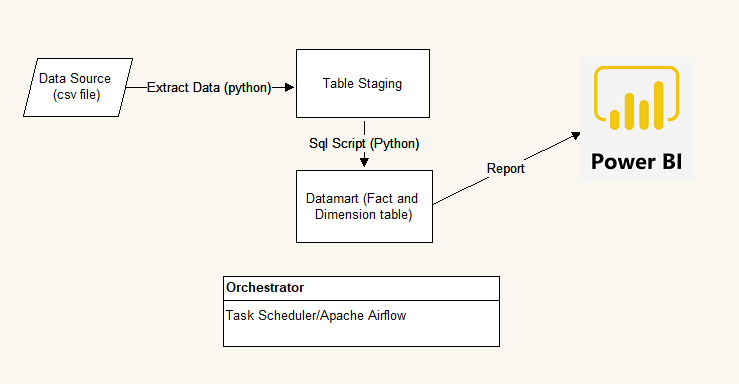
Documentation

1. Background
   1. This project to create a datamart whose data comes from <https://www.kaggle.com/datasets/datascientist97/e-commerece-sales-data-2024>.
   2. ETL process using python script and sqlite as data storage for datamart.
   3. Automate jobs using task scheduler in windows system by running batch file that calls python script or put the python/bat script in apache airflow
   4. Datamart results in sqlite databases are used to build dashboards with Power BI
2. ETL Process
   1. Data collecting from E-commerce Sales Data 2024 dataset on Kaggle. consist of customer\_details.csv,E-commerece sales data 2024.csv, product\_details.csv
   2. Data extraction using python script. This stage is a process
      1. Checking data
      2. Cleaning data
      3. Remove duplicate
      4. Remove unused column
      5. Transform data
   3. Load data into sqlite databases from pandas dataframes, the steps are
      1. Create table for datamart purpose (staging,dimension and fact table).
      2. Insert into table staging for temporary data
      3. Load data from table staging to datamart (dimension and fact table).
3. Datawarehouse Architecture



1. Databases Structure
   1. Customer\_Staging

CREATE TABLE "customer\_staging" (

"Customer ID" INTEGER,

"Age" INTEGER,

"Gender" TEXT,

"Item Purchased" TEXT,

"Category" TEXT,

"Purchase Amount (USD)" INTEGER,

"Location" TEXT,

"Size" TEXT,

"Color" TEXT,

"Season" TEXT,

"Review Rating" REAL,

"Subscription Status" TEXT,

"Shipping Type" TEXT,

"Discount Applied" TEXT,

"Promo Code Used" TEXT,

"Previous Purchases" INTEGER,

"Payment Method" TEXT,

"Frequency of Purchases" TEXT

)

* 1. Sales Staging

CREATE TABLE "sales\_staging" (

"user id" REAL,

"product id" TEXT,

"Interaction type" TEXT,

"Time stamp" TEXT

)

* 1. Product\_Staging

CREATE TABLE "product\_staging" (

"Uniqe Id" TEXT,

"Product Name" TEXT,

"Brand Name" REAL,

"Asin" REAL,

"Category" TEXT,

"Upc Ean Code" TEXT,

"List Price" REAL,

"Selling Price" TEXT,

"Quantity" REAL,

"Model Number" TEXT,

"About Product" TEXT,

"Product Specification" TEXT,

"Technical Details" TEXT,

"Shipping Weight" TEXT,

"Product Dimensions" TEXT,

"Image" TEXT,

"Variants" TEXT,

"Sku" REAL,

"Product Url" TEXT,

"Stock" REAL,

"Product Details" REAL,

"Dimensions" REAL,

"Color" REAL,

"Ingredients" REAL,

"Direction To Use" REAL,

"Is Amazon Seller" TEXT,

"Size Quantity Variant" REAL,

"Product Description" REAL

)

* 1. Customer\_dim

CREATE TABLE customer\_dim (

customer\_id INTEGER PRIMARY KEY UNIQUE,

Age INTEGER,

Gender TEXT,

Item\_Purchased TEXT,

Category TEXT,

Purchase\_Amount\_USD INTEGER,

Location TEXT,

Size TEXT,

Color TEXT,

season TEXT,

Review\_Rating INTEGER,

Subscription\_Status TEXT,

Shipping\_Type TEXT,

Discount\_Applied TEXT,

Promo\_Code\_Used TEXT,

Previous\_Purchases INTEGER,

Payment\_Method TEXT,

Frequency\_of\_Purchases TEXT)

* 1. Product\_dim

CREATE TABLE product\_dim

(Uniqe\_Id TEXT PRIMARY KEY UNIQUE,

Product\_Name TEXT,

Category TEXT,

Upc\_Ean\_Code TEXT,

Selling\_Price NUMERIC,

Model\_Number TEXT,

About\_Product TEXT,

Product\_Specification TEXT,

Technical\_Details TEXT,

Shipping\_Weight TEXT,

Product\_Dimensions TEXT,

Image TEXT,

Variants TEXT,

Product\_Url TEZT,

Is\_Amazon\_Seller TEXT)

* 1. Sales\_fact

CREATE TABLE sales\_fact (

user\_id INTEGER PRIMARY KEY UNIQUE,

product\_id TEXT,

Interaction\_type TEXT,

Time\_stamp TEXT,

datekey TEXT UNIQUE)

* 1. Date\_dim

CREATE TABLE date\_dim (

Datekey TEXT PRIMARY KEY,

Date TEXT,

Year INTEGER,

MonthNumber INTEGER,

Month TEXT,

Days\_In\_Month INTEGER,

Quarter\_Number INTEGER,

Quarter TEXT,

Week\_of\_Year TEXT,

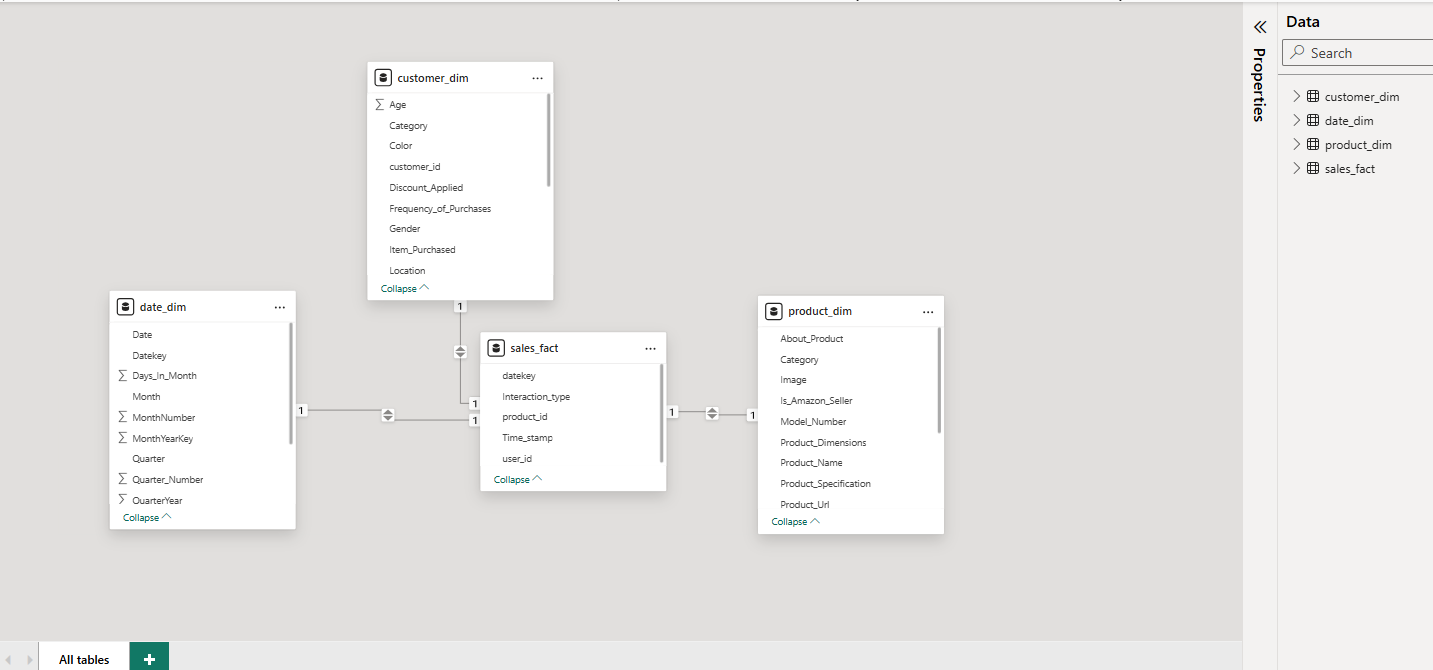
MonthYearKey INTEGER,

WeekNumber INTEGER,

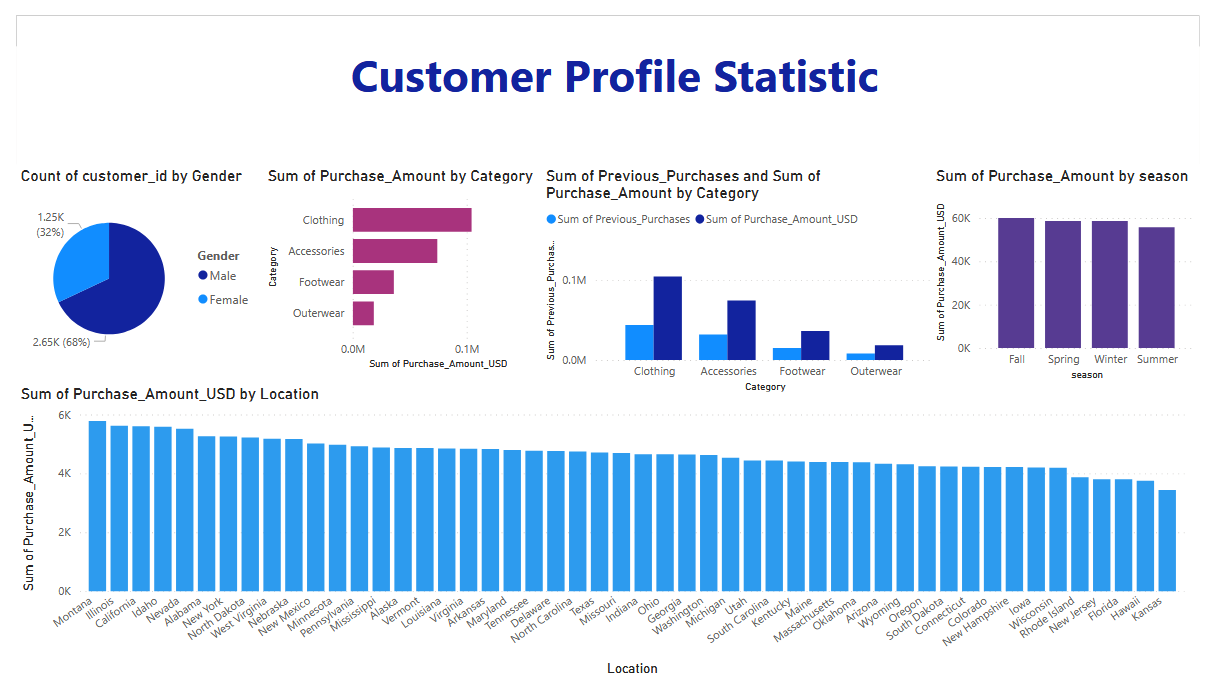
QuarterYear INTEGER,

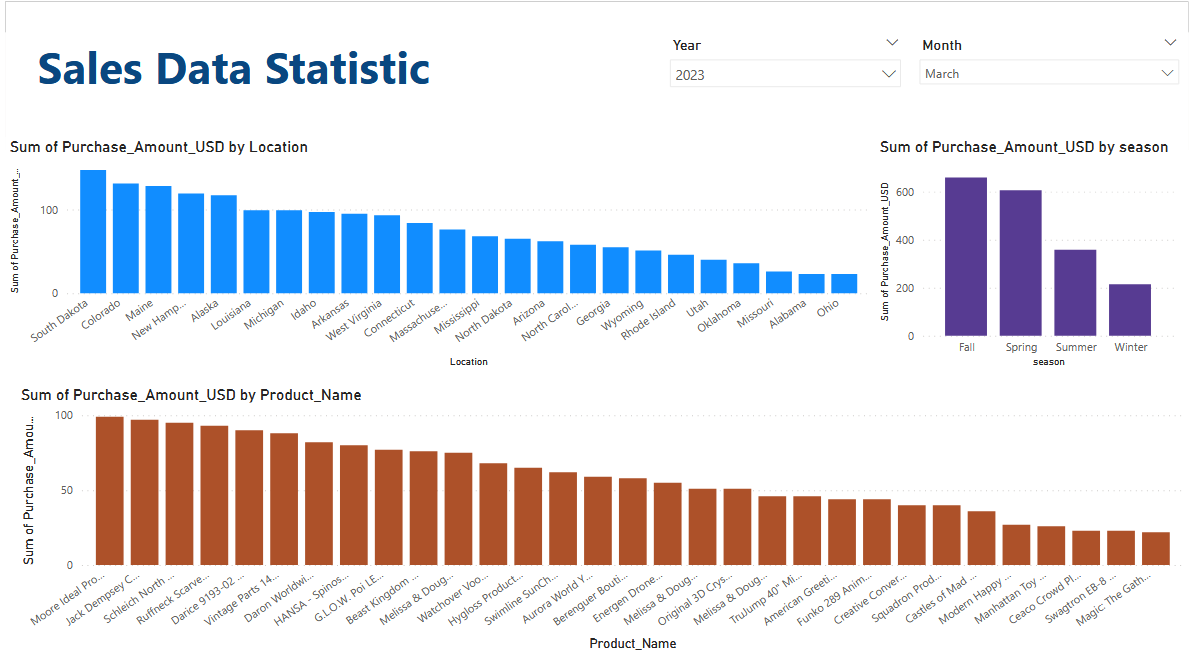
WeekYearKey INTEGER)

1. Datawarehouse Design using star schema model, check picture below

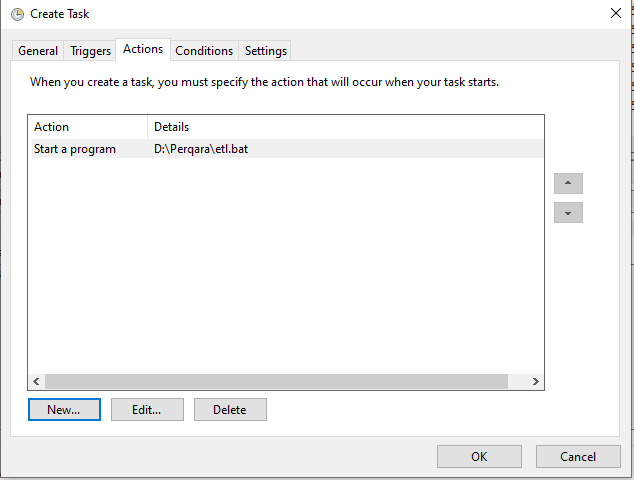


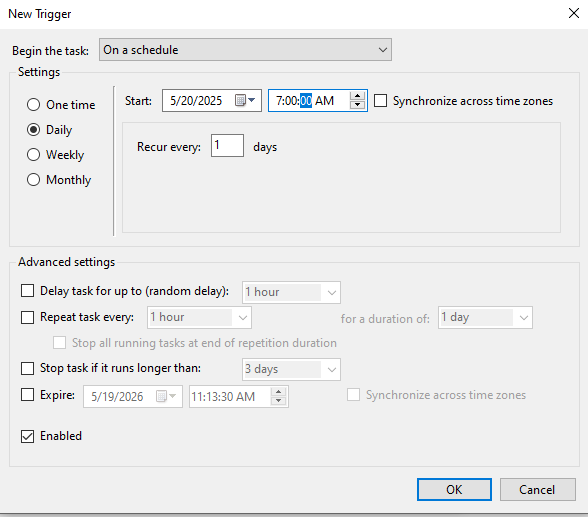
1. Visualization and sample report (Power BI)





1. Job Scheduled. Using task schedule in windows system to run program autamatically (call batch file) or put the python/bat script in apache airflow.





1. Attachment file, find in github repository (https://github.com/firmanrachman/datawarehouse)
   1. data.zip (source data)
   2. script\_databases\_sqlite.zip (sql file script)
   3. exctract\_data.zip (python and batch file script)
   4. sales\_data\_perqara\_report.zip (Power BI report)
   5. Documentation Project (docx file)
2. Installation instruction
   1. Extract source data (data.zip)
   2. Create sqlite database with the name sales\_data.db and then execute sql file to generate table (script\_databases\_sqlite.zip)
   3. Extract python and batch file script then modified code for adjust data source location and sqlite databases (exctract\_data.zip)
   4. Modified connection in power BI report to access your datamart and install odbc diver for sqlite if needed (sales\_data\_perqara\_report.zip).
   5. Modified batch file to adjus script python location
   6. Testing batch file to run program
   7. Create task schedule in windows system to run program autamatically (call batch file) or put the python/bat script in apache airflow.