Alternative Fuel Consumption Evaluation Analysis.
Case Study

Problem Statement:

The objective of this assignment is to develop a data engineering application using Azure cloud services and the Spark framework to analyze the usage of alternate fuels and the factors influencing customers' fuel choices. The dataset used for this project is published by the US and Canada Transportation and consists of millions of observations and over 30 variables. The focus of the analysis is on low carbon fuels (LCF), such as sustainable aviation fuel (SAF), biodiesel, bioethanol, and renewable compressed natural gas (R-CNG), which aim to reduce carbon emissions from transportation.

What is Expected?

As a Data Engineer, the initial step in the exploratory analysis involves examining the dataset to extract meaningful insights and identify any issues that require further investigation. The output of the exploratory analysis will consist of a comprehensive report detailing the dataset's key characteristics, such as data distribution, missing values, and outliers. Additionally, the report will highlight any potential data quality problems that need follow-up, such as inconsistent or erroneous data entries. Basic descriptive analysis will be conducted to showcase essential outcomes and findings from the data, including trends, patterns, and correlations between different variables. These insights will be used to inform the subsequent data engineering solution for classifying successful and unsuccessful campaigns.

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. State with the most alternate fuel stations.
2. Most used Alternative fuel-based vehicle.
3. Number of hybrid electric vehicles (HEVs), by model, sold in the United States between 1999 and 2019.
4. Economy based on Drivetrain, Category and Car manufacturer.
5. Economy (highway, city) according to fuel type.
6. Engine performance based on engine type, size, and description.
7. Trend of buying biodiesel state wise.

Data Dictionary:

https://github.com/manojkumarsingh77/Shell2023/tree/main/AlternativeFuel/DataDictionary

Data Sets:

https://github.com/manojkumarsingh77/Shell2023/tree/main/AlternativeFuel/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.

Alternative Fuel Consumption Evaluation Analysis.
Case Study

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.

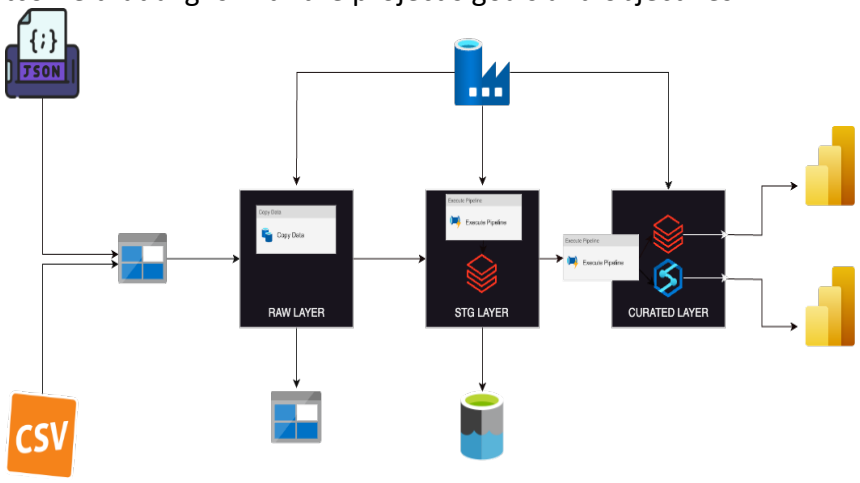
RAW Sub-folder: The RAW sub-folder within each team's parent folder will be used to store the raw and unprocessed data acquired from various sources. This includes the data ingested through Azure Data Factory or any other data ingestion mechanism.

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.

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.



Alternative Fuel Consumption Evaluation Analysis.
Case Study

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 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
▪ GitHub account creation (5 Marks) ▪ Proposing your own Architecture design and details (15Marks)	Activities	20
▪ Data Management and Storage (10 marks)	Activities	10
▪ Data ingestion and Transformation technique details (20 marks)	Activities	20
▪ Visualization of data, by keeping scope of Business User (10 Marks) ▪ Story telling by visualizing data (10 marks)	Activities	20
▪ Live presentation of Solution on Azure portal (15 marks)	Activities	30

Alternative Fuel Consumption Evaluation Analysis.
Case Study

▪ Viva (15 marks)		
	Total Marks	**100**