Create Delta Live Tables in Azure Databricks

Business Overview

The process of analyzing and measuring data as soon as it enters the database is referred to as real-time analytics. Thus, users gain insights or may conclude as soon as data enters their system. Businesses can react quickly using real-time analytics. They can grasp opportunities and avert issues before they occur.

On the other hand, Batch-style analytics might take hours or even days to provide findings. As a result, batch analytical systems frequently produce only static insights based on lagging indications. Real-time analytics insights may help organizations stay ahead of the competition. These pipelines for streaming data generally follow a 3 step process, i.e., Ingest, Analyze and Deliver.

We aim to build a Delta Live Tables pipeline in Azure Databricks to process streaming and batch data coming from different sources. The batch data is stored in Azure Data Lake Storage, whereas the streaming data is ingested using Azure Event Hub. We will perform various kinds of transformations in each layer of the Delta Live Table. Using Power BI, we will also visualize the data stored in the gold layer of the Delta Live Tables.

Tech Stack

➔
Language: SQL, Spark, Python

➔
Services: Azure Event Hub, Azure Data Lake Storage, Azure Databricks, Delta Live Tables, Power BI

Delta Live Tables:

Delta Live Tables is a system for creating dependable, manageable, and tested data processing pipelines. Delta Live Tables controls task orchestration, cluster management, monitoring, data quality, and error handling while you specify the data transformations to be applied to your data.

Delta Live Tables regulates how your data is transformed based on a target schema you designate for each processing stage instead of creating your data pipelines using several different Apache Spark tasks. Additionally, you can impose data quality standards using Delta Live Tables. Expectations let you declare the expected level of data quality and how to deal with records that don't meet them.

Key Process

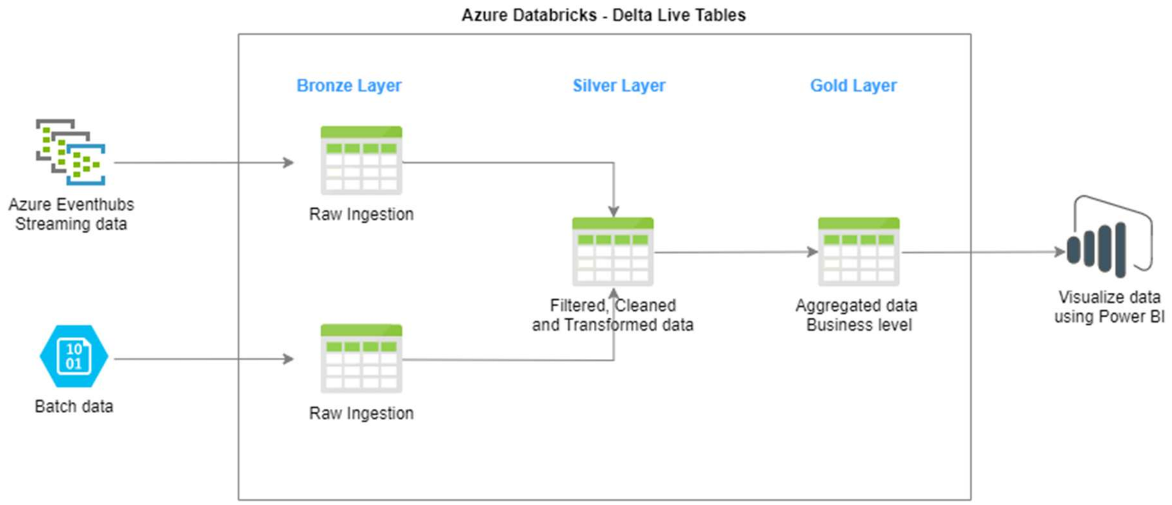
- Understanding the Airline Dataset
- Understanding the concept of Delta Live Tables
- Multiple use cases of Delta Live Tables
- Creating Azure Resource Group

- Creating Azure Data Lake Storage account
- Creating Azure Event Hub namespace
- Ingest streaming data into Event Hub
- Understanding the python code to ingest streaming data
- Upload batch data into Azure Data Lake Storage container
- Creating a Azure Databricks workspace
- Creating a computing cluster in Databricks workspace
- Importing notebooks from local system into Databricks workspace
- Loading batch data from Azure Data Lake Storage container into Databricks table
- Loading streaming data from Event Hub into Databricks file storage
- Creating a Delta Live Tables Pipeline
- Understanding the configurations of Delta Live Tables Pipeline
- Understanding the Bronze, Silver, and Gold layers of Delta Live Tables Pipeline
- Apply different types of transformations to data stored in each layer of pipeline
- Load data from Delta Live Tables into Power BI
- Creating visualizations in Power BI

Approach

1) Create an azure data lake storage account and upload the batch data in the container.
2) Create an azure event hub and ingest streaming data into it using python script.
3) Create an azure databricks workspace.
4) Create a computing cluster in databricks workspace.
5) Load batch data from Azure Data Lake Storage container into Databricks table
6) Load streaming data from Event Hub into Databricks file storage
7) Create a Delta Live Tables pipeline to process the streaming and batch data.
8) Apply transformations on tables stored in Bronze and Silver layers.
9) Store the cleaned data into Gold layer of Delta Live Tables pipeline.
10) Load data from Gold layer tables into Power BI.
11) Creating visualizations in Power BI.

Architecture Diagram:


Azure Databricks - Delta Live Tables
Bronze Layer
Silver Layer
Gold Layer
Azure Eventhubs
Streaming data
Raw Ingestion
Batch data
Raw Ingestion
Filtered, Cleaned
and Transformed data
Aggregated data
Business level
Visualize data
using Power BI