**Project Title: Data Ingestion and Transformation with Microsoft Azure for PayPal Data**

**Objective**:  
The goal of this project was to build a robust ETL pipeline in Microsoft Azure to ingest, transform, and prepare PayPal transaction data for analysis. This involved data integration, cleaning, transformation, and validation, ensuring the data was ready for further analysis.

**Project Steps:**

**1. Data Ingestion**

* **Data Source**: PayPal data files stored in Azure Blob Storage.
* **Process**: Used Azure Data Factory (ADF) to connect to Blob Storage and ingest the data. This involved setting up linked services and datasets in ADF to read and handle the raw transaction data.
* **Result**: Successfully extracted raw data from Blob Storage into the ETL pipeline.

**2. Data Transformation**

* **Objective**: Clean and format the data for consistency, merge multiple sources, and apply necessary filters.
* **Steps**:
  + **Data Flow Setup**: Created a data flow in Azure Data Factory, enabling the flow of data from Blob Storage to ADF.
  + **Merging Data**: Combined multiple data sources within the data flow to create a unified dataset. This step ensured all relevant data was available in a single flow for easier manipulation.
  + **Filtering by Transaction Status**: Applied data flow expressions to filter transactions based on specific statuses, such as "Completed," "Pending," or "Failed." This allowed for targeted analysis of relevant transactions only.
* **Outcome**: Created a clean, merged, and filtered dataset, suitable for storage in the database and ready for further transformation steps.

**3. Data Loading (Sink)**

* **Purpose**: Load the transformed data into a storage destination for analysis and reporting.
* **Implementation**: Used the Sink feature in ADF to transfer the final transformed dataset to an Azure SQL Database. This involved configuring ADF to securely load data from the pipeline into SQL.
* **Result**: The data was successfully stored in the Azure SQL Database, providing a structured and queryable data repository.

**4. Validation and Error Handling**

* **Objective**: Ensure data integrity and process reliability.
* **Process**:
  + Configured validation checkpoints within the pipeline to monitor data flow and detect any breakages or irregularities.
  + Implemented error handling to capture and report issues, enabling quick troubleshooting.
* **Outcome**: This validation layer provided a safeguard, ensuring data flows were uninterrupted or promptly addressed if any issues arose.

**5. Pipeline Activation and Scheduling**

* **Pipeline**: Created a pipeline in ADF that orchestrates each ETL step, from ingestion to loading.
* **Trigger**: Added a trigger to activate the pipeline at designated intervals or on-demand, ensuring timely data processing.
* **Outcome**: The pipeline now runs automatically, enabling regular ingestion and transformation of PayPal data.

**6. ETL Preprocessing Benefits**

* **Value**: By preprocessing the data through ETL, the pipeline ensured clean, validated, and transformed data was ready for analysis before being integrated into any systems. This not only optimized data quality but also improved efficiency for downstream analytics.

**Key Learnings**

* Gained experience in data ingestion, transformation, and validation using Azure Data Factory.
* Developed a comprehensive understanding of ETL best practices, from data integration to validation.
* Enhanced skills in troubleshooting and error handling within Azure to ensure a resilient data pipeline.

**Screenshots:**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**