| **Project Title:** | **Dmart analysis using pyspark** |
| --- | --- |
| **Technologies** | **Python, SQL , Pyspark** |

**Dataset :**[**Dmart**](https://drive.google.com/drive/folders/1Y7Q27S00milOYuPnO7NHLl3sQGd49Ng6?usp=sharing)

**Problem Statement :**

The task is to create a data pipeline using PySpark to integrate and analyze sales data from three different sources: product information, sales transactions, and customer details. The goal is to establish a connection with PySpark, load the datasets, perform data transformations, and answer a set of analytical questions.

**Tasks:**

Task 1: Establish PySpark Connection

* Set up a PySpark environment.
* Create a connection to read CSV files into PySpark DataFrames.

Task 2: Load Data into PySpark DataFrames

* Load the products.csv, sales.csv, and customer.csv files into separate PySpark DataFrames.

Task 3: Data Transformation and Cleaning

* + Perform necessary data cleaning and transformation:
  + Rename columns for consistency if needed.
  + Handle missing values appropriately.
  + Ensure data types are correctly set for each column.
  + Join the DataFrames on relevant keys (Product ID and Customer ID).

#### Task 4: Data Analysis and Querying

* Formulate 10 analytical questions based on the integrated dataset.
* Write PySpark code to answer these questions.

### Task 5 : Run Queries on the Pyspark

1. What is the total sales for each product category?
2. Which customer has made the highest number of purchases?
3. What is the average discount given on sales across all products?
4. How many unique products were sold in each region?
5. What is the total profit generated in each state?
6. Which product sub-category has the highest sales?
7. What is the average age of customers in each segment?
8. How many orders were shipped in each shipping mode?
9. What is the total quantity of products sold in each city?
10. Which customer segment has the highest profit margin?

**Submission:**

* Provide a well-commented Python file containing the complete code for the project, organized into sections for data Pipeline and Analysis.
* Upload the same into github with a proper Readme file.
* Presentation on the entire project, including Problem Statement, Tools Used, Approaches and Insights Found.

**Evaluation Metrics:**

* Project evaluation will be done in the live session and have to showcase the approaches done to complete the project
* You are supposed to write a code in a modular fashion (in functional blocks)
* Maintainable: It can be maintained, even as your codebase grows.
* Portable: It works the same in every environment (operating system)
* You have to maintain your code on GitHub.(Mandatory)
* You have to keep your GitHub repo public so that anyone can check yourcode.(Mandatory)
* Proper readme file you have to maintain for any project development(Mandatory)
* Follow the coding standards:
  + https://www.python.org/dev/peps/pep-0008/
* You should include basic workflow and execution of the entire project in the readme file on GitHub

**GitHub Repo:**

The attached reference document will help you use GitHub effectively. - [Link](https://docs.google.com/presentation/d/1XHCbgUOqbcXNUyQ87vTlKdKRgAbBxtkA/edit?usp=sharing&ouid=106590842700357786537&rtpof=true&sd=true)

**Reference Material:**

Official Documentation:

* <https://www.python.org/doc/>
* <https://docs.streamlit.io/>
* <https://dev.mysql.com/doc/>

FAQs : [FAQs for Dmart Analysis using PySpark](https://docs.google.com/document/d/1t3PV5EaLyoOa4IYe32anso9KIPmS9FFA6VqyrLzhtK8/edit?usp=sharing)