Airline data ingestion using azure synapse analytics

- Bhushan Sagar

……

Business Overview:

Apache Spark, a parallel processing framework that enhances the performance of big data analytical applications through in-memory processing, is supported by Azure Synapse Analytics, one of Microsoft's cloud implementations of Apache Spark, in addition to Databricks. Azure Synapse offers serverless Apache Spark pools that are easy to set up and configure. These Spark pools are compatible with Azure Storage and Azure Data Lake Generation 2 Storage, allowing for the processing of data stored in Azure. In this project, we will create a pipeline using Azure Synapse Analytics, Azure Storage, an Azure Synapse Spark pool, and Power BI to transform an airline dataset. The transformed data will be stored in tables associated with the Spark pool and visualized using Power BI.

Data Description:

The dataset utilized in this project is an Airline dataset comprising approximately 4,000 records with various fields. Some of the key parameters included in the dataset are:

- Date

- Flight carrier details

- Origin details

- Destination details

- Delay reason

- Distance

- Elapsed time

Tech Stack:

- Framework: Spark

- Languages: SQL, Python

- Services: Azure Synapse Analytics, Azure Storage, Azure Synapse Spark Pool, Power BI

Azure Synapse Analytics:

Azure Synapse is a cloud-based analytics service that integrates enterprise data warehousing and big data analytics. It enables querying data according to your preferences, utilizing either serverless resources or provisioned resources at scale. Azure Synapse bridges the gap between these domains by offering a unified experience for ingesting, preparing, managing, and delivering data for business intelligence (BI) and machine learning needs.

Azure Storage:

Microsoft's Azure Storage platform offers a cloud solution for contemporary data storage needs. Azure Storage provides highly available, massively scalable, durable, and secure cloud storage for various data objects.

Approach:

- Set the location and type of the data.

- Read all the airline data stored under the folder "airport" spread across multiple files.

- Query the data as a Spark DataFrame.

- Create a temporary view or table for analyzing the data.

- Create a persistent, permanent table.

Key Takeaways:

- Introduction to Synapse Analytics

- Understanding Spark Pool

- Difference between SQL Pool and Spark Pool

- Understanding the Project Architecture

- Creating a Spark Pool

- Loading data into Azure Storage

- Ingesting data using Synapse Data Factory

- Processing the data using Spark

- Storing transformed data in permanent tables

- Visualization using Power BI

Architecture Diagram: