**Data Pipeline Project – Orders Processing**

**📌 Project Overview**

This project demonstrates a **data pipeline** that ingests, validates, and processes order data from multiple sources using **Azure Data Lake (ADLS Gen2)**, **Azure Data Factory (ADF)**, **Databricks**, **Azure SQL Database**, and **Amazon S3**.

The pipeline ensures:

1. Incoming files (orders.csv) are validated.
   * **No duplicate order\_id**.
   * **Valid order\_status** based on a reference table.
2. Valid files are moved to the **staging folder**, and invalid ones go to the **discarded folder**.
3. Additional integration with **Amazon S3** for order\_items and **Azure SQL DB** for customers.
4. Aggregated results (orders per customer & spending) are pushed back to SQL DB for reporting.

**⚙️ Architecture**

**Data Flow:**  
Third-party service → ADLS Gen2 (Landing) → ADF (trigger) → Databricks (validation & transformation) → ADLS Gen2 (Staging/Discarded) → Azure SQL DB (validations, reporting)

**Components Used:**

* **Azure Data Lake Gen2 (ADLS)** – Landing, Staging, Discarded folders.
* **Azure Data Factory (ADF)** – Orchestration & Storage Event Triggers.
* **Databricks** – Data validation, transformations, Spark jobs.
* **Azure SQL Database** – Stores valid order\_status & customer details.
* **Azure Key Vault** – Secrets management (DB password, tokens).
* **Amazon S3** – Source for order\_items.json.

**🗂️ Resources to be Created**

* **Storage Account**
  + Container with folders: landing, staging, discarded
* **Databricks Workspace**
* **Data Factory**
* **Key Vault**
* **Azure SQL Database**
  + Table: valid\_order\_status
  + Table: customers

**📊 Datasets**

* **Orders** → CSV dropped into ADLS landing folder.
* **Order Items** → JSON from Amazon S3.
* **Customers** → Published in Azure SQL Database.

**✅ Validations Implemented**

1. **Order ID check** – No duplicates allowed.
2. **Order Status check** – Must exist in valid\_order\_status lookup table.

👉 If validations pass → Move to **staging**.  
👉 If validations fail → Move to **discarded**.

**🔑 Security & Secrets**

* All credentials (DB password, Databricks token, Storage keys) are stored in **Azure Key Vault**.
* Databricks accesses Key Vault using **secret scopes**.

**🚀 Future Enhancements**

* Support for **multiple file types dynamically** (not just orders.csv).
* Automated mounting/unmounting for Databricks file system.
* Advanced transformations and reporting.

**📂 Folder Structure (ADLS)**

/sales

├── landing

│ └── orders.csv

├── staging

└── discarded