Technical Document: Azure Data Engineering Project Overview

Introduction

This technical document provides an overview of a comprehensive data engineering project using Azure technologies. The project involves migrating data from an on-premises SQL Server database to Azure cloud using various Azure resources. This document outlines the tools and steps involved in this project.

Project Overview

In this data engineering project, we will leverage Azure resources to demonstrate an end-to-end data pipeline. The project's use case is the common scenario of migrating an on-premises SQL Server database to the Azure cloud. This project aims to help viewers gain a clear understanding of using Azure tools for data engineering and prepare for data engineering interviews.

Tools and Azure Resources Used

The following Azure tools and resources will be utilized in this project:

Azure Data Factory: This is an Extract, Transform, Load (ETL) tool used for data ingestion. It connects to the on-premises SQL Server database and transfers tables to the cloud.

Azure Data Lake Gen 2: Azure Data Lake Storage Gen 2 is used as the storage solution to store data transferred from the on-premises database.

Azure Databricks: Azure Databricks is a big data analytics tool that is utilized for data transformation. It allows for data processing using languages like SQL, PySpark, and Python.

Lakehouse Architecture: The project implements a Lakehouse architecture, which includes different layers (Bronze, Silver, Gold) for data transformation. The Bronze layer holds the raw data as-is, while the silver and gold layers involve progressive transformations and cleaning.

Azure Synapse Analytics: Azure Synapse Analytics is used to create databases and tables to house the transformed data, similar to an on-premises SQL Server database.

Power BI: Power BI is employed for data reporting and visualization. It creates reports and charts based on the data stored in Azure Synapse Analytics.

Azure Active Directory: Azure Active Directory serves as an identity and access management tool for security-related configurations.

Azure Key Vault: Azure Key Vault is used to securely store sensitive information such as usernames and passwords.

Project Workflow

The project follows a structured workflow to achieve data migration and reporting:

Data Ingestion (Azure Data Factory): Azure Data Factory is used to connect to the on-premises SQL Server database and copy tables to Azure Data Lake Gen 2.

Data Transformation (Azure Databricks): Azure Databricks is employed for data transformation. Data is processed and cleaned in multiple stages, from Bronze to Silver to Gold layers, following the Lakehouse architecture.

Data Loading (Azure Synapse Analytics): Azure Synapse Analytics is utilized to create databases and tables mirroring the on-premises database structure. Data from the Gold layer is loaded into these tables.

Data Reporting (Power BI): Power BI is used to create reports and visualizations based on the data in Azure Synapse Analytics. This step allows for data analysis and reporting.

Automation (Azure Pipelines): Automation is crucial in maintaining data pipelines. Azure Pipelines can be configured to automate data updates and transformations, ensuring the data remains up to date.

Conclusion

This technical document provides an overview of the Azure data engineering project, including the tools and Azure resources used, as well as the workflow involved. The project's goal is to demonstrate how to migrate on-premises data to Azure and create meaningful reports using Power BI. Understanding these concepts can be valuable for data engineering professionals and job interviews.