****

**COLLEGE OF COMPUTING**

**DEPARTMENT OF SOFTWARE ENGINEERING**

**FUNDAMENTAL OF**

**BIG DATA ANALYTICS AND BI**

**INDIVIDUAI ASSIGNMNT**

**NAME ID**

**Abayneh bekele DBU/037/14**

**SUBMITION DATE : 13/02/2025**

**SUBMITED TO : Derbew Felasman (MSc).**

**End-to-End Data Pipeline Documentation**

**Assignment Overview**

Title: Building an End-to-End Data Pipeline for E-Commerce Sales Data

Objective:

This project focuses on extracting, transforming, storing, and visualizing e-commerce sales data using a structured ETL process. The goal is to provide meaningful business insights by processing large-scale sales transactions and storing them efficiently in PostgreSQL, followed by analysis in Power BI.

1, **Technology Stack Used**:

* Step
* Technology Used
* Purpose
* Extract (E)
* Python (Pandas)
* Load CSV sales dataset
* Transform (T)
* Pandas
* Clean and preprocess the data
* Load (L)
* PostgreSQL
* Store structured data for efficient querying
* Visualization
* Power BI
* Create interactive dashboards

2 **,Data Source Identification & Understanding**

* Large Dataset: Sales Transactions Dataset
* Dataset Name: dataset.csv
* Source: Provided CSV file
* Size: Large dataset with multiple transaction records
* File Format: CSV
* Description of Fields
* Column Name
* Data Type
* Description
* order\_id
* STRING

**Unique identifier for the order**

* customer\_id
* STRING
* Unique customer identifier
* product\_id
* Unique product identifier
* product\_name

**Name of the product sold**

* category
* STRING
* Product category
* price
* FLOAT
* Price per unit of the product
* quantity
* INTEGER
* Number of units purchased
* total\_price
* FLOAT

**Total amount paid for the product**

* discount
* FLOAT
* Discount applied to the product
* order\_date
* DATETIME
* Date when the order was placed
* payment\_method
* STRING
* Method used for payment
* country
* STRING
* Customer's country
* his dataset provides a comprehensive view of sales transactions, customer behavior, and product demand trends.

**3,Data Pipeline (ETL Process)**

* Extract Phase
* Loading the Dataset
* The dataset is loaded from dataset.csv using Pandas.
* Code snippet:
* import pandas as pd
* def extract\_data(file\_path):
* data = pd.read\_csv(file\_path, encoding='ISO-8859-1')
* return data
* file\_path = 'data/dataset.csv'
* df = extract\_data(file\_path)
* print(df.head())

**Ensures data is successfully extracted into a Pandas DataFrame.**

* Transform Phase
* Data Cleaning Techniques
* Handling Missing Values: Drop rows with missing customer\_id.
* Formatting Date Columns: Convert order\_date to datetime format.
* Removing Negative/Zero Quantities: Transactions where quantity <= 0 are removed.
* Code Snippet:
* from transform import transform\_data
* def transform\_data(df):
* df = df.dropna(subset=['customer\_id'])
* df['order\_date'] = pd.to\_datetime(df['order\_date'], errors='coerce')
* df = df[df['quantity'] > 0]
* return df

**Ensures data integrity and correct formatting for analysis**.

Load Phase

Database Schema Design

The cleaned data is stored in PostgreSQL using the following schema:

CREATE TABLE sales\_transactions (

order\_id TEXT PRIMARY KEY,

customer\_id TEXT,

product\_id TEXT,

product\_name TEXT,

category TEXT,

price NUMERIC,

quantity INT,

total\_price NUMERIC,

discount NUMERIC,

order\_date TIMESTAMP,

payment\_method TEXT,

country TEXT

);

Optimized for efficient querying in PostgreSQL.

Storing Data in PostgreSQL

Data is inserted into the database using Python and psycopg2.

from load import insert\_data

insert\_data(transformed\_df)

Ensures all sales transactions are stored securely for further analysis.

4, **Data Visualization and Insights**

* Connecting Power BI to PostgreSQL
* Enable PostgreSQL connection in Power BI.
* Use PostgreSQL ODBC driver to import the dataset.
* Load the sales\_transactions table for visualization.

Key Visualizations in Power BI

* Visualization
* Chart Type
* Objective
* Sales Trends Over Time
* Line Chart
* Track total sales across months/years

Top-Selling Categories

1. Bar Chart
2. Identify the most popular product categories

Order Status Distribution

1. Pie Chart
2. Compare sales trends by category
3. Average Order Value by Category
4. Column Chart
5. Analyze which categories generate higher revenue
6. Customer Locations
7. Map Visualization
8. See where the most purchases occur geographically
9. Power BI helps uncover key sales insights through interactive reports.

**5 Key Business Insights**

* Trends Found in the Data
* Top Categories: Electronics and Apparel account for the highest sales volume.
* Payment Trends: Most purchases are made using credit cards.
* Discount Impact: Products with higher discounts tend to have increased sales.
* Potential Business Decisions
* Focus on high-performing product categories.
* Optimize pricing strategies based on discount impact analysis.
* Improve marketing efforts in top-performing customer regions.

6**, Documentation & Code Quality**

* Ensuring High-Quality Code:
* Well-structured scripts for extraction, transformation, and loading.
* Error handling for missing data, duplicates, and incorrect data types.
* Efficient queries for fast execution.

7**, Conclusion & Future Enhancements**

1. Summary of the ETL Pipeline
2. Extracted data using Pandas.
3. Cleaned and enriched data with transformations.
4. Stored data in PostgreSQL.
5. Visualized insights in Power BI dashboards.

**Future Enhancements**

Real-time ingestion with Apache Kafka for live sales tracking.

Machine Learning Model to predict future sales trends.