Case Study– SQL, Python, Numpy and Pandas – (GitHub)

(A Brief Introduction)

Dr. Ehibhahiemen Nelson Ughele

10Alytics

What is Numpy?

- NumPy is a powerful Python library used for numerical computing. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently. NumPy is widely used in various fields such as data science, machine learning, and scientific computing due to its speed, versatility, and ease of use.

How to install and import into your Python notebook?

- You can install NumPy using pip, the Python package manager. Simply run the following command in your terminal or command prompt:

```
pip install numpy
```

- Once installed, you can import NumPy into your Python notebook using the following import statement:

```
import numpy as np
```

Basic operations of Numpy for data preprocessing

- Some basic operations in NumPy for data preprocessing include creating arrays, accessing elements within arrays, performing mathematical operations on arrays, and handling missing values.

- NumPy provides functions for array manipulation, such as reshaping, slicing, concatenation, and more, making it essential for data preprocessing tasks.

Case Study For Yanki Ecommerce

EXECUTIVE SUMMARY

Yanki Ecommerce is an online store specializing in the sale of a diverse range of perfumes and perfume-related products. The company aims to provide customers with high-quality products from various brands while delivering an exceptional shopping experience. To optimize its operations and enhance customer satisfaction, Yanki Ecommerce seeks to leverage data-driven insights through advanced data engineering techniques.
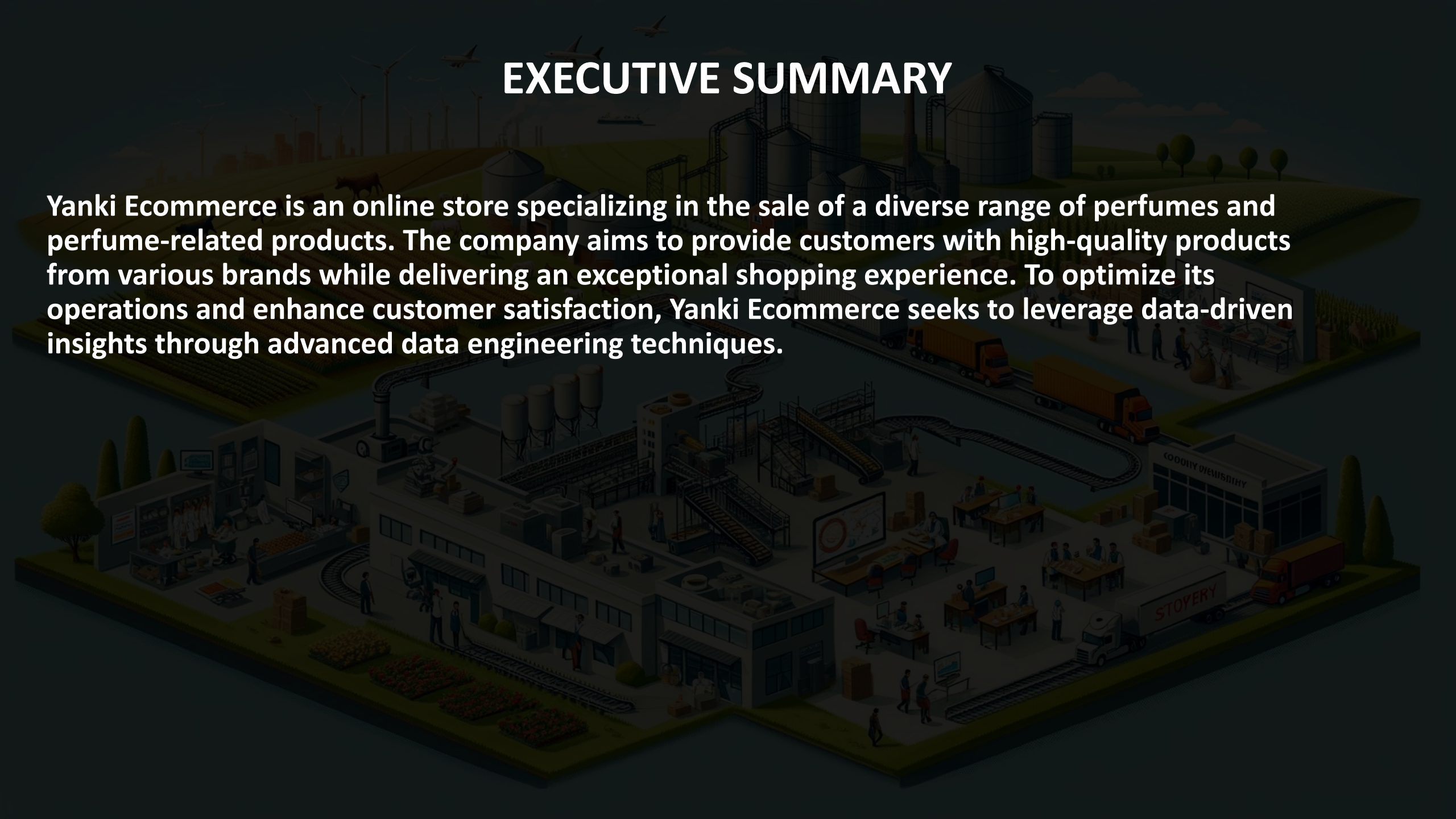
BUSINESS PROBLEM STATEMENT

Yanki Ecommerce faces challenges in effectively managing and analyzing its vast amounts of data, including historical sales data, customer information, and inventory data. The company requires efficient data engineering solutions to streamline data processing, perform comprehensive analysis, and derive actionable insights to improve decision-making processes and enhance business performance.
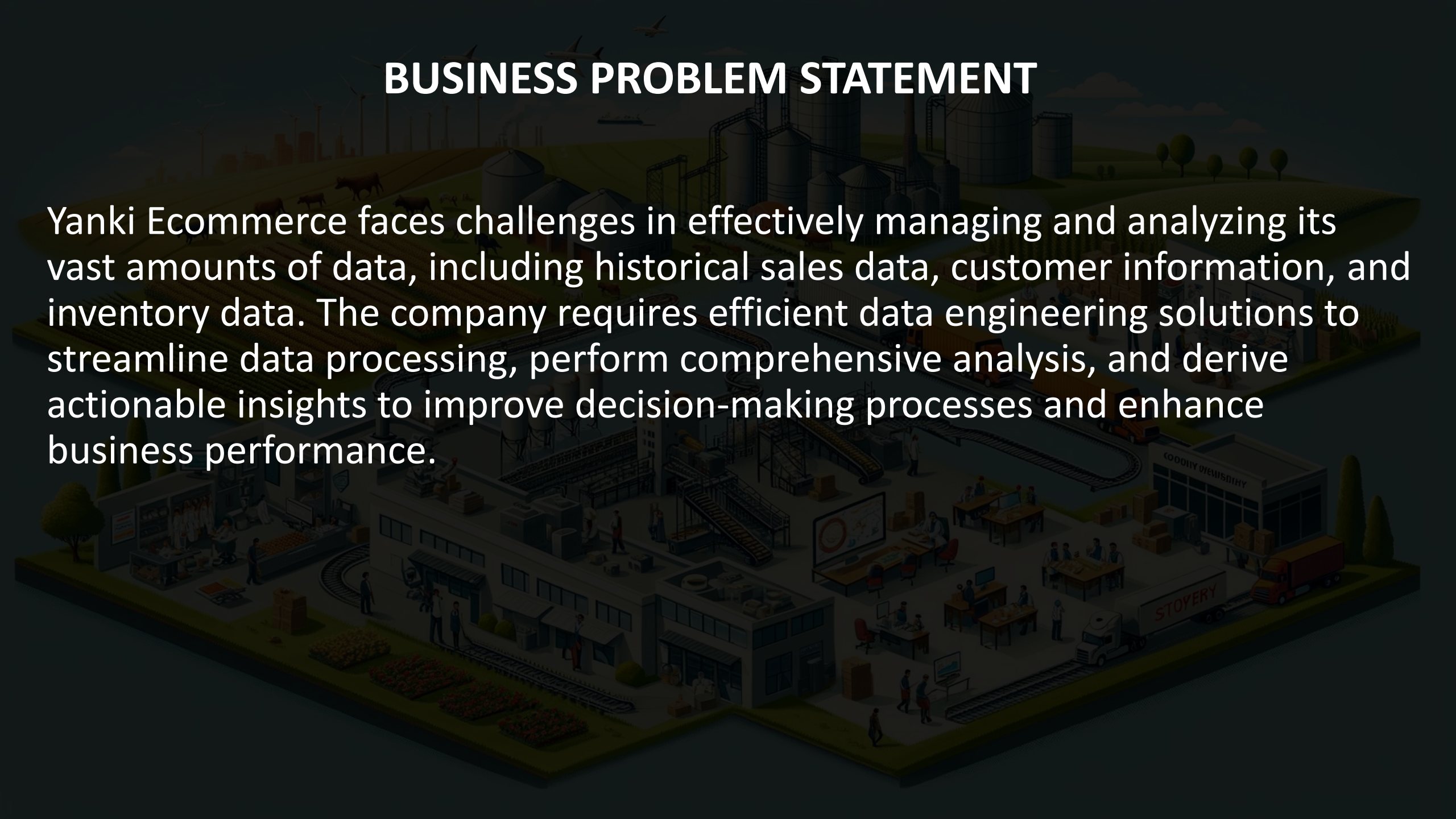
Objectives

1. Extract historical sales data from CSV files and other relevant sources.

2. Clean and preprocess the extracted data to ensure consistency, accuracy, and completeness.

3. Transform the data using NumPy and Pandas to prepare it for analysis and modeling.

4. Load the processed data into a PostgreSQL database for storage and easy access.

5. Implement version control using GitHub Desktop to manage codebase changes and collaborate effectively.

Benefits

1. Improved efficiency in data processing and analysis.

2. Enhanced data quality and integrity.

3. Streamlined data loading and storage processes.

4. Facilitated collaboration and version control.

5. Empowered decision-making through data-driven insights.
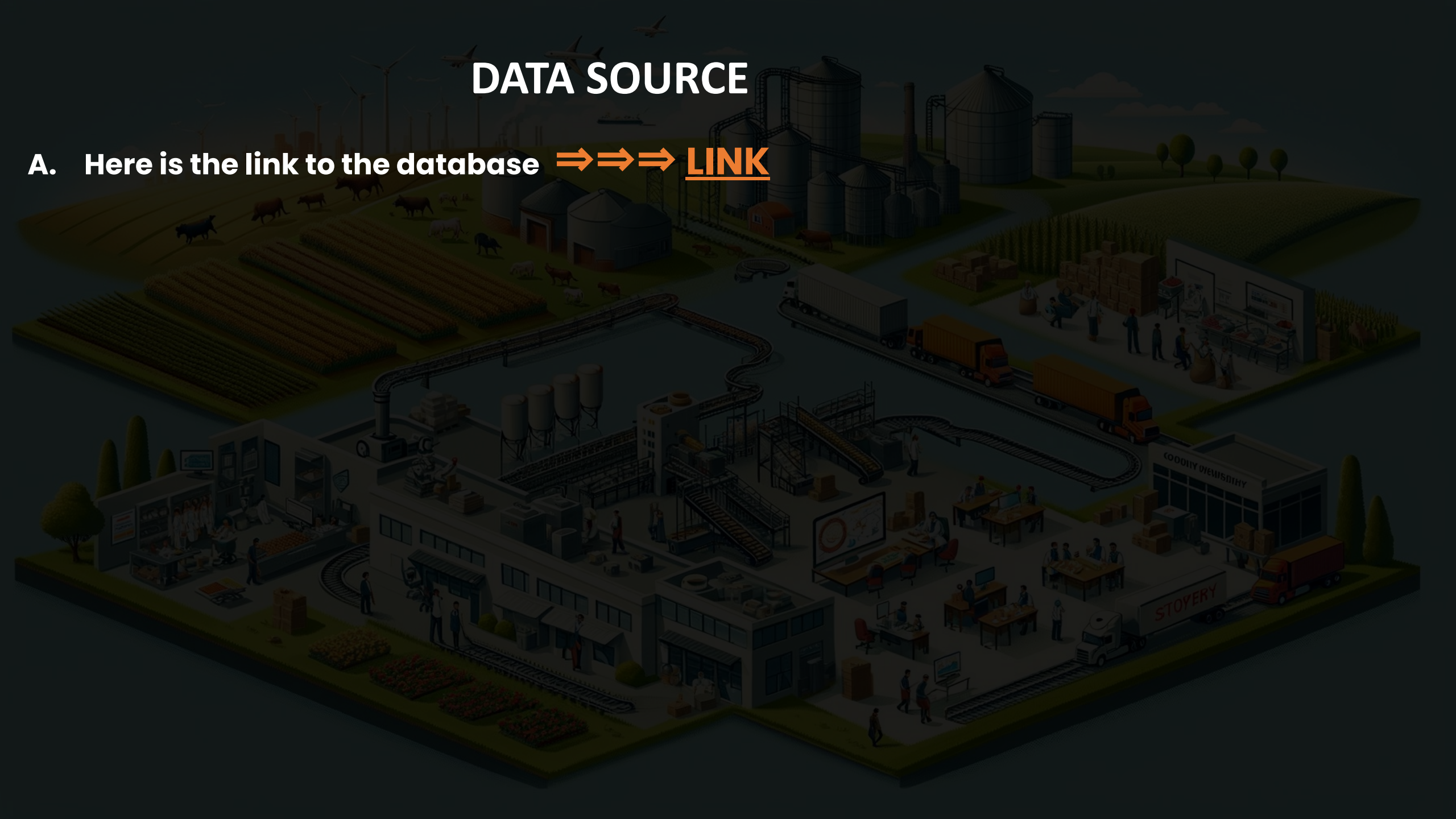
TECH STACK

A. **Python**

B. **NumPy**

C. **Pandas**

D. **GitHub (using GitHub Desktop)**

E. **PostgreSQL**

DATA SOURCE

A. Here is the link to the database ⇒⇒⇒ **LINK**

PROJECT SCOPE

A. Data Extraction from Historical CSV Data
   a. Utilize Python libraries such as Pandas for extracting data from CSV files.
   b. Implement efficient data extraction methods to handle large volumes of historical sales data.

B. Data Cleaning and Transformation (using NumPy and Pandas)
   a. Utilize NumPy and Pandas for data cleaning tasks such as handling missing values, removing duplicates, and standardizing data formats.
   b. Perform data transformation operations such as normalization, encoding, and feature engineering to prepare the data for analysis.

C. Data Loading to a PostgreSQL Database
   a. Use PostgreSQL as the database management system for storing the processed data.
   b. Design and implement database schemas to accommodate the data structure and relationships.

D. Version Control using GitHub Desktop
   a. Set up a GitHub repository for the project to enable version control and collaboration.
   b. Use GitHub Desktop for managing code changes, branching, and merging workflows.

By successfully executing these tasks, the data engineering team at Yanki Ecommerce can effectively leverage Python, NumPy, Pandas, GitHub, and PostgreSQL to optimize data processing, analysis, and decision-making processes, ultimately driving business growth and success.

Now We Proceed To Coding!!!

Happy Coding!!!

GOODLUCK