Databricks Real-Time Streaming with Event Hubs and Snowflake

Business Overview
Real-time data processing and analytics using Azure Event Hubs, Databricks, and Snowflake involve ingesting, processing, storing, and analyzing large volumes of data in real-time, providing near-instantaneous insights into business operations.
Here are some advantages of using these technologies together:

- Event Hubs can handle millions of events per second, Databricks can process large volumes of data, and Snowflake can store and manage petabytes of data. Together, they provide a highly scalable platform for processing and analyzing large amounts of data.
- Event Hubs and Databricks support real-time data processing, allowing organizations to respond quickly to changing data and generate real-time insights. Snowflake also supports real-time data ingestion, allowing organizations to load and process data in near real time.
- Databricks provide a powerful platform for performing complex data analysis, including machine learning and data visualization. Snowflake provides a highly scalable data warehouse platform for storing and querying data. Together, they provide a comprehensive platform for data analytics.
- All of them are cloud-based services, meaning organizations can take advantage of cloud-based pricing models, including pay-as-you-go pricing and flexible scaling. This can help organizations to reduce costs and improve the return on investment for their data processing and analytics projects.
- They provide robust security features, including data encryption, access controls, and auditing. This can help organizations to protect their data and comply with regulatory requirements.

One example of using Azure Event Hubs and Databricks for real-time data processing and analytics is monitoring IoT devices which will be demonstrated in this project.

Dataset Description:
In this project, we will use a Microsoft Sample Dataset capturing the GPS coordinates of the phone with the following fields:

- Index - User ID
- Arrival_Time - GPS tracker arrival time
- Creation_Time - GPS tracker captures time
- x - GPS co-ordinate
- y - GPS co-ordinate
- z - GPS co-ordinate
- User - dummy column

- Model - Phone Model
- Device - Phone version
- gt - IOT device location
- Id -ID of the reading
- Geolocation - City, Country of the GPS reading

Tech Stack:
Framework: Spark
Language: Scala, Python
Services: Azure Blob Storage, Azure Databricks, Azure Event Hubs, Snowflake

Azure Event Hubs

Azure Event Hubs is a large data streaming platform and event ingestion service. It can receive and process millions of events per second. Data delivered to an event hub may be converted and saved using any real-time analytics provider or batching/storage adapters.

Databricks

Databricks is a cloud platform for large-scale data engineering and collaborative data science. Databricks combines best-in-class data warehousing and data lakes into a laser house architecture. Databricks is developing a web-based Spark framework that provides automated cluster management and IPython-style notebooks. In addition to building the Databricks platform, the company co-hosted the Spark Massive Open Online Courses and a conference for the Spark community. Databricks also provides a platform for other workloads, including machine learning, data warehousing, streaming analytics, and business intelligence.

Snowflake

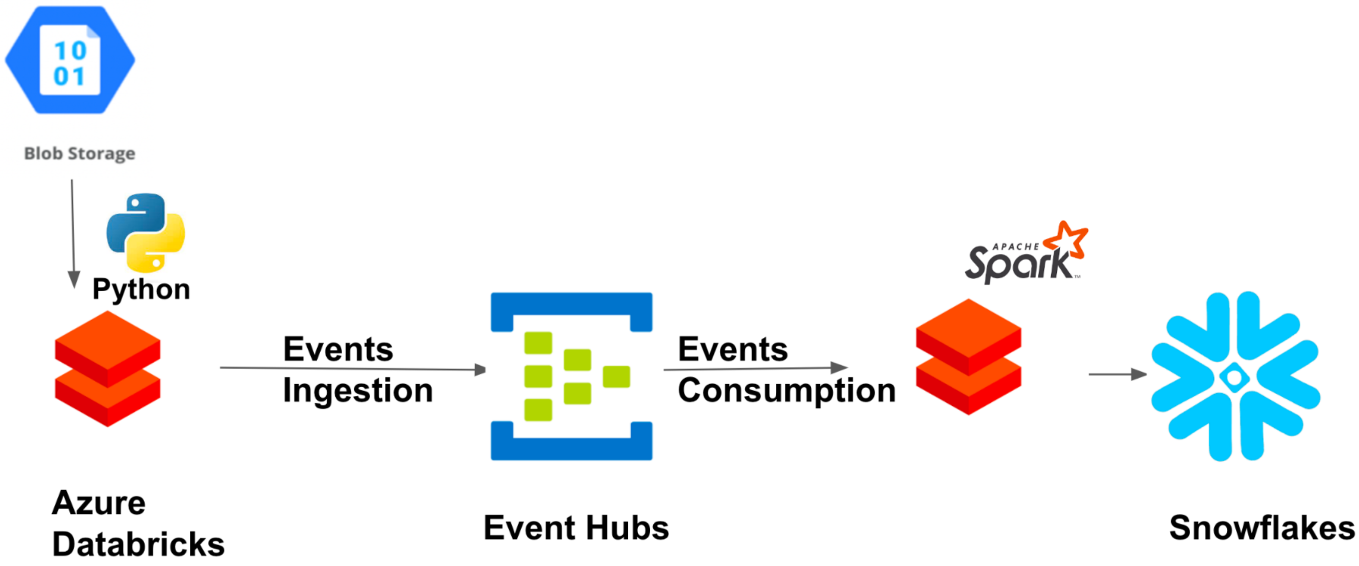
Snowflake is a data storage, processing, and analytics platform that blends a unique SQL query engine with a cloud-native architecture. Snowflake delivers all the features of an enterprise analytic database to the user. Snowflake components include:

- Warehouse/Virtual Warehouse
- Database and Schema
- Table
- View
- Stored procedure
- Snowpipe
- Stream
- Task

Key Takeaways:

- Understanding Databricks advantages
- Understanding Event Hubs advantages
- Understanding Snowflake Architecture
- Pros and Cons of Blob Storage
- Create Resource Groups
- Configure Compute Cluster in Databricks
- Mount dataset in Blob using Scala in Databricks
- Understanding Structured Streaming
- Create Event Hub
- Setup Snowflake Account
- Stream data into Event Hub using Databricks
- Consume data from Event Hub
- Load data in Snowflake

Architecture


10
01
Blob Storage
Python
Events
Ingestion
Events
Consumption
APACHE Spark
Azure
Databricks
Event Hubs
Snowflakes