**Project Title: Covid Data Pipeline using Microsoft Azure**

**Overview**

This project is an implementation of a Covid data pipeline using various Microsoft Azure services, such as Azure Data Lake Gen 2 Storage, Blob Storage, Azure Databricks, PySpark, and Azure Data Factory. The goal of this project is to extract, transform, and load (ETL) Covid data from the European Centre for Disease Prevention and Control (ECDC) website for further analysis and visualization in PowerBI.

**Table of Contents**

1. Data Source
2. Data Pipeline Architecture
3. Azure Services and Components
4. Data Flow and Transformation
5. Output and Visualization
6. Limitations and Future Improvements

**1. Data Source**

The data for this project is sourced from the ECDC website, which provides up-to-date and reliable information on Covid cases across Europe. The data is ingested into Azure Blob Storage for initial storage.

**2. Data Pipeline Architecture**

The data pipeline architecture consists of the following steps:

Ingesting data from the ECDC website into Azure Blob Storage.

Copying the data from Azure Blob Storage to Azure Data Lake Gen 2 Storage.

Performing ETL operations on the data using Azure Databricks and PySpark.

Storing the processed data into an Azure SQL database.

Visualizing and analyzing the data using PowerBI.

**3. Azure Services and Components**

Azure Blob Storage: Used for initial storage of the Covid data.

Azure Data Lake Gen 2 Storage: Serves as a scalable and cost-effective storage solution for the data pipeline.

Azure Databricks: Provides a managed Spark environment for ETL operations and complex data transformations.

PySpark: A Python library used in Azure Databricks for data processing and transformation tasks.

Azure Data Factory: Orchestration tool used for data movement and ETL activities.

Azure SQL Database: Final storage location for the processed data, which can be queried for analytics and reporting purposes.

PowerBI: Business intelligence tool used for data visualization and analysis.

**4. Data Flow and Transformation**

Data is ingested from the ECDC website and stored in Azure Blob Storage.

Data is copied from Azure Blob Storage to Azure Data Lake Gen 2 Storage, with three main folders: lookup, raw, and processed.

Azure Databricks performs ETL activities on the raw data using PySpark, including filtering, cleaning, and transforming the data.

Processed data is stored in the Azure SQL database for further querying and analysis.

**5. Output and Visualization**

The processed data stored in the Azure SQL database is used for visualization and analysis using PowerBI. Custom reports and dashboards can be created to display important metrics and trends related to Covid cases in Europe.

**6. Limitations and Future Improvements**

Azure Synapse was not used as a data warehouse in this project due to the limited volume of data. For larger datasets, Azure Synapse could be considered for better scalability and performance.

This project focuses on data from the ECDC website. In the future, additional data sources could be integrated to provide a more comprehensive view of Covid cases worldwide.

The data pipeline could be further optimized and expanded to include additional ETL operations or machine learning models for predicting trends or identifying patterns in the data.