

Lab 11 - Create a stream processing solution with Event Hubs and Azure Databricks

In this lab, you will learn how to ingest and process streaming data at scale with Event Hubs and Spark Structured Streaming in Azure Databricks. You will learn the key features and uses of Structured Streaming. You will implement sliding windows to aggregate over chunks of data and apply watermarking to remove stale data. Finally, you will connect to Event Hubs to read and write streams.

After completing this lab, you will be able to:

- Know the key features and uses of Structured Streaming
- Stream data from a file and write it out to a distributed file system
- Use sliding windows to aggregate over chunks of data rather than all data
- Apply watermarking to remove stale data
- Connect to Event Hubs read and write streams

Concepts

Apache Spark Structured Streaming is a fast, scalable, and fault-tolerant stream processing API. You can use it to perform analytics on your streaming data in near real time.

With Structured Streaming, you can use SQL queries to process streaming data in the same way that you would process static data. The API continuously increments and updates the final data.

Event Hubs and Spark Structured Streaming

Azure Event Hubs is a scalable real-time data ingestion service that processes millions of data in a matter of seconds. It can receive large amounts of data from multiple sources and stream the prepared data to Azure Data Lake or Azure Blob storage.

Azure Event Hubs can be integrated with Spark Structured Streaming to perform processing of messages in near real time. You can query and analyze the processed data as it comes by using a Structured Streaming query and Spark SQL.

Streaming concepts

Stream processing is where you continuously incorporate new data into Data Lake storage and compute results. The streaming data comes in faster than it can be consumed when using traditional batch-related processing techniques. A stream of data is treated as a table to which data is continuously appended. Examples of such data include bank card transactions, Internet of Things (IoT) device data, and video game play events.

A streaming system consists of:

- Input sources such as Kafka, Azure Event Hubs, IoT Hub, files on a distributed system, or TCP-IP sockets
- Stream processing using Structured Streaming, forEach sinks, memory sinks, etc.

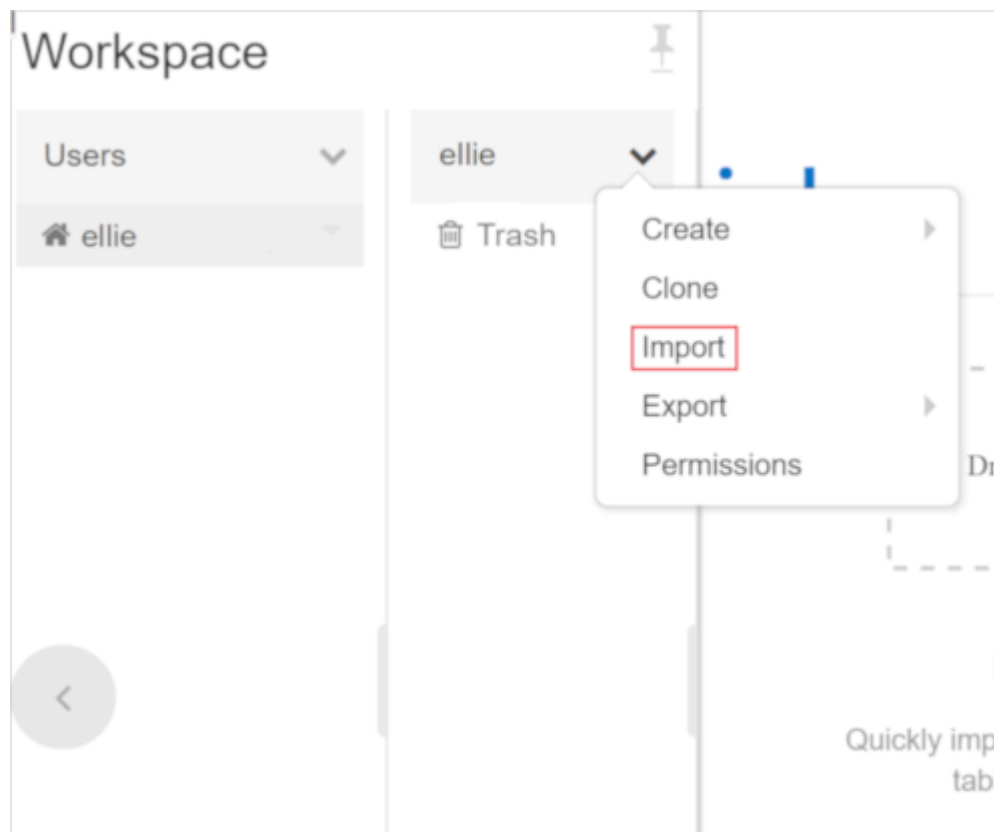
Lab setup and pre-requisites

Before starting this lab, ensure you have successfully completed the setup steps to create your lab environment.

Exercise 1 - Explore Structured Streaming Concepts

Task 1: Clone the Databricks archive

1. If you do not currently have your Azure Databricks workspace open: in the Azure portal, navigate to your deployed Azure Databricks workspace and select **Launch Workspace**.
2. In the left pane, select **Compute**. If you have an existing cluster, ensure that it is running (start it if necessary). If you don't have an existing cluster, create a single-node cluster that uses the latest runtime and **Scala 2.12** or later.
3. When your cluster is running, in the left pane, select **Workspace** > **Users**, and select your username (the entry with the house icon).
4. In the pane that appears, select the arrow next to your name, and select **Import**.



5. In the **Import Notebooks** dialog box, select the URL and paste in the following URL:

```
https://github.com/MicrosoftLearning/DP-203-Data-Engineer/raw/master/Allfiles/microsoft-learning-paths-databricks-notebooks/data-engineering/DBC/10-Structured-Streaming.dbc
```

1. Select **Import**.
2. Select the **10-Structured-Streaming** folder that appears.

Task 2: Complete the notebook

1. Open the **1.Structured-Streaming-Concepts** notebook. Make sure you attach your cluster to the notebook before following the instructions and running the cells within.

Within the notebook, you will:

- Stream data from a file and write it out to a distributed file system
- List active streams
- Stop active streams

Exercise 2 - Work with Time Windows

Task 1: Complete the notebook

1. In your Azure Databricks workspace, open the **10-Structured-Streaming** folder that you imported within your user folder.
2. Open the **2.Time-Windows** notebook. Make sure you attach your cluster to the notebook before following the instructions and running the cells within.

Within the notebook, you will:

- Use sliding windows to aggregate over chunks of data rather than all data
- Apply watermarking to throw away stale old data that you do not have space to keep
- Plot live graphs using `display`

Exercise 3 - Use Structured Streaming with Azure EventHubs

Task 1: Create an event hub

1. In the Azure portal (<https://portal.azure.com>), in the **data-engineering-synapse-xxxxxxx** resource group that contains your Azure resources for this course, open the **eventhubxxxxxxx** Event Hub namespace.
2. Add a new event hub by selecting the **+ Event Hub** button on the toolbar.
3. On the **Create Event Hub** pane, create a new event hub with the following details:
 - **Name:** `databricks-demo-eventhub`
 - **Partition Count:** `2`
 - **Message Retention:** `1`
 - **Capture:** *Off*

Select **Create**.

Create Event Hub

Event Hubs

* Name ⓘ

databricks-demo-eventhub ✓

Partition Count ⓘ

2

Message Retention ⓘ

1

Capture ⓘ

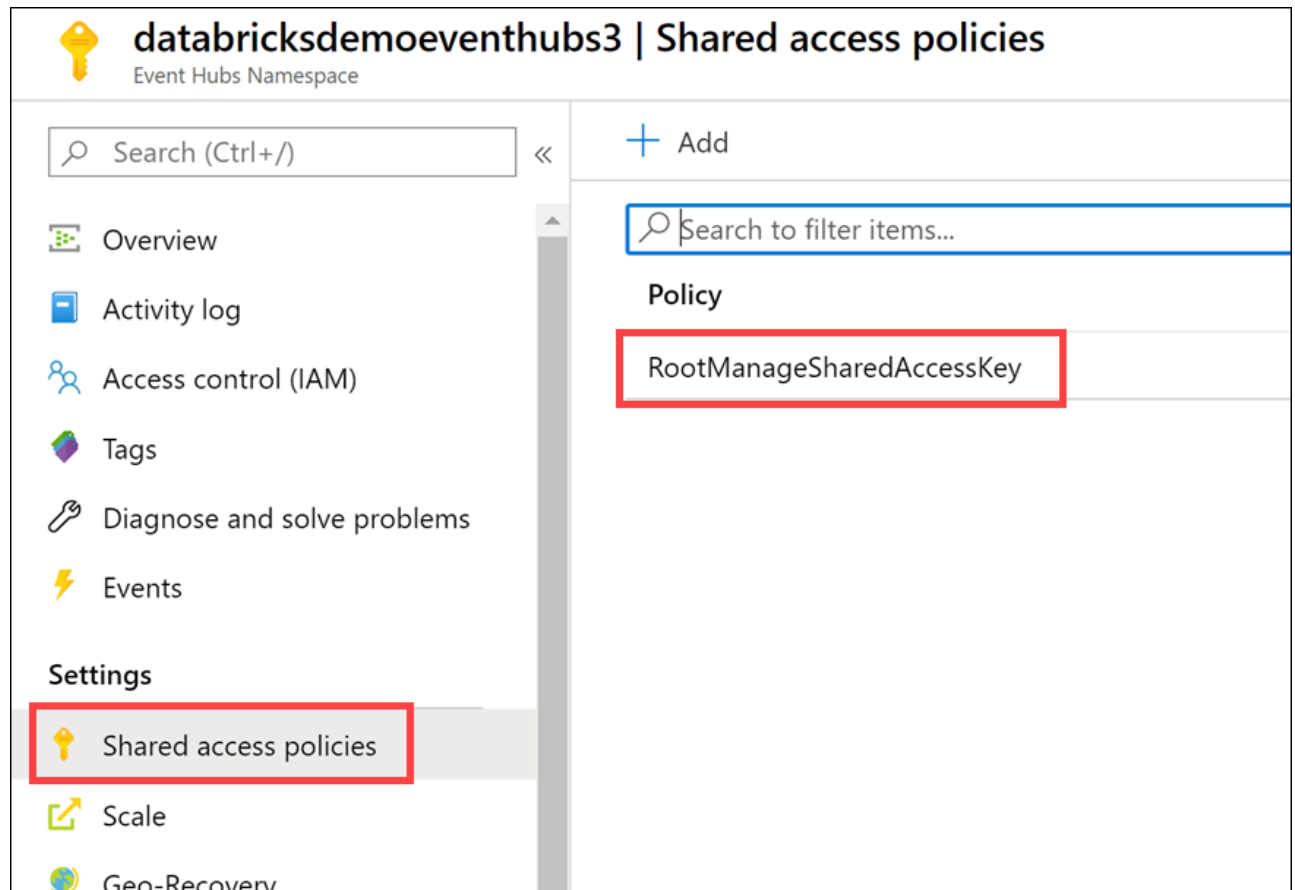
On

Off

Create




Task 2: Copy the connection string primary key for the shared access policy

1. On the left-hand menu in your Event Hubs namespace, select **Shared access policies** under **Settings**, then select the **RootManageSharedAccessKey** policy.



2. Copy the connection string for the primary key by selecting the copy button.

SAS Policy: RootManageSharedAccessKey ×


 Save  Discard  Delete ...

☒ Manage


☐ Send

☐ Listen


Primary key

GuAj2zdcehfXNJXjeeXB6eEOWR4xANcGwra0LaG9N... 


Secondary key

3V5l+g4CMFkw5SPhmHh9uOFWIP2nEe7RfomOJOX... 

Connection string–primary key

Endpoint=sb://databricksdemoeventhubs3.serviceb... 

Connection string–secondary key

Endpoint=sb://databricksdemoeventhubs3.serviceb... 

3. Save the copied primary key to Notepad or another text editor for later reference.

Task 3: Run the notebook

1. Switch back to the browser tab containing your Azure Databricks workspace, and open the **10-Structured-Streaming** folder that you imported within your user folder.
2. Open the **3.Streaming-With-Event-Hubs-Demo** notebook. Make sure you attach your cluster to the notebook before following the instructions and running the cells within.

Within the notebook, you will:

- Connect to Event Hubs and write a stream to your event hub
- Read a stream from your event hub
- Define a schema for the JSON payload and parse the data to display it within a table

Shut down your cluster

1. After you've completed the lab, in the left pane, select **Compute** and select your cluster. Then select **Terminate** to stop the cluster.