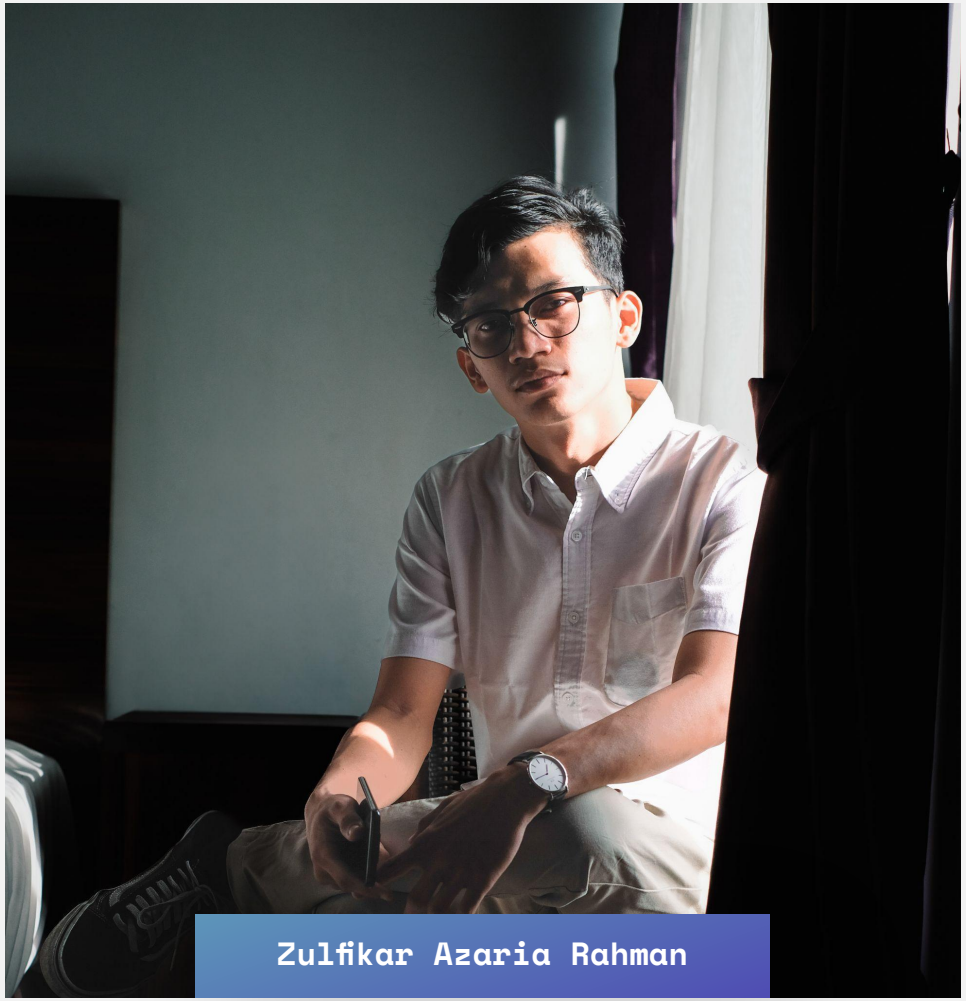
- Best Student Batch 22
- Best Final Project

MORE

Management

Universitas Muhammadiyah Surakarta
Jun 2018 - Jan 2022
IPK : 3.64

MORE



Zulfikar Azaria Rahman

Work Experience

Jan 2020 - Present

Tanarasa Photography
Owner

Jul 2022 - Jan 2023

PT. Maybank Indonesia Finance Tbk
Credit Marketing Officer

May 2019 - Jul 2021

Arto Moro
Videographer

Aug 2021 - Mar 2021

PT. Bangkit Laju Jaya
Social Media Marketing

Certification

Mimo

SQL

Revo U

Intro To Data Analytics

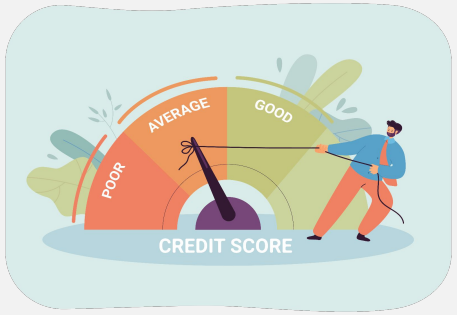
Dibimbing.id

Bootcamp Data Science

ID/X Partners X Rakamin Academy

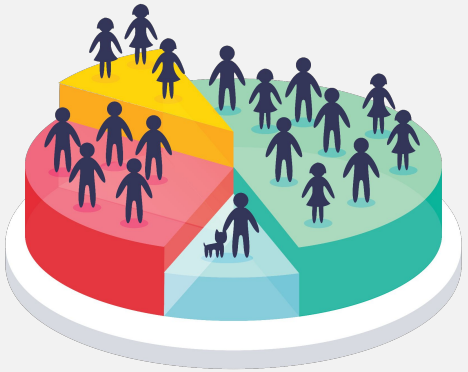
Project based Internship

Data Science Projects



Predictive Credit Risk Modeling : Leveraging Decision Trees, XGBoost, and Random Forest

MORE



Enhancing Airline Customer Strategies Through K-Means Clustering and LRFMC Analyis

MORE

Building Automated ETL Workflows with Airflow, MySQL, and PostgreSQL

Zulfikar Azaria Rahman

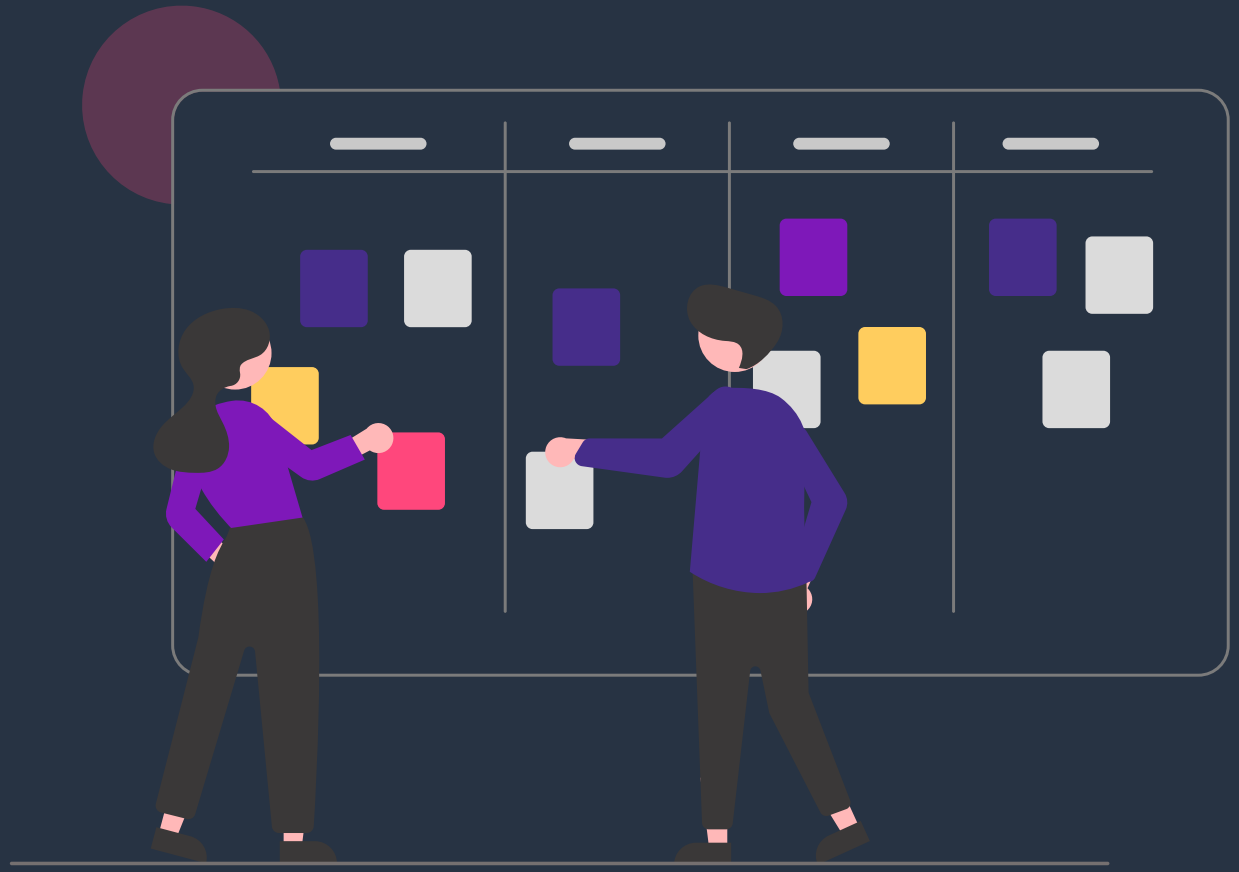
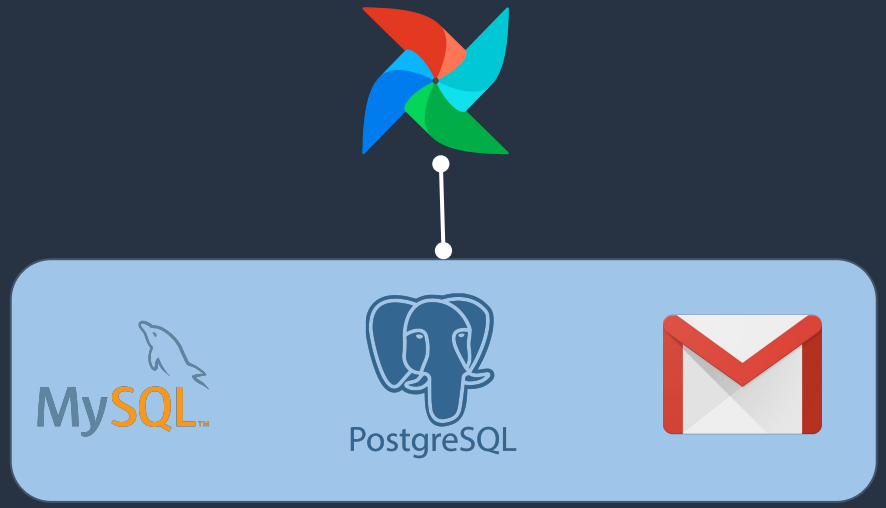


Table of Content

Project Overview

This project involves implementing a simple ETL using Apache Airflow to extract data from MySQL and load it into PostgreSQL every day at 7am. The processed data will be stored in a staging area in Parquet format. After the ETL process is complete, a simple analysis job is run and the results are emailed at 9am. There are 3 DAGs involved: ETL DAG, aggregation DAG, and report delivery DAG.



Project Background

Why This Project?

This project was designed to automate the Extract, Transform, Load (ETL) process using Airflow, with a focus on efficiency and automation of the data transfer process between databases. The goal was to make it easier to analyze data on a scheduled basis and automate the delivery of reports based on the latest data.

Who Benefits:

This project provides benefits to companies or teams that require data movement between database platforms as well as scheduled analysis that can be accessed daily.

Problem Statement

The main challenge in this project was how to automatically move data from MySQL to PostgreSQL on a daily basis, ensuring that the latest data was always available for analysis. In addition, the project also needed to be able to generate automated reports based on the results of the analysis performed, and deliver the reports at a predetermined time.

To address these issues, the goal of the project is to build a scheduled ETL pipeline using Apache Airflow. This pipeline would run the data extraction process from MySQL every day at 7 am, load it into the staging area in PostgreSQL, perform the aggregation process on the processed data, and send the analysis report via email at 9 am. Thus, this pipeline ensures that data is always up-to-date and reports can be received in a timely manner.

Data Platform Understanding

Data was extracted from MySQL and loaded into PostgreSQL in a staging area with Parquet format for easy processing and storage. PostgreSQL tables were used to store the final results after the aggregation process.

		Dag
01	**ETL** Every morning, the system retrieves sales transaction data from the previous day stored in the MySQL database, the data is temporarily stored in a special file called Parquet for efficiency.	
02	**Data Aggregation** The data is processed to calculate the total number of products sold and the total money received, and the results are stored in the PostgreSQL database.	
03	**Sending Report via Email** After all was done, an Excel report was sent automatically to email every 9am.	

id	transaction_date	product_name	quantity	total_amount
1	2024-09-18	Product A	5	100
2	2024-09-18	Product B	3	75
3	2024-09-18	Product C	10	200
4	2024-09-18	Product D	2	50
5	2024-09-18	Product A	7	140
18	2024-09-23	Product A	3	60
19	2024-09-23	Product A	4	80
20	2024-09-23	Product B	5	125
21	2024-09-23	Product C	4	80
22	2024-09-23	Product D	6	150
23	2024-09-23	Product D	2	50

Data Understanding

In this project, I designed to move transaction data from MySQL to PostgreSQL based on the transaction date, then I analyzed the data to produce a new table called sales_aggregated, which contains transaction_date, product_name, total_quantity, and total_sales columns. Later, the results of the data analysis will be sent via email to related parties automatically and on a scheduled basis.

MySQL

id	transaction_date	product_name	quantity	total_amount
1	2024-09-18	Product A	5	100
2	2024-09-18	Product B	3	75
3	2024-09-18	Product C	10	200
4	2024-09-18	Product D	2	50
5	2024-09-18	Product A	7	140
18	2024-09-23	Product A	3	60
19	2024-09-23	Product A	4	80
20	2024-09-23	Product B	5	125
21	2024-09-23	Product C	4	80
22	2024-09-23	Product D	6	150
23	2024-09-23	Product D	2	50

PostgreSQL

Before

id	transaction_date	product_name	quantity	total_amount
1	2024-09-18	Product A	5	100
2	2024-09-18	Product B	3	75
3	2024-09-18	Product C	10	200
4	2024-09-18	Product D	2	50
5	2024-09-18	Product A	7	140

After

id	transaction_date	product_name	quantity	total_amount
1	2024-09-18	Product A	5	100
2	2024-09-18	Product B	3	75
3	2024-09-18	Product C	10	200
4	2024-09-18	Product D	2	50
5	2024-09-18	Product A	7	140
18	2024-09-23	Product A	3	60
19	2024-09-23	Product A	4	80
20	2024-09-23	Product B	5	125
21	2024-09-23	Product C	4	80
22	2024-09-23	Product D	6	150
23	2024-09-23	Product D	2	50

Transformation & Consideration

The system retrieves data from the data source, using MySQL hooks in Airflow, SQL queries extract transaction data from the previous day, then saved to the staging area in Parquet format for space efficiency and speed in loading and storing temporary data. The data in Parquet is then loaded into PostgreSQL.

transaction_date	product_name	total_quantity	total_sales
2024-09-23	Product A	7	140
2024-09-23	Product B	5	125
2024-09-23	Product C	4	80
2024-09-23	Product D	8	200

Data Modeling

The DAG aggregates sales data by transaction_date and product_name, calculating:

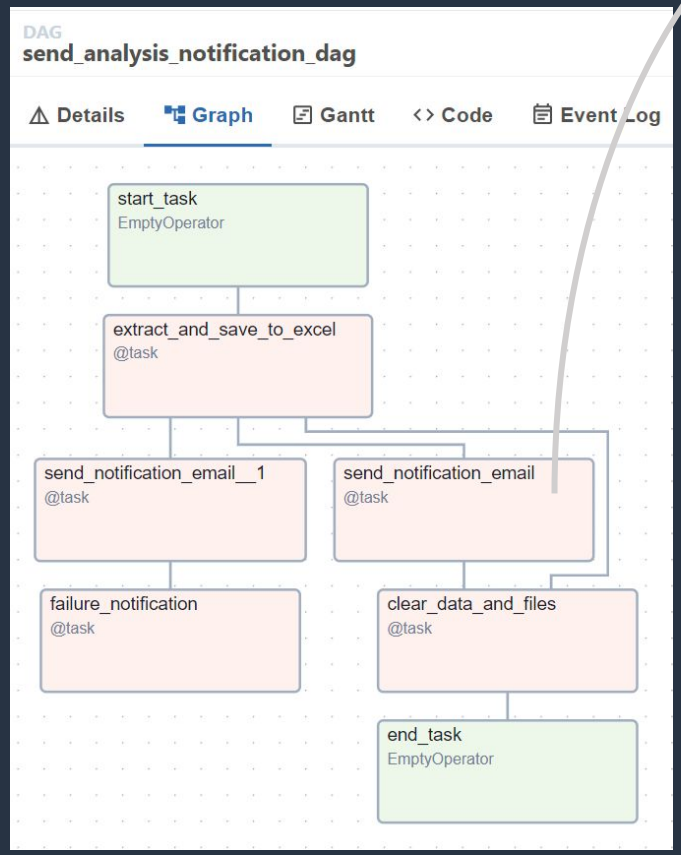
- **Total Quantity**: Sum of the quantity sold per product per day.
- **Total Sales**: Sum of the total_amount per product per day.

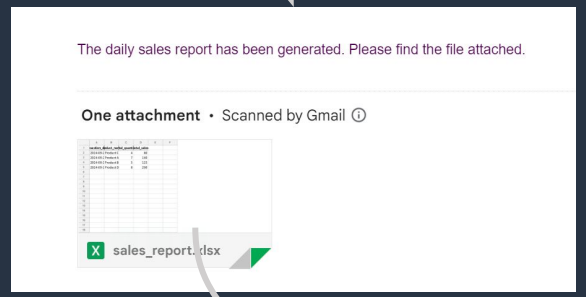
This summarizes daily sales totals and stores them in the sales_aggregated table.

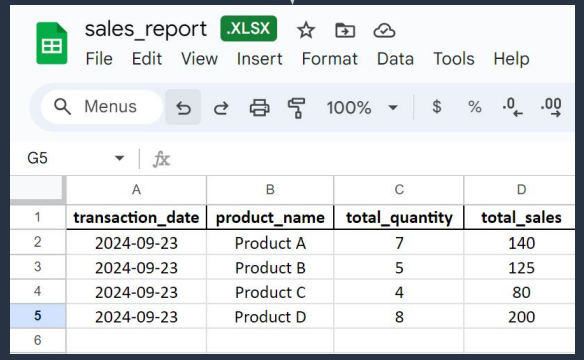
Table Design:

- **Primary Key**: Composite key (transaction_date, product_name) ensures unique entries and fast access.
- **Efficient Storage**:
  - Uses NUMERIC for sales and INT for quantity.
  - Parquet format is used for staging, offering efficient processing and compression.
  - Batch inserts via SQLAlchemy enhance write performance.

This setup allows for fast queries and efficient storage of large sales datasets.







	A	B	C	D
1	transaction_date	product_name	total_quantity	total_sales
2	2024-09-23	Product A	7	140
3	2024-09-23	Product B	5	125
4	2024-09-23	Product C	4	80
5	2024-09-23	Product D	8	200
6				

Report & Sent to Email

This process automatically retrieves sales data from the sales_aggregated table in PostgreSQL, stores it in an Excel file, and then sends the report via email every day. After the report is sent, the Excel file, Parquet staging data, and sales_aggregated table are deleted to save space. If an error occurs in the process, the system will automatically send a failure notification. This flow ensures that daily sales reports are generated and sent without the need for manual intervention.

Conclusion & Recommendations

This pipeline processes data automatically on a daily basis, delivering updated reports efficiently. Challenges faced include ensuring data is accurate and non-duplicate. Solutions such as Airflow's automatic scheduling and cleaning of the staging area after each process helped overcome this.

In the future, this daily ETL can be further improved, develop a real-time dashboard that displays aggregated data directly from the PostgreSQL database, so that business decisions can be made based on more up-to-date data.

Get Connected

Zulfikarazaria

Zulfikarazaria

Zulfikarazaria@gmail.com

Thank You

FOR YOUR ATTENTION