

In this exercise, write the data from the event hub created in the previous exercise into a location in databricks. Then, stream the data from databricks into a table called youtube\_videos in Snowflake.

Now, query the snowflake table, and you should be able to see the data.

**Title:** Streaming Data from Azure Event Hub to Snowflake via Databricks

**Objective:** Continuing from the previous exercise, the objective of this exercise is to stream data from the Azure Event Hub, write it into a location in Databricks, and then stream the data from Databricks into a Snowflake table called "youtube\_videos." Finally, validate the data by querying the Snowflake table.

**(SOL)**

#### Tasks:

1. Configure Event Hub to Databricks Integration: Establish the connection between Azure Event Hub and Databricks. Set up the necessary credentials and configurations to access the Event Hub.

```
event_hub_connection_string = "<event_hub_connection_string>"
event_hub_name = "<event_hub_name>"
```

2. Stream Data to Databricks: Implement Spark code in Databricks to consume data from the Event Hub. Write the data to a specific location in Databricks, such as a Delta table or Parquet files.

```
from pyspark.sql import SparkSession
```

```
spark = SparkSession.builder.appName("EventHubToDatabricks").getOrCreate()
```

```
df = spark.readStream \
    .format("eventhubs") \
    .option("eventhubs.connectionString", event_hub_connection_string) \
    .option("eventhubs.consumerGroup", "$Default") \
    .option("eventhubs.startingPosition", "earliest") \
    .load()
```

```
# Apply necessary transformations to df
```

```
output_path = "/mnt/<mount_point>/eventhub_data"
df.writeStream \
```

```
.format("delta") \
.option("path", output_path) \
.option("checkpointLocation", "/mnt/<mount_point>/eventhub_checkpoint") \
.start()
```

3. Create Snowflake Table: Set up a table named "youtube\_videos" in Snowflake with the appropriate schema that matches the data from the Event Hub. Define the column names and data types accordingly.

Create a Snowflake Table "youtube\_videos" with the necessary schema as how it is going to be written and defined in Databricks

4. Stream Data from Databricks to Snowflake: Configure a data streaming pipeline in Databricks to read the data from the location where it was written and stream it into the Snowflake table. Use Snowflake Connector to establish the connection and write data to the table.

```
from pyspark.sql import SparkSession
```

```
spark = SparkSession.builder.appName("DatabricksToSnowflake").getOrCreate()
```

```
df = spark.readStream \
    .format("delta") \
    .load(output_path)
```

```
df.writeStream \
    .format("snowflake") \
    .option("sfURL", "<snowflake_url>") \
    .option("sfDatabase", "<snowflake_database>") \
    .option("sfWarehouse", "<snowflake_warehouse>") \
    .option("sfSchema", "<snowflake_schema>") \
    .option("sfRole", "<snowflake_role>") \
    .option("dbtable", "youtube_videos") \
    .option("checkpointLocation", "/mnt/<mount_point>/snowflake_checkpoint") \
    .start()
```

5. Validate Data in Snowflake: Execute SQL queries in Snowflake to retrieve and examine the data stored in the "youtube\_videos" table. Verify that the data streaming process from Databricks to Snowflake was successful.

Execute SQL Statements and validate the Data load and do analysis on top of it.

2023-02-07 3:36pm

ACCOUNTADMIN - COMPUTE\_WH

TEST\_SRC\_OUTPUT

```
1 | select * from iot_with_geo
```

Updated 6 seconds ago

Objects Editor Results Chart

	ARRIVAL_TIME	CREATION_TIME	DEVICE	INDEX	MODEL	USER	GT	X
1	2015-02-23 10:27:53.000	2015-02-23 10:58:39.111	nexus4_2	83,306	nexus4	g	sit	0.0007781982
2	2015-02-23 10:27:52.000	2015-02-23 10:27:50.272	nexus4_1	85,992	nexus4	g	sit	0.014007568
3	2015-02-23 10:27:51.000	2015-02-23 10:58:37.913	nexus4_2	83,068	nexus4	g	sit	0.013595581
4	2015-02-23 10:27:51.000	2015-02-23 10:58:37.510	nexus4_2	82,988	nexus4	g	sit	0.024276733
5	2015-02-23 10:27:51.000	2015-02-23 10:58:37.108	nexus4_2	82,908	nexus4	g	sit	0.013595581
6	2015-02-23 10:27:50.000	2015-02-23 10:58:36.704	nexus4_2	82,828	nexus4	g	sit	-0.005630493
7	2015-02-23 10:27:50.000	2015-02-23 10:27:48.253	nexus4_1	85,591	nexus4	g	sit	0.011871338
8	2015-02-23 10:27:49.000	2015-02-23 10:27:47.855	nexus4_1	85,512	nexus4	g	sit	-0.015899658
9	2015-02-23 10:27:49.000	2015-02-23 10:58:35.496	nexus4_2	82,588	nexus4	g	sit	-0.0013580322
10	2015-02-23 10:27:48.000	2015-02-23 10:58:35.093	nexus4_2	82,508	nexus4	g	sit	-0.0077667236

Result limit exceeded  
Only 10,000 rows of the results are displayed. Please download the results for all of the rows.

Query Details

Query duration 1.7s

Rows 207.9K

Calculating stats...

## Learning Outcomes:

- Integrating Azure Event Hub with Databricks.
- Streaming data from Event Hub to Databricks.
- Writing data from Databricks to a specific location.
- Setting up a Snowflake table for data ingestion.
- Streaming data from Databricks to Snowflake using the Snowflake Connector.
- Querying and validating data in Snowflake.