

1. Create a pipeline to fetch data for 5 countries (India, US, UK, China, Russia) from a REST API and save it as separate JSON files.
2. Add a trigger to the above pipeline to run automatically two times a day (12:00 AM and 12:00 PM IST).
3. Create a pipeline to copy customer data from a database to Azure Data Lake Storage Gen2 (ADLS Gen2) only if the record count is more than 500. Once copied, it should call a child pipeline that copies product data if the customer record count is greater than 600.
4. Design the pipeline to pass the customer count from the parent pipeline to the child product pipeline via a pipeline parameter.

## Part 1: Database Setup (MySQL)

Before setting up the Azure Data Factory pipelines, the source database (MySQL) needs to be prepared with the necessary tables and sample data.

### 1.1 Create Database and Tables

The following SQL script was executed in MySQL Workbench to create the customer\_data\_db database and the CUST\_MSTR (Customer) and Products tables.

**Action:** Execute this script in MySQL Workbench.

-- Drop the database if it exists to ensure a clean start

```
DROP DATABASE IF EXISTS customer_data_db;
```

-- Create the database

```
CREATE DATABASE customer_data_db;
```

-- Use the newly created database

```
USE customer_data_db;
```

-- Create the CUST\_MSTR table

```
CREATE TABLE CUST_MSTR (
  CustomerID INT PRIMARY KEY AUTO_INCREMENT,
  FirstName VARCHAR(100),
  LastName VARCHAR(100),
  Email VARCHAR(255) UNIQUE,
  PhoneNumber VARCHAR(20),
  AddressLine1 VARCHAR(255),
  City VARCHAR(100),
  State VARCHAR(100),
  PostalCode VARCHAR(20),
  LoadDate DATE NOT NULL
);
```

-- Create the Products table

```
CREATE TABLE Products (
  ProductID INT PRIMARY KEY AUTO_INCREMENT,
  ProductName VARCHAR(100) NOT NULL,
  Category VARCHAR(50),
  Price DECIMAL(10, 2) NOT NULL,
  StockQuantity INT NOT NULL,
  Description TEXT,
  LastUpdated DATETIME DEFAULT CURRENT_TIMESTAMP ON UPDATE
  CURRENT_TIMESTAMP
);
```

### 1.2 Insert Sample Data into CUST\_MSTR Table

To test the conditional logic of the ADF pipelines (customer count > 500 and > 600), a stored procedure was created and called to insert 700 sample records into the CUST\_MSTR table.

**Action:** Execute this script in MySQL Workbench. Ensure the entire block, including DELIMITER statements, is selected and executed together.

```
USE customer_data_db;
```

```
DELIMITER //
```

```
CREATE PROCEDURE InsertManyCustomers()
```

```
BEGIN
```

```
    DECLARE i INT DEFAULT 0;
```

```
    WHILE i < 700 DO
```

```
        INSERT INTO CUST_MSTR (
```

```
            FirstName,
```

```
            LastName,
```

```
            Email,
```

```
            PhoneNumber,
```

```
            AddressLine1,
```

```
            City,
```

```
            State,
```

```
            PostalCode,
```

```
            LoadDate
```

```
        )
```

```
        VALUES
```

```
        (
```

```
            CONCAT('TestFirstName_', i),
```

```
            CONCAT('TestLastName_', i),
```

```
            CONCAT('test.email_', i, '@example.com'),
```

```
            CONCAT('555-123-', LPAD(i, 4, '0')),
```

```
            CONCAT(i, ' Test Address'),
```

```
            CASE FLOOR(RAND() * 5)
```

```
                WHEN 0 THEN 'Mumbai'
```

```
                WHEN 1 THEN 'Delhi'
```

```
                WHEN 2 THEN 'Bangalore'
```

```
                WHEN 3 THEN 'Chennai'
```

```
                WHEN 4 THEN 'Hyderabad'
```

```
            END,
```

```
            'MH',
```

```
            '12345',
```

```
            CURDATE()
```

```
        );
```

```
        SET i = i + 1;
```

```
    END WHILE;
```

```
END //
```

```
DELIMITER ;
```

```
CALL InsertManyCustomers();
```

```
-- Verify the count of records after insertion
```

```
SELECT COUNT(*) AS TotalCustomers FROM CUST_MSTR;
```

### 1.3 Insert Sample Data into Products Table

To ensure the product data pipeline copies actual data, sample records were inserted into the Products table.

**Action:** Execute this script in MySQL Workbench.

```
USE customer_data_db;
```

```
INSERT INTO Products (ProductName, Category, Price, StockQuantity, Description) VALUES
('Laptop Pro', 'Electronics', 1200.00, 50, 'High performance laptop'),
('Wireless Mouse', 'Accessories', 25.50, 200, 'Ergonomic wireless mouse'),
('Mechanical Keyboard', 'Accessories', 80.00, 100, 'RGB mechanical keyboard with clicky switches'),
('Gaming Headset', 'Accessories', 50.00, 150, 'Immersive sound for gaming'),
('Webcam HD', 'Electronics', 75.00, 80, 'Full HD video calls');
```

```
SELECT * FROM Products;
```

## Part 2: Azure Data Factory Setup

This section details the configuration of Linked Services, Datasets, Pipelines, and Triggers within Azure Data Factory Studio.

### 2.1 Linked Services

Linked Services define the connection information to external data stores and compute resources.

- **Ls\_RestCountriesApi (REST API):**
  - **Type:** REST
  - **Base URL:** <https://restcountries.com/v3.1/>
  - **Authentication type:** Anonymous
- **Ls\_ADLSGen2 (Azure Data Lake Storage Gen2):**
  - **Type:** Azure Data Lake Storage Gen2
  - **Authentication:** (e.g., Account key or Managed Identity)
  - *(Note: This is reused across pipelines.)*
- **mysql1 (MySQL Database):**
  - **Type:** MySQL
  - **Connection via integration runtime:** (Select your Self-Hosted IR if MySQL is on-premises, or AutoResolveIntegrationRuntime if in Azure and publicly accessible).
  - **Server name:** (e.g., localhost, 192.168.1.100, or Azure FQDN like your-mysql-server.mysql.database.azure.com)
  - **Database name:** customer\_data\_db
  - **User name & Password:** (Your MySQL credentials)

### 2.2 Datasets

Datasets define the structure and location of the data.

- **Ds\_RestCountryData (REST API Source):**
  - **Type:** REST
  - **Linked Service:** Ls\_RestCountriesApi
  - **Relative URL:** (Dynamic, set in pipeline)
  - **Request method:** GET
- **Ds\_ADLSCountryJson (JSON Sink - ADLS Gen2):**
  - **Type:** Azure Data Lake Storage Gen2 (JSON format)
  - **Linked Service:** Ls\_ADLSGen2
  - **File path:** raw-data/countries/ (Directory)
  - **File name:** (Dynamic, set in pipeline)
  - **File format settings:** Array of objects
- **Ds\_CustomerTable (MySQL Customer Source):**
  - **Type:** MySQL
  - **Linked Service:** mysql1
  - **Table:** CUST\_MSTR
- **Ds\_ADLSCustomerData (ADLS Gen2 Customer Sink):**
  - **Type:** Azure Data Lake Storage Gen2 (e.g., Parquet or DelimitedText)
  - **Linked Service:** Ls\_ADLSGen2
  - **File path:** processed-data/customer-data/customers\_{formatDateTime(pipeline().TriggerTime, 'yyyyMMddHHmmss')}.{extension}
- **Ds\_ProductTable (MySQL Product Source):**

- **Type:** MySQL
- **Linked Service:** mysql1
- **Table:** Products
- **Ds\_ADLSProductData (ADLS Gen2 Product Sink):**
  - **Type:** Azure Data Lake Storage Gen2 (e.g., Parquet or DelimitedText)
  - **Linked Service:** Ls\_ADLSGen2
  - **File path:** processed-data/product-data/products\_@{formatDateTime(pipeline().TriggerTime, 'yyyyMMddHHmmss')}. {extension}

## 2.3 Pipelines

### 2.3.1 CountryDataIngestionPipeline (Addresses Question 1)

This pipeline fetches country data from the REST API and saves it to ADLS Gen2.

- **Pipeline Name:** CountryDataIngestionPipeline
- **Variables:**
  - CountryList (Type: Array, Default value: ["india", "us", "uk", "china", "russia"])
- **Activities:**
  - **ForEach\_Country (ForEach Activity):**
    - **Items:** @variables('CountryList')
    - **Inside ForEach:**
      - **Copy\_CountryDataToADLS (Copy Data Activity):**
        - **Source:** Ds\_RestCountryData
          - **Relative URL:** @concat('name/', item())
        - **Sink:** Ds\_ADLSCountryJson
          - **File name:** @concat(item(), '.json')

### 2.3.2 CustomerDataProcessingPipeline (Parent - Addresses Questions 3 & 4)

This pipeline conditionally copies customer data and calls the child pipeline based on customer count.

- **Pipeline Name:** CustomerDataProcessingPipeline
- **Activities:**
  - **Lookup\_CustomerCount (Lookup Activity):**
    - **Source Dataset:** Ds\_CustomerTable
    - **Query:** SELECT COUNT(\*) AS CustomerCount FROM CUST\_MSTR;
    - **First row only:** Checked
  - **If\_CustomerCountGreaterThan500 (If Condition Activity):**
    - **Expression:**

```
@greater(activity('Lookup_CustomerCount').output.firstRow.CustomerCount, 500)
```
    - **Inside "True" Branch:**
      - **Copy\_CustomerDataToADLS (Copy Data Activity):**
        - **Source:** Ds\_CustomerTable
        - **Sink:** Ds\_ADLSCustomerData
      - **If\_CustomerCountGreaterThan600 (Nested If Condition Activity):**
        - **Expression:**

```
@greater(activity('Lookup_CustomerCount').output.firstRow.CustomerCount, 600)
```
        - **Inside "True" Branch (Nested):**
          - **Execute\_ProductPipeline (Execute Pipeline Activity):**
            - **Invoked pipeline:** ProductDataCopyPipeline
            - **Parameters:**
              - **Name:** customerCount
              - **Value:**

```
@activity('Lookup_CustomerCount').output.firstRow.CustomerCount
```

### 2.3.3 ProductDataCopyPipeline (Child - Part of Question 3 & 4)

This pipeline copies product data.

- **Pipeline Name:** ProductDataCopyPipeline
- **Parameters:**
  - customerCount (Type: Integer)
- **Activities:**
  - **Copy\_ProductDataToADLS (Copy Data Activity):**
    - **Source:** Ds\_ProductTable
    - **Sink:** Ds\_ADLSProductData

### 2.4 Triggers (Addresses Question 2)

Triggers schedule pipeline executions.

- **Trigger\_CountryData\_Midnight (Schedule Trigger):**
  - **Type:** Schedule
  - **Pipeline:** CountryDataIngestionPipeline
  - **Recurrence:** Daily
  - **Time zone:** (UTC+05:30) Chennai, Kolkata, Mumbai, New Delhi
  - **Execute at these times:** 00:00 (12:00 AM IST)
  - **Activated:** Yes
- **Trigger\_CountryData\_Noon (Schedule Trigger):**
  - **Type:** Schedule
  - **Pipeline:** CountryDataIngestionPipeline
  - **Recurrence:** Daily
  - **Time zone:** (UTC+05:30) Chennai, Kolkata, Mumbai, New Delhi
  - **Execute at these times:** 12:00 (12:00 PM IST)
  - **Activated:** Yes

### Part 3: Verification Steps

After deploying all components (by clicking **Publish All** in ADF Studio), verify the functionality:

1. **Verify MySQL Data:**
  - In MySQL Workbench, run `SELECT COUNT(*) FROM CUST_MSTR;` to confirm the customer table has 700+ records.
  - Run `SELECT COUNT(*) FROM Products;` to confirm the product table has data.
2. **Trigger Pipelines in ADF:**
  - Manually trigger CountryDataIngestionPipeline and CustomerDataProcessingPipeline using the "Debug" or "Trigger Now" options in ADF Studio.
3. **Monitor Pipeline Runs:**
  - Go to the **Monitor** section in ADF Studio > **Pipeline runs**.
  - Confirm that all triggered pipelines (CountryDataIngestionPipeline, CustomerDataProcessingPipeline, ProductDataCopyPipeline) show a **"Succeeded"** status.
  - For CustomerDataProcessingPipeline, click on the run to verify that Lookup\_CustomerCount returned 700, and that all subsequent conditional activities (Copy Data for Customer, Execute Pipeline for Product) executed.
4. **Check ADLS Gen2 Output:**
  - Navigate to your Azure Data Lake Storage Gen2 account in the Azure Portal.
  - **For Country Data:** Check the raw-data/countries/ container. You should find JSON files named india.json, us.json, uk.json, china.json, russia.json.
  - **For Customer Data:** Check the processed-data/customer-data/ container. You should find a customer data file (e.g., customers\_YYYYMMDDHHMMSS.extension).
  - **For Product Data:** Check the processed-data/product-data/ container. You should find a product data file (e.g., product\_data\_YYYYMMDDHHMMSS.csv)