Technical Document: Environment Setup for Azure Data Engineering Project

Introduction

This technical document outlines the data ingestion process as part of the comprehensive Azure Data Engineering Project. Data ingestion involves transferring tables from an on-premises SQL Server database to Azure Data Lake Gen 2 using Azure Data Factory. This document provides a step-by-step guide to perform data ingestion.

Prerequisites

Before proceeding with data ingestion, ensure the following prerequisites are met:

Azure Resources: Confirm that the required Azure resources, including Azure Data Factory, Azure Data Lake Gen 2, and Azure Synapse Analytics, have been created as outlined in the project setup.

SQL Server Database: Verify the existence of the on-premises SQL Server database from which data will be ingested. Ensure that a login and user have been created for accessing the database.

Key Vault: Secrets for the SQL Server username and password should be stored securely in Azure Key Vault.

Environment Setup

The data ingestion process involves establishing a connection between the on-premises SQL Server database and Azure Data Factory. Data Factory is then used to copy tables from the SQL Server database to Azure Data Lake Gen 2.

1. Creating a Login and User

Before initiating data ingestion, a login and user should be created for accessing the on-premises SQL Server database. This is done using SQL Server Management Studio (SSMS). Follow these steps:

Open SSMS and connect to the SQL Server.

Expand the database section and select the database containing the tables to be ingested (e.g., Adventure Works LT 2017).

Create a login and user with appropriate permissions. For example:

sql

CREATE LOGIN <username> WITH PASSWORD = <password>;

USE [AdventureWorksLT2017] – at the top and then in the script

CREATE USER <name> FOR LOGIN <username>;

2. Assigning Roles to the User

The user created in the previous step requires appropriate permissions to read the tables in the SQL Server database. Assign the DB_datareader role to the user by following these steps:

In SSMS, right-click on the user under Security and select Properties.

Under Membership, assign the DB_datareader role.

3. Storing Secrets in Key Vault

To securely manage the SQL Server username and password, store them as secrets in Azure Key Vault. This ensures that credentials are not exposed in the Azure environment. Follow these steps:

Access the Azure portal and navigate to the Azure Key Vault in the resource group.

Create secrets for the username and password and store their values.

4. Data Factory Configuration

Azure Data Factory is used to establish a connection with the on-premises SQL Server database and copy tables to Azure Data Lake Gen 2. Configure Data Factory as follows:

Create a Data Factory pipeline.

Configure linked services for the SQL Server database and Azure Data Lake Gen 2.

Define datasets representing the tables to be copied.

Create copy activities within the pipeline to copy data from SQL Server to Data Lake.

Conclusion

Data ingestion is a critical step in the Azure Data Engineering Project, enabling the transfer of data from an on-premises SQL Server database to Azure. By creating logins, assigning roles, and securely storing credentials in Key Vault, you ensure the security and integrity of your data. The use of Azure Data Factory simplifies the data transfer process and allows for efficient data management. With data successfully ingested, you can proceed to the next phases of data transformation, loading, and reporting.