Fashion Product Image Classifier System

Overview

The Fashion Product Image Dataset contains a dataset of nearly around 44000 fashion products with category labels and images. The project aims to create a classification system machine learning model capable of predicting the type of product shown in the image.

The total size of the dataset is 25 GB. The dataset contains one folder named “Images” which contains images of the fashion product. Another folder named “styles” contains the Metadata of each image in JSON format. Finally, the “styles.csv” file maps all the products and contains key product categories.

Scope

The scope of the project would be to (1) Acquire the data from the data source and store it into data lake (2) Perform ETL process and build data warehouse based on star schema (3) Analyse the data using the data warehouse.

Data Source(s)

The data is sourced from https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset.

The total size of the dataset is 25 GB. The dataset contains one folder named “Images” which contains images of the fashion product. Another folder named “styles” contains the Metadata of each image in JSON format. Finally, the “styles.csv” file maps all the products and contains key product categories.

Architecture of the Solution

1. The data would be acquired and stored in a data lake
2. The data from data lake would be then put through ETL process and data warehouse based on Kimball’s approach would be formed. Cloud solutions will be used to scale the database
3. The data warehouse would be then used to build product classifier machine learning model.

Technology

The choice of technology would depend on the stage of project. The initial choice of technologies is:

1) Data Acquisition: Python, API modules
2) Data Exploration: Python, Pandas, SQL
3) ETL: Python, SQL
4) Machine Learning and Analysis: Python, SQL