**Document Title:** Project Development Architecture & Flow

**Project Title:** Customer Transactions Data Pipeline

**Project Development Flow (Planned Approach):**

This document outlines the project execution strategy after the requirement analysis.

**Phase 1: Initial Setup and Database Design**

* Create an Azure Storage account and add three containers:
  + raw: for original files
  + processed: for cleaned files
* Will create Errorlogs table: for rejected or error files
* Set up an Azure SQL Database to store final clean data.
* Design six tables (Table and column Description):

**Fact table: Transactions:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| TransactionID | INT (Primary Key) | Unique identifier for each transaction. |
| CustomerID | INT | Direct reference to Customers.CustomerID. |
| ProductID | INT | Direct reference to Products.ProductID. |
| PaymentMethodID | INT | Direct reference to PaymentMethods.PaymentMethodID. |
| LocationID | INT | Direct reference to Locations.LocationID. |
| TransactionDate | DATE | The actual date of the transaction. |
| Quantity | INT | Number of items or services purchased in the transaction. |
| Amount | DECIMAL(10,2) | Price per unit or per service. |
| TotalAmount | DECIMAL(10,2) | Computed value = Quantity \* Amount. |

**dimension tables: Customers, Products, Payment Methods, and Locations:**

**1. Customers:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| CustomerID | INT (Primary Key) | Unique identifier for each customer. |
| CustomerName | NVARCHAR(50) | Full name of the customer. |
| Email | NVARCHAR(50) | Email address of the customer. |
| PhoneNumber | NVARCHAR(15) | Contact number of the customer. |

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| ProductID | INT (Primary Key) | Unique identifier for each product or service. |
| ProductName | NVARCHAR(50) | Name or title of the product or service. |
| ProductCategory | NVARCHAR(50) | Category of the product (e.g., electronics, books). |

**2.Products:**

**3.Payment\_Methods:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| PaymentMethodID | INT (Primary Key) | Unique identifier for each payment method. |
| PaymentMethodType | NVARCHAR(20) | Type of payment (e.g., UPI, credit card, net banking). |

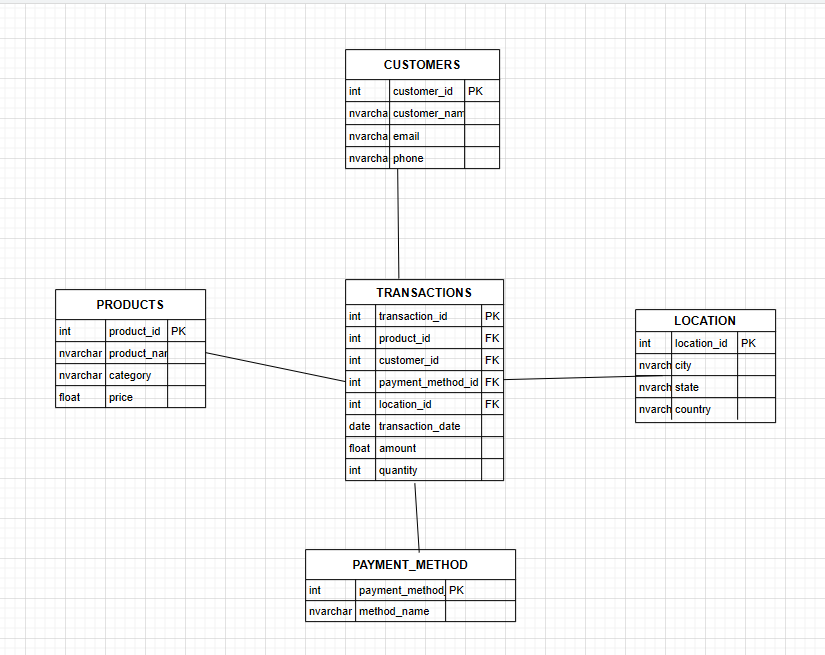
**4.Locations:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| LocationID | INT (Primary Key) | Unique identifier for each location. |
| City | NVARCHAR(50) | Name of the city. |
| State | NVARCHAR(50) | Name of the state or province. |
| Country | NVARCHAR(50) | Name of the country. |

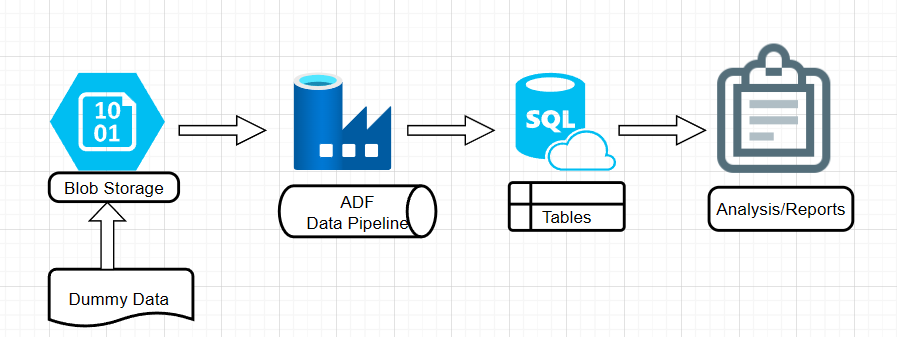
**ErrorLog Table (for error loging):**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| PipelineId | VARCHAR(100) | Unique identifier of the ADF pipeline run that attempted the data load. |
| FileName | VARCHAR(100) | Name of the file being processed when the failure occurred. |
| Status | VARCHAR(20) | Status of the load operation (e.g., 'Failed'). |
| Log | VARCHAR(MAX) | Detailed error message returned by the pipeline on failure. |
| LogTime | DATETIME | Timestamp when the failure log entry was created. |

**Schema Diag.:**



**Flow Diag.:**



**Phase 2: Create Dummy Data Using Python**

* Will Use Python libraries like Faker and Pandas to create sample data.
* Generate:
  + Will Generate records for Fact Transactions Table.
  + Will generate records for each dimension table
* Cleaning the data:
  + Will remove missing and duplicate records
  + Will format date, email, and numbers correctly
* Save all files as CSVs and upload them into the appropriate Azure Blob container.

**Phase 3: Data Ingestion Using Azure Data Factory (ADF)**

* Will Build pipelines in ADF to move data from Blob Storage to Azure SQL DB.
* Will Use triggers to automatically run pipelines when new files arrive.
* Will Add error handling to log any problems to the logs/ container.

**Phase 4: Stored proc. (for error log)**

* Will Develop a stored procedure to insert error details into the log table dynamically during pipeline execution.

**Phase 5: SQL Queries and Data Checks**

* Will write SQL queries to:
  + Check if all data is loaded correctly
  + Verify links between fact and dimension tables
* Will create analysis queries such as:
  + Revenue by location
  + Top spending customers
  + Revenue by product category

**Phase 6: Documentation**

* Will Create and update a GitHub repository with all:
  + Python code
  + SQL scripts
  + ADF JSON pipeline

**Prepared by:** Shanu Bharti  
**Date:** April 17, 2025