Capstone Project: Retail Sales Performance Dashboard

Objective:
Build a system that collects and analyzes retail sales data across multiple stores and products. The final output should help monitor store performance and identify underperforming products.

Week 1 – Database Foundations: MySQL & MongoDB

Tools: MySQL, MongoDB

Capstone Tasks:
- Create MySQL tables for `products`, `stores`, `sales`, and `employees`
- Insert sales data with CRUD operations
- Write a stored procedure to calculate daily sales for a store
- Store promotional campaign feedback in MongoDB
- Add indexes to search by product and region

Deliverables:
- SQL script with schema, CRUD, and stored procedure
- MongoDB script for campaign data and indexing

Week 2 – Data Collection & Cleanup Using Python

Tools: Python (Pandas, NumPy, Requests)

Capstone Tasks:
- Load sales and product data from CSV or API
- Use `pandas` to clean missing values, correct data types
- Use `numpy` to calculate revenue, discount percentages, and profit margins
- Show summary of total revenue by product and store

Sample Code Snippet:

```python
import pandas as pd import numpy as np

df = pd.read_csv('sales.csv') df['revenue'] = df['quantity'] * df['price']
df['profit'] = df['revenue'] - df['cost']

summary = df.groupby('store_id')[['revenue', 'profit']].sum() print(summary)
```

Deliverables:
- Cleaned dataset with calculated fields
- Python script summarizing key metrics

Week 3 – PySpark for Store-Level Insights

Tools: PySpark

Capstone Tasks:
- Load large sales data into PySpark
- Filter data for underperforming products (e.g., low sales, high returns)
- Group by store and calculate average monthly revenue

Deliverables:
- PySpark script with filtering, grouping, and aggregation
- Output file showing underperforming products/store summary

Week 4 – ETL Pipeline in Azure Databricks

Tools: Azure Databricks

Capstone Tasks:
- Upload cleaned data to Databricks
- Transform and join product + sales data
- Save final metrics (e.g., profit margin by category) in Delta or CSV
- Use a Databricks SQL cell to find top 3 best-selling products

Deliverables:
- Databricks notebook with ETL logic
- Saved output table/file for dashboard use

Week 5 – Pipeline Automation with Azure DevOps

Tools: Azure DevOps

Capstone Tasks:
- Create a pipeline that runs the full analysis weekly
- Output results to a CSV or log file
- Add a step to email or log top 5 lowest performing stores

Deliverables:
- YAML pipeline file
- Output file showing key sales insights

Final Outcome by Week 5:

- Structured data storage for retail products, sales, and store performance
- Python script to analyze and clean raw sales data
- PySpark transformation for store-wise insights
- ETL job in Databricks producing clean metrics
- Azure DevOps pipeline automating reporting