

Data Engineering Program Capstone 3 - Outline Property of Rizky Fajar Aditya

Table of Contents:

Background and Study Case

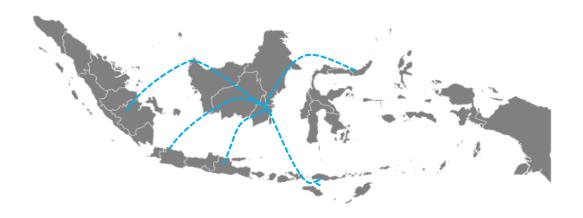
Development Cycle and Process

Project Results 3

Data Engineering Program Capstone 3 - Background and Study Case Property of Rizky Faiar Aditya

Study Case: B40 Inc. faces growing challenges in Data Access and Quality, hindering decision-making and business agility.

Background and Study Case:



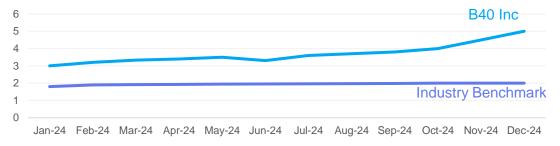
B40 Inc is a major energy company in Indonesia that distributes Biodiesel product to Mining Companies in every corner in Indonesia.

The Marketing and Sales Team at B40 Inc. is facing challenges in gathering data for their reports and insights, which is impacting their decision-making processes. To address this, they have enlisted the help of Data Engineering team to design and implement a robust data pipeline that will streamline data collection and processing, enabling faster reporting and more informed decision-making.

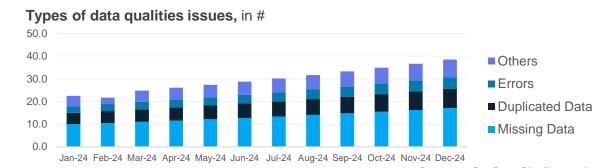
Several challenges that makes B40 Inc is on the verge to develop a robust data pipeline:

The time required to access critical data increases by 4% each month, as the duration from data collection to reporting grows, which delays their ability to respond to market changes

Time needed teed to collect data to reporting, in hours



Inconsistent data quality makes it challenging to perform comprehensive data analysis



Please note: The case study and data presented in this slide are hypothetical and created for illustrative purposes only

To address the issue, The SDLC cycle is applied to develop a robust data pipeline tailored to meet B40 Inc.'s needs.

Maintenance

Continuously monitor and improve the data pipeline based on user feedback and evolving business needs



Planning

Defining scope and requirements of data pipeline to address current challenges



Design

Design the data pipeline architecture and outline the steps to ensure scalability, data quality, and efficient reporting



Deploy the data pipeline into a production environment for ongoing use.



Development

Building data pipeline based on design specifications



SDLC

Testing

Test the data pipeline for accuracy, reliability, and performance under real-world conditions

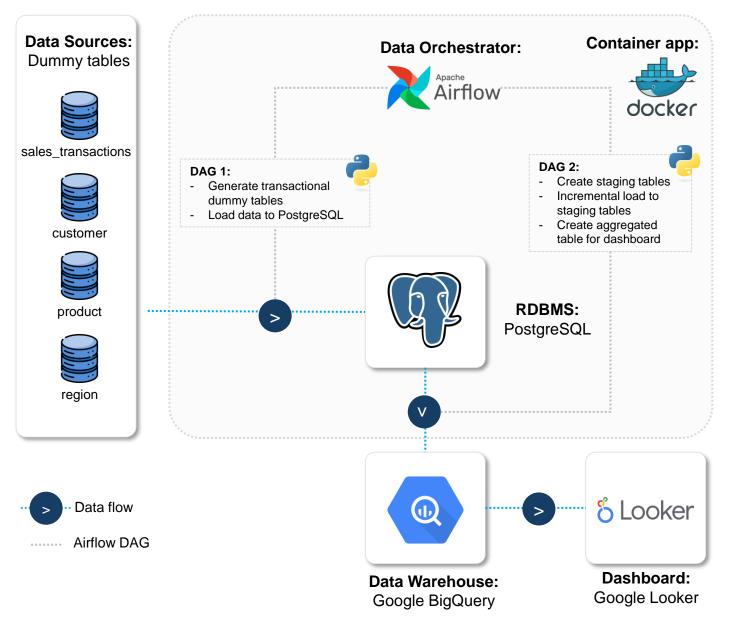
Stage 1: Planning and gathering design requirements

Defining scope and requirements of data pipeline to address current challenges:

Key Requirement	Description		
Key Stakeholders	Marketing and Sales Team		
Data Sources / Schema	Generate 4 dummy tables: 1. sales_transactions (fact) 2. product (dim) 3. customers (dim) 4. region (dim)		
Transactional data storage	PostgreSQL Server		
Data Warehouse	BigQuery		
Batching Method	Data will be updated in daily basis at 10 AM		
Incremental Load	Load only new or updated data to reduce processing time		
Data Granularity	Daily		
Reporting	Data and report can be accessed via Dashboard (Google Looker)		

Data Engineering Program **Capstone 3 – Development Cycle and Process** Property of Rizky Fajar Aditya

Stage 2: Data Pipeline Architecture Design



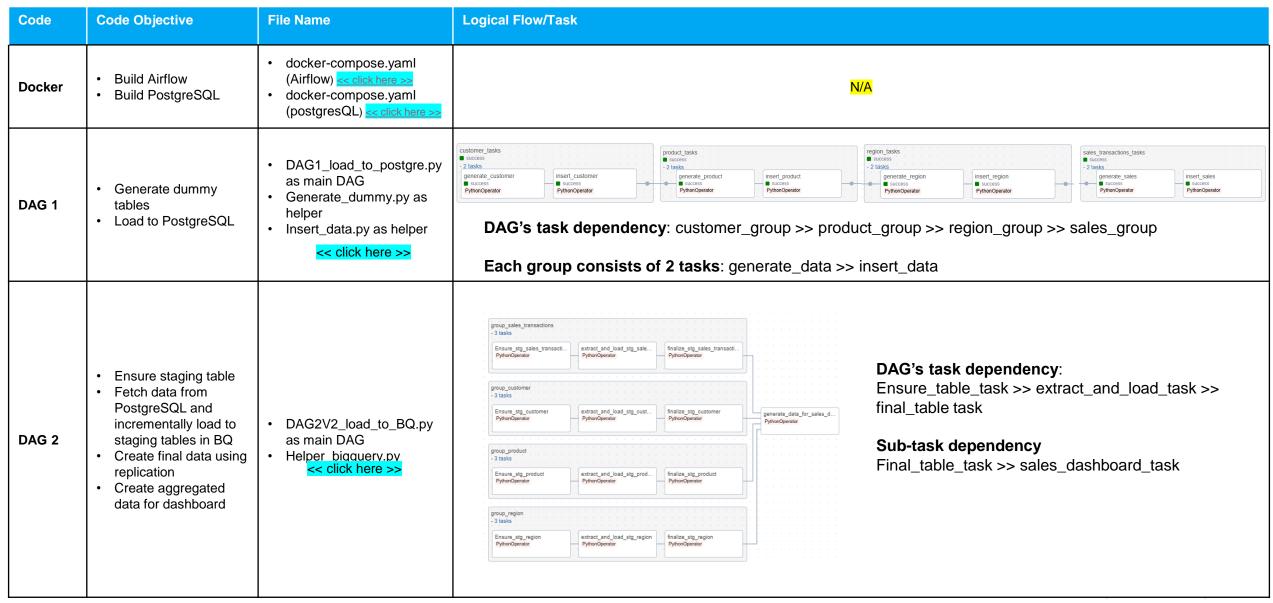
Tools utilized in this project:

Function	Tools	Explanation
Data Orchestrator	Apache Airflow	Able to define task dependencySuitable for batching pipeline
Program Script	Python	Versatile programming language for building data pipeline and processing
Containerization	Docker	Enables all apps run consistently for development
RDBMS	PostgreSQL	 The data sources are structured data, suitable for storing data with relationships.
Data Warehouse	Google BigQuery	 Partitioning enables faster querying, and improving performance Good performance warehouse that's scalable and efficient
Dashboard	Google Looker	Seamlessly integrates with Google BigQuery, allowing real- time data updates

Data Engineering Program

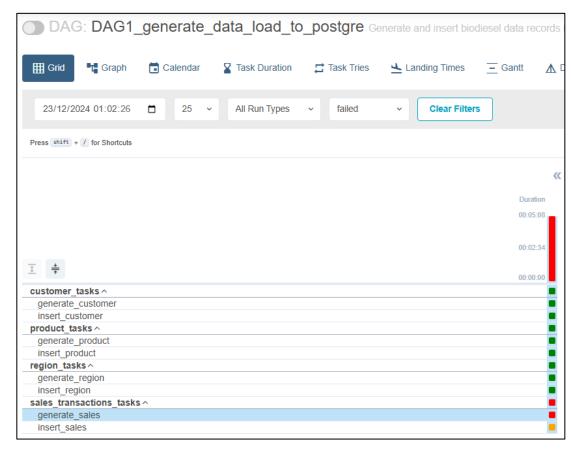
Capstone 3 – Development Cycle and Process

Stage 3: Code Development



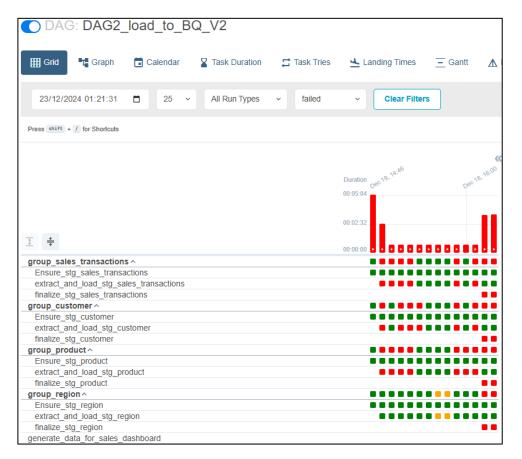
Data Engineering Program Capstone 3 – Development Cycle and Process Property of Rizky Fajar Aditya

Stage 4: Testing



DAG1 Testing Log

During testing, only one error was recorded. The task failed because the function to connect to PostgreSQL was not available in the code. As a result, the code was updated and successfully worked at this stage.



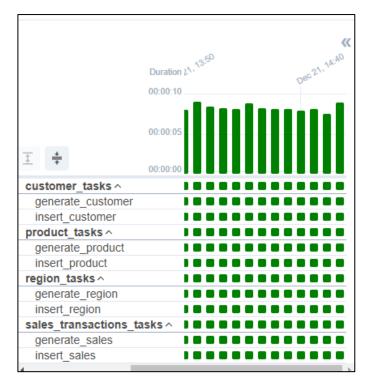
DAG2 Testing Log

During testing, 13 errors was recorded. The task failed due to various reasons: failed fetching, failed ensuring staging tables, etc. Thus, every error was thoroughly learned and revised. As a result, the code was updated and successfully worked at this stage.

Data Engineering Program **Capstone 3 – Development Cycle and Process** Property of Rizky Fajar Aditya

Stage 5 & 6: Deployment and Maintenance

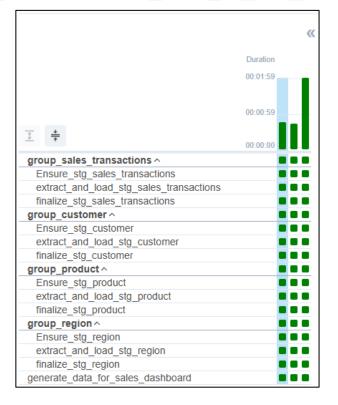
DAG: DAG1_generate_data_load_to_postgre



DAG1 Testing Log

DAG1 was deployed with a scheduled run on December 19, 2024. The task will execute automatically every 5 minutes to generate data and load to PostgreSQL. The scheduled tasks have been successfully performed and are running smoothly. Monitoring and maintenance will be closely overseen.





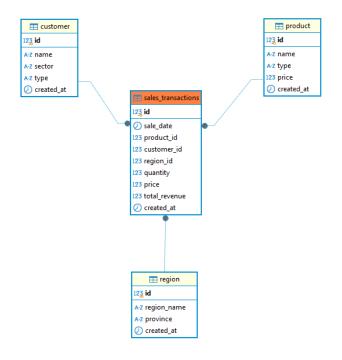
DAG2 Testing Log

DAG2 were deployed using scheduled run in 20 December 2024. The task will run automatically daily at 10.00 AM local Jakarta time. The scheduled tasks have been successfully performed and are running smoothly. Monitoring and maintenance will be closely overseen.

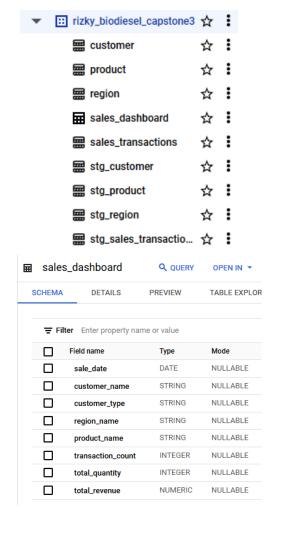
Data Engineering Program Capstone 3 - Project Results Property of Rizky Fajar Aditya

Project Results

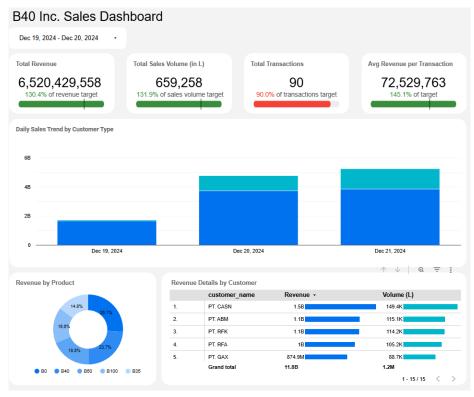
Successfully generate data, load data, and build relationship between tables in PostgresSQL- all orchestrated using Airflow



Successfully fetched data from Postgre, incrementally load to staging stables in BigQuery, create sales_dashboard table as source for dashboard



Created dashboard to generate business insights and informations: reducing reporting time through automation, increase data quality, and enhancing decision-making capabilities



<< Dashboard link click here >>

