Project Proposal: Milestone Breakdown for ETL Pipeline

By

Hoan Lam

CIS 660 Data Engineering

Instructor: Dr. Sorio Bolt

Due: 4/6/2025

Project Content

Project Overview

For this project, I will examine an e-commerce consumer behavior dataset. The goal is to uncover insights into the datasets that will be beneficial to online business owners. To accomplish this efficiently, I will create an ETL pipeline using Kestra. The data source will be available on my GitHub for this project. After the data is processed, the cleaned data will be sent to PostgreSQL. Both Kestra and PostgreSQL servers will be running via Docker. At this stage, I will retrieve the data from PostgreSQL and import it into my Python notebook on my local computer for further data analysis. Interesting graphs or findings will then be shown via Google Looker Studio for a more user-friendly environment.

Data Source Selection – Kaggle

Link: https://www.kaggle.com/datasets/salahuddinahmedshuvo/ecommerce-consumer-behavior-analysis-data/data

Technology & Tool Requirements

- **ETL Platform –** Kestra
- **Database –** PostgreSQL
- **Server –** Docker
- **Visualization –** Google Looker Studio

Implementation

- **Extract -** A .cvs file will be uploaded to my GitHub page. Kestra will extract the data from my GitHub page.
- **Transform -** The data has missing and currency values. Both values will be cleaned accordingly using substitution and reformatting.
- **Load –** In the Kestra workflow, a table will be created if it does not exist in Postgres. All the data will be loaded to Postgres.
- **Data Analysis –** Jupiter Notebook to explore data patterns, distribution, and correlations.
- **Data Visualization –** Google Looker Studio to display relevant charts.

Project Timeline

Week #	Milestone	Description
1	**Research and Proposal Planning**	• Set up the working environment – Kestra, Postgre, Docker • Set up the basic workflow – Add Python, Google Looker Studio • Improve the basic workflow – Trial and error to get better results
2	**Complete the project**	• Optimize the basic workflow • Create useful data analysis with Python • Create visualizations with Google Looker Studio
3	**Documentation**	• Clean up • Documentation

Expected Challenges

- Set up the working environment and have everything working together
- Find useful information from the data

References

- https://www.kaggle.com/datasets/salahuddinahmedshuvo/ecommerce-consumer-behavior-analysis-data/data