**Case Study: Building a Sales Analytics Pipeline in Azure**

**Objective:**

Design and implement a fully automated solution that integrates, processes, and analyzes sales data from multiple sources (CSV, JSON) using Azure Data Factory (ADF), Azure Data Lake, Azure Data Factory and Synapse.

**Business Scenario:**

A retail company collects sales, product, and customer data. The sales data comes from transactional systems as CSV files, customer data is available in JSON format from an API, and product details are maintained in CSV files. All data is stored in a GitHub repository.

The goal is to build a centralized and automated data pipeline that:

1. **Consolidates data:** Brings data into a central repository (Azure Data Lake).
2. **Processes and cleanses data:** Ensures high-quality, integrated datasets.
3. **Provides analytics:** Generates insights such as average sales trends, product performance, and rank products based on sales.

**Pipeline Design Overview**

**1. Data Sources**

**Source Files:**

1. **Sales Data (CSV)**  
   Contains details of each transaction.
   * **Fields:** SaleID, ProductID, CustomerID, SalesAmount, Quantity, Timestamp
2. **Customer Data (JSON)**  
   Provides demographic and regional data about customers.
   * **Fields:** CustomerID, FirstName, LastName, Gender, Region, SSN
3. **Product Data (CSV)**  
   Describes products sold.
   * **Fields:** ProductID, ProductName, Category

**Storage Location:**  
All files are hosted in a GitHub repository, accessible via an HTTP URL.

* [DeebaLakshmiCT/Sales-for-ADF](https://github.com/DeebaLakshmiCT/Sales-for-ADF/tree/main)
* [DeebaLakshmiCT/Customers-for-ADF](https://github.com/DeebaLakshmiCT/Customers-for-ADF)
* [DeebaLakshmiCT/products-for-adf](https://github.com/DeebaLakshmiCT/products-for-adf)

**2. Ingestion Layer: Azure Data Factory**

**Workflow:**

1. **Create ADF Pipelines:**
   * **Step 1:** Define HTTP-linked services to connect to the GitHub repository.
   * **Step 2:** Use copy activities to download files into Azure Data Lake Gen2 (ADLS).
   * **Destination:** Raw zone in ADLS with a structured folder hierarchy:
     + /raw/sales/
     + /raw/customers/
     + /raw/products/
2. **Trigger Automation:**
   * Schedule the ADF pipeline to run on the 25th of every month using time-based triggers.

**3. Transformation Layer: Azure Data Factory**

**Process:**

1. **Data Cleansing:**

**Customer Data**

* + Change the Schema of Customer file – Cus\_ID, Name, Gender, Country, SSN\_ID
  + Replace male /female with 0/1
  + Nomalize SSN format

**Sales Data**

* + Provide time stamp in DD-MM\_YYYY format
  + Work with nulls on sales amount and ensure they’re non-negative

**Products Data**

* + Validate if products table is non empty

.

1. **Data Integration:**
   * **Join Datasets:**
     + Merge sales data with customer data on CustomerID.
     + Merge the result with product data on ProductID.
2. **Data Transformation:**
   * Extract Year and Month from the Timestamp field.
   * Calculate metrics

AverageSalesPerProduct

Create Rank on sales partitioning based on Product and categories

1. **Store as table in Dedicated SQL :**
   * Save the rank table and average sales calculated table.