**Title: Building end to end analytical solution for Sales Data in Microsoft Fabric**

**Objective:**

Design and implement a fully automated end to end analytical solution that integrates, processes, and analyses sales data from multiple sources (CSV, JSON) using Microsoft fabric. The system will generate actionable insights by putting together visualizations like charts and graphs, create an interactive report, and present them to the decision-makers in the organization.

**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.

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

1. **Consolidates data:** Brings data into a central repository into OneLake.
2. **Processes and cleanses data:** Ensures high-quality, integrated datasets.
3. **Provides analytics:** Generates insights such as regional sales trends, product performance, and monthly sales summaries.
4. **Automates updates:** Incorporates new data monthly without manual intervention.

**Pipeline Design Overview**

**1. Data Sources**

**Source Files:**

1. **Sales Data (CSV):**

This data comes from the ERP system and is stored in a Microsoft Azure Blob Storage.

Contains details of each transaction.

* + **Fields:** SaleID, ProductID, CustomerID, SalesAmount, Quantity, Timestamp

1. **Customer Data (JSON):**

This data comes from Customer Insights and is stored in Dataverse which is downloaded in JSON format and available in local folder. The data is always upto date, and it provides demographic and regional data about customers.

* + **Fields:** CustomerID, FirstName, LastName, Gender, Region, SSN

1. **Product Data (CSV)**  
   Describes products sold. it is stored as an export file in a SharePoint folder
   * **Fields:** ProductID, ProductName, Category

**2. Ingestion Layer:**

**Workflow:**

* + **Step 1:** Create shortcut to connect to the Azure Data Lake Gen2 (ADLS) Blog Storage.
  + **Step 2:** Use Dataflow Gen2 in pipeline activities to get data from SharePoint.
  + **Step 3:** upload customer.csv file into File Folder
  + **Destination:** Raw zone in Lakehouse with a structured folder hierarchy:
    - /raw/sales/
    - /raw/customers/
    - /raw/products/

**3. Storage Layer:**

* **Raw Data:** Unprocessed data ingested from different sources.
* **Processed Data:** Cleaned and transformed data stored in Delta format.
* **Output Data:** Aggregated metrics for visualization and analytics.

**4. Transformation Layer:**

**Process:**

1. **Read Raw Data:**
   * Load CSV and JSON files from the raw zone in Lakehouse.
2. **Data Cleansing:**
   * Handle missing or null values.
   * Normalize inconsistent formats
   * Validate numerical fields (e.g., ensure non-negative sales amounts).
3. **Data Integration:**
   * **Join Datasets:**
     + Merge sales data with customer data on CustomerID.
     + Merge the result with product data on ProductID.
4. **Data Transformation:**
   * Extract Year and Month from the Timestamp field.
   * Calculate additional metrics if required (e.g., AverageSalesPerProduct).
5. **Store in Delta Lake:**
   * Save the cleaned and transformed data in Delta tables.

**5. Aggregated Metrics**

**Metrics:**

* Calculate total sales and quantity grouped by:
  + Region.
  + Product Category.
  + Year and Month.

**Output:**

Save the aggregated metrics as Delta tables

**6. Visualization Aggregated and Reporting:**

**Visualization Outputs:**

1. Bar plots for total sales by region and category.
2. Line charts for monthly sales trends.
3. Pie charts showing sales distribution by product category.

**7. Data Security:**

Security in Fabric must meet the following requirements:

The data engineers must have read and write access to the Lakehouse, including the underlying files.

The data analysts must only have read access to the Delta tables.

**8. Automation:**

1. **Pipeline Scheduling:**
   * Automate the transformation and aggregation.

**Challenges Addressed**

1. **Data Integration Across Formats:** Unified raw data stored in multiple formats (CSV, JSON) into a consistent format.
2. **Scalability:** OneLake ensures fast querying and storage optimization for large datasets.
3. **Automation:** Monthly updates using Pipeline reduce manual effort.
4. **Security:** workspace level and item level.