The Data Engineer Learning Path

Author: Mr. Thanaton Booncherd

Advisor: Asst. Prof. Dr.Prapaporn Techa-Angkoon

Abstract

This research aims to study the data engineering process to meet the increasing demand in the IT labor market. Today, organizations increasingly need to use data to drive decision-making and operations, resulting in a need for personnel with expertise in managing data effectively. This study selected Apache Airflow, an open-source tool for managing ETL (Extract, Transform, Load) processes, because it is popular and widely used in leading organizations. The selection of this topic and tool allows for study and learning without the need for internal company data, which avoids the disclosure of potentially confidential information. The results of this research will increase the understanding of the data engineering process and the use of Apache Airflow, which is beneficial for the development of professional skills in data engineering and can be applied in real work in the future.

Introduction

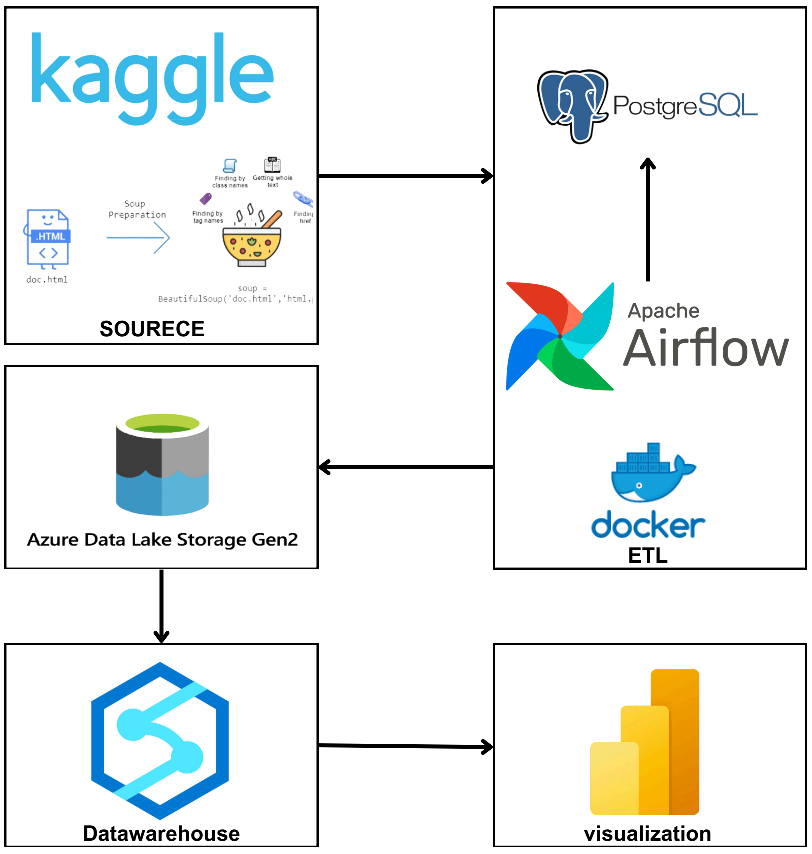
In today's data-driven business environment, Data Engineering is essential for effective decision-making. This Cooperative Education Project explores the Data Engineering process, focusing on Extract, Transform, Load (ETL) using Apache Airflow for managing data pipelines. Power BI is used for data visualization, while Data Lake Gen2 and Azure Synapse handle data storage and processing. The project enhances understanding of systematic data management and provides insights applicable to future Data Engineering work.

Technology

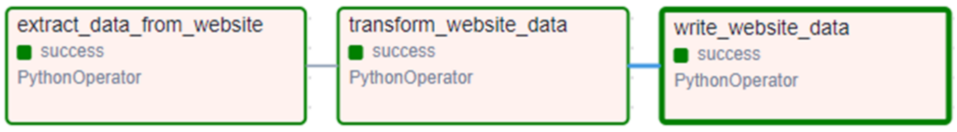
Methodology

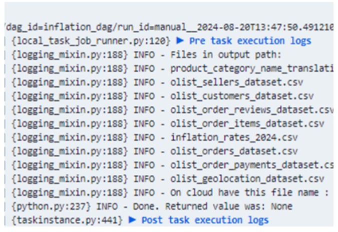
This study explores Data Engineering processes, focusing on Dim and Fact data in real-world scenarios. Using datasets from Kaggle and public sources, the research followed five key steps. Data Collection involved selecting high-quality datasets on e-commerce transactions and inflation rates. The ETL process utilized BeautifulSoup for web scraping, data cleaning, and formatting before loading. Data was stored in Azure Data Lake Gen2 for large-scale storage and structured in a Star Schema within Azure Synapse Analytics. Finally, Power BI was used to develop dashboards, providing insights into trading behavior and inflation impacts.

Methodology

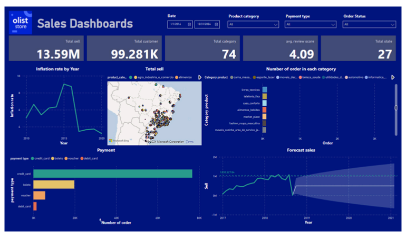


Result





```
dag_id=inflation_dag/run_id=manual__2024-08-20T13:47:50.491210
{local_task_job_runner.py:120} ▶ Pre task execution logs
{logging_mixin.py:188} INFO - Files in output path:
{logging_mixin.py:188} INFO - product_category_name_translati
{logging_mixin.py:188} INFO - olist_sellers_dataset.csv
{logging_mixin.py:188} INFO - olist_customers_dataset.csv
{logging_mixin.py:188} INFO - olist_order_reviews_dataset.csv
{logging_mixin.py:188} INFO - olist_order_items_dataset.csv
{logging_mixin.py:188} INFO - inflation_rates_2024.csv
{logging_mixin.py:188} INFO - olist_orders_dataset.csv
{logging_mixin.py:188} INFO - olist_order_payments_dataset.cs
{logging_mixin.py:188} INFO - olist_geolocation_dataset.csv
{logging_mixin.py:188} INFO - On cloud have this file name :
{python.py:237} INFO - Done. Returned value was: None
{taskinstance.py:441} ▶ Post task execution logs
```



Conclusion

This study successfully implemented a data engineering solution using Apache Airflow for ETL, integrating a data lake and data warehouse on Azure Synapse Analytics. Power BI was used for data visualization, creating a robust pipeline for efficient data analysis and decision-making. The project highlights how modern cloud-based architecture enhances data processing and management. The insights gained offer valuable knowledge for Business Intelligence and Data Engineering, with real-world applications.

Reference

[1] https://learn.microsoft.com/bs-latn-ba/azure/cloud-services/cloud-services-choose-me

[2] https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction

Co-operation Project, 204496 Department of Computer Science, 2024

Department of Computer Science, Faculty of Science Chiang Mai University