Build a Data Pipeline with Azure Synapse and Spark Pool

Business Overview:

In-memory processing is supported by Apache Spark, a parallel processing framework, to improve the performance of big data analytical applications. One of Microsoft's cloud implementations of Apache Spark is in Azure Synapse Analytics, apart from Databricks. A serverless Apache Spark pool on Azure that is simple to set up and configure owing to Azure Synapse. Azure Storage and Azure Data Lake Generation 2 Storage are compatible with the Spark pools in Azure Synapse. Therefore, one can process Azure-stored data using Spark pools. In this project, we will build a pipeline in Azure using Azure Synapse Analytics, Azure Storage, Azure Synapse Spark pool, and Power BI to perform data transformations on an Airline dataset, store them in tables associated with the Spark pool and visualize the data in Power BI.

Data Description:

The dataset used in this project is an Airline dataset with around 4000 records and multiple fields. A few of the parameters included in the dataset are:

- Date
- Flight_carrier details
- Origin details
- Destination details
- Delay reason
- Distance
- Elapsed_time

Tech Stack:

Framework: Spark

Language: 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 combines enterprise data warehouse and big data analytics. It allows you to query data on your terms, whether using serverless resources or being provisioned at scale. Azure Synapse bridges the gap between these two worlds by providing a unified experience for ingesting, preparing, managing, and delivering data for BI and machine learning needs.

Azure Storage

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

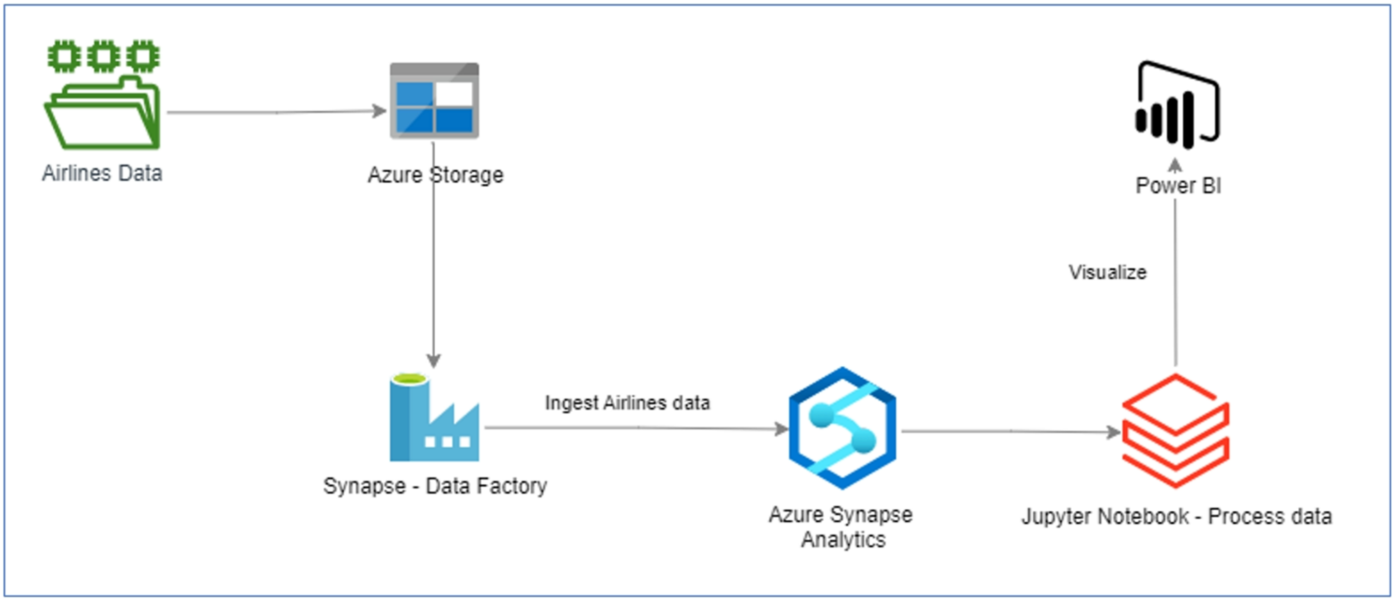
Approach

- Set the location and type of the data
- Read all the airline data 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 in Azure Storage
- Ingest Data using Synapse Data Factory
- Process the data using Spark
- Store transformed data in permanent tables
- Visualization using Power BI

Architecture Diagram:


Airlines Data
Azure Storage
Power BI
Visualize
Ingest Airlines data
Synapse - Data Factory
Azure Synapse Analytics
Jupyter Notebook - Process data