

Comprehensive Retail Sales Analysis.

Case Study

Problem Statement:

In the realm of retail, making informed decisions based on historical data is crucial. The challenge arises when strategic choices are influenced by limited historical records, especially for events that occur annually, such as holidays and major promotions. This challenge is exacerbated by the fact that significant events, like holidays and markdowns, have a considerable impact on sales, making it necessary to anticipate the effects on specific departments and overall performance.

This project focuses on leveraging historical sales data to engineer a data-driven solution for the company's decision-making processes. The dataset encompasses sales data from 45 stores distributed across various regions, each housing multiple departments. The company's operational strategy includes several promotional markdown events, each aligned with key holidays like the Super Bowl, Labour Day, Thanksgiving, and Christmas. Worth noting is the fact that week's containing these holidays hold five times the significance of non-holiday weeks during evaluations.

What is Expected?

The objective of this data engineering project is to construct a robust data pipeline that effectively integrates, cleanses, and transforms the provided data. The pipeline should be designed to prepare the data for subsequent analysis, focusing specifically on quantifying the impact of holidays and markdowns on sales performance across different departments and stores. By engineering this data pipeline, the project aims to facilitate comprehensive decision-making based on historical trends, enabling the company to better anticipate and manage the effects of promotional events and holidays on its bottom line.

For the data engineering solution, a comparative study of several approaches will be performed to build an appropriate system for campaign classification. The first step in this process will involve data pre-processing and cleaning to address data quality issues identified during the exploratory analysis.

1. Holiday Impact Assessment:

- a. How do sales figures during holiday weeks compare to non-holiday weeks?
- b. Which departments experience the most significant sales spikes during holiday periods?

2. Markdown Analysis:

- a. To what extent do markdown events influence overall sales and departmental performance?
- b. Can we identify specific departments that exhibit a more pronounced response to markdowns?

3. Store Performance Comparison:

- a. How does the sales performance vary across different stores during holiday and non-holiday weeks?
- b. Are there specific stores that consistently outperform others during significant events?

4. Departmental Sales Patterns:

- a. What are the typical sales patterns for each department during holiday and non-holiday weeks?
- b. Are there departments that consistently show strong sales regardless of the time of year?

5. Long-Term Impact of Promotions:

- a. Do promotional events like markdowns have a lasting impact on sales beyond the event period?
- b. Can we identify trends indicating sustained customer interest after promotional periods?

6. Region-Specific Analysis:

- a. How do regional variations impact the effects of holidays and promotions on sales?
- b. Are there regions where holiday impact is more pronounced, and if so, why?

Comprehensive Retail Sales Analysis.

Case Study

Data Dictionary:

<https://github.com/manojkumarsingh77/Shell2023/tree/main/RetailSales/DataDictionary>

Data Sets:

<https://github.com/manojkumarsingh77/Shell2023/tree/main/RetailSales/DataSets>

Case Study Execution Plan:

- The execution of each Case Study will involve a group of 4 or 5 members, with each member assigned specific tasks to align with the project's objectives.
- Each group member will work concurrently on their designated tasks, ensuring parallel progress, and the integration of individual contributions will occur during the Final Stage of the project.
- On the Final day, the completed Case Study will be presented to the Shell Subject Matter Experts (SME) and UNext Mentors, providing an opportunity to showcase the project's outcomes and achievements.
- The entire project development process will be implemented using a Continuous Integration/Continuous Deployment (CI/CD) pipeline. This approach ensures seamless integration of code changes, automated testing, and efficient deployment, promoting collaboration and efficiency throughout the project lifecycle.

Technicalities:

In order to address the given problem statement, we will adhere to a standard data pipeline pattern. This structured approach will ensure a systematic and efficient workflow for data processing and transformation.

The data pipeline will consist of the following key stages:

- Data Ingestion.
- Data Processing.
- Data Storage.
- Data Visualization and Reporting.

Data Layers:

As part of a structured data storage approach, you will implement measures to ensure efficient data organization and management. The data will be divided into separate parent folders, one for each team, with sub-folders for **RAW**, **STG (Staging)**, and **CURATED** data:

Parent Folders: Each team involved in the project will have its dedicated parent folder to manage their data processing activities. This ensures data isolation and promotes collaboration within the team.

RAW Sub-folder: The RAW sub-folder within each team's parent folder will be used to store the raw and unprocessed data acquired from various sources. This includes the data ingested through Azure Data Factory or any other data ingestion mechanism.

STG (Staging) Sub-folder: The STG sub-folder will serve as an intermediate storage location where data from the RAW sub-folder is transformed and prepared for further processing. This staging step ensures data quality and consistency before moving it to the CURATED sub-folder.

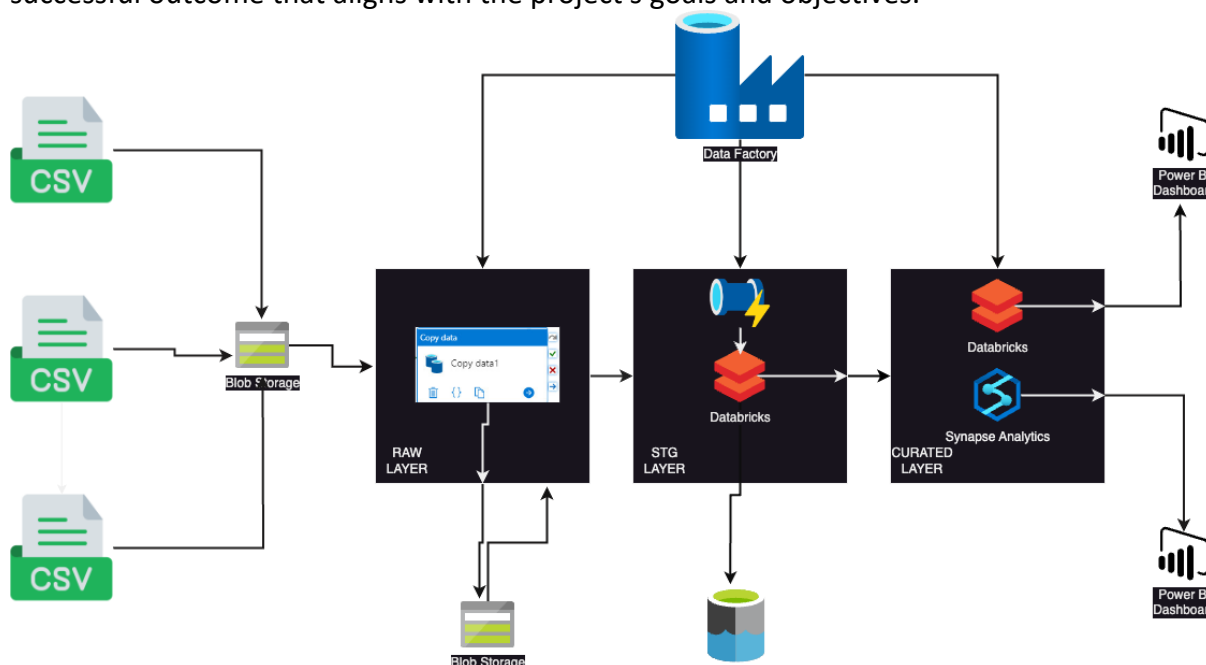
CURATED Sub-folder: The CURATED sub-folder will hold the processed and curated data ready for visualization and analysis. This data is transformed, cleansed, and enriched to meet specific business requirements.

Comprehensive Retail Sales Analysis.

Case Study

Reference architecture diagram:

The provided architecture diagram serves as a foundational reference for each team to envision their own version while building upon it. This diagram presents a clear and structured visualization of the system's components and their interactions. Each team is tasked with developing their own iteration of the architecture diagram, using the provided sample as a foundation. This approach fosters creativity and empowers teams to tailor the solution to meet specific requirements and address unique challenges. By building upon the initial reference, teams can explore diverse design choices and leverage individual expertise, resulting in a comprehensive and adaptable solution. This collaborative process ensures a successful outcome that aligns with the project's goals and objectives.



Activity Breakdown:

In the case study, data engineers will perform data ingestion and cleansing activities to ensure data quality and integrity. They will create a reusable and secured connection for data ingestion and handle tasks like removing duplicate records, handling missing values through imputation techniques, and correcting data anomalies.

For ETL and analysis, data engineers will filter out irrelevant or incomplete data, aggregate data to calculate summary statistics, transform data types and create derived columns, perform data joining based on common keys, and apply data partitioning for improved query performance. They will also conduct data deduplication and implement validation checks to ensure data quality and adherence to business rules.

The Case Study is divided into two parallel streams, each handled by separate teams:

- i. **Stream 1:** This stream utilizes SQL Data Warehouse/Database (SQL DW/DB) as the data storage and management solution. The team in charge of this stream will leverage the capabilities of Power BI for data visualization and creating interactive dashboards. The combination of SQL DW/DB and Power BI ensures efficient data processing, storage, and analysis, providing stakeholders with valuable insights to support data-driven decision-making.
- ii. **Stream 2:** In this stream, the team will employ Azure Databricks with SQL End-point (ADB SQL End-point) as the data processing and analysis platform. Power BI will be used for data visualization and interactive dashboard creation. By leveraging the distributed data processing capabilities of Azure Databricks and combining it with Power BI's visualization

Comprehensive Retail Sales Analysis. Case Study

capabilities, this stream enables efficient and scalable data processing, ensuring stakeholders have access to timely and insightful information.

By splitting the case study into these two streams, the project benefits from parallel efforts, maximizing efficiency and expertise in both SQL-based and Databricks-based data processing approaches. This approach allows for a comprehensive exploration of different technologies, resulting in a well-rounded and robust solution for meeting the specified data processing and visualization requirements.

Deliverables:

Create a presentation which has:

Slide 1: BatchName_FirstName_SecondName

Slide 2: Problem statement

Slide 3: Implemented data flow diagram showing various technical components and Layers.

Slide 4-6: Snapshots of developments in each layer (RAW, STAGING(STG), CURATED)

Slide 7: Screenshot of dashboards built on Power BI.

Slide 8: GitHub link where solution is available

Slide 9: System Demo

Slide 10: Q&A

Slide 11: Challenges faced, learnings, suggestions, and feedback.

Rubrics for Case Study Evaluation:

Deliverables / milestones	Remarks	Max Marks
<ul style="list-style-type: none"> GitHub account creation (5 Marks) Proposing your own Architecture design and details (15Marks) 	Activities	20
<ul style="list-style-type: none"> Data Management and Storage (10 marks) 	Activities	10
<ul style="list-style-type: none"> Data ingestion and Transformation technique details (20 marks) 	Activities	20
<ul style="list-style-type: none"> Visualization of data, by keeping scope of Business User (10 Marks) Story telling by visualizing data (10 marks) 	Activities	20
<ul style="list-style-type: none"> Live presentation of Solution on Azure portal (15 marks) Viva (15 marks) 	Activities	30
	Total Marks	100