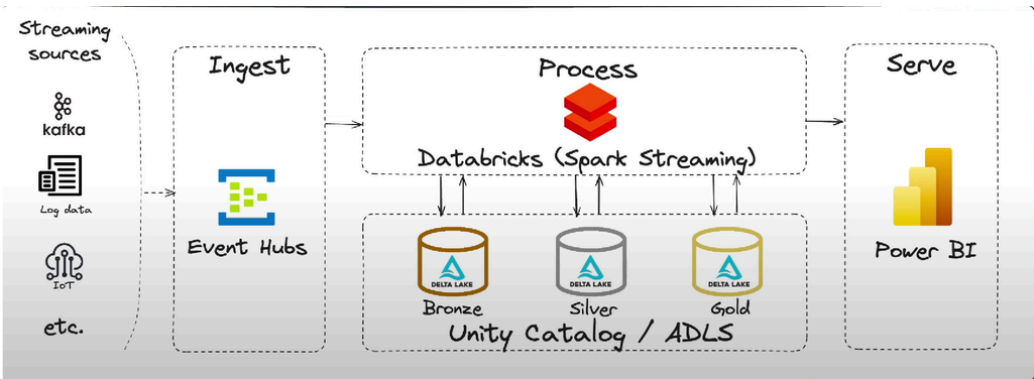


Real-time Data Processing with Azure Databricks (and Event Hubs)



- Data Sources: Streaming data from IoT devices or social media feeds. (Simulated in Event Hubs)
- Ingestion: Azure Event Hubs for capturing real-time data.
- Processing: Azure Databricks for stream processing using Structured Streaming.
- Storage: Processed data stored Azure Data Lake (Delta Format).
- Visualisation: Data visualized using Power BI

Azure Services Required

- Databricks Workspace (Unity Catalog enabled)
- Azure Data Lake Storage (Premium)
- Azure Event Hub (Basic Tier)

Event hubs namespace	eventhub-vidya
Event hubs Instance	eh-demo
Resource group	vidyademo
Azure Databricks service	databrick-vidya
Storage Account	evstorageaccountvidya

- Create an event hubs namespace

Microsoft Azure

eventhub-vidya | Overview

Deployment

Search resources, services, and docs (G+I)

Copilot

W0828878@myscc.ca
ST CLAIR COLLEGE (STCLAIRCON...)

Home >

eventhub-vidya | Overview

Deployment

Search

Delete Cancel Redeploy Download Refresh

Overview

Inputs

Outputs

Template

✓ Your deployment is complete

Deployment name : eventhub-vidya
Subscription : Azure for Students
Resource group : vidyademo

Start time : 29/8/2024, 1:01:52 pm
Correlation ID : a1103902-309c-4de8-a2a5-56ac10581b52

> Deployment details

Next steps

Go to resource

Cost management
Get notified to stay within your budget and prevent unexpected charges on your bill.
[Set up cost alerts >](#)

Microsoft Defender for Cloud
Secure your apps and infrastructure
[Go to Microsoft Defender for Cloud >](#)

Free Microsoft tutorials
[Start learning today >](#)

Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
[Find an Azure expert >](#)

- Create an event hub instance

Microsoft Azure

eventhub-vidya | Event Hubs

Event Hubs Namespace

Search resources, services, and docs (G+I)

Copilot

W0828878@myscc.ca
ST CLAIR COLLEGE (STCLAIRCON...)

Home > eventhub-vidya

eventhub-vidya | Event Hubs

Event Hubs Namespace

Search

+ Event Hub Refresh Give feedback

Search to filter items by name...

Event Hubs 0

Name	Status	Message retention	Partition count
No Event Hubs yet.			

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Generate data (preview)

Events

Settings

Entities

Event Hubs

Monitoring

Automation

Help

<https://portal.azure.com/#/@stclairconnect.onmicrosoft.com/resource/subscriptions/6167d1d0-c1d0-4809-80df-1f7ac899aea/resourceGroups/vidyademo/providers/Microsoft.EventHub/namespaces/eventhub-vidya/eventhubs>

Microsoft Azure

eventhub-vidya | Event Hubs

Event Hubs Namespace

Search resources, services, and docs (G+I)

Copilot

W0828878@myscc.ca
ST CLAIR COLLEGE (STCLAIRCON...)

Home > eventhub-vidya

eventhub-vidya | Event Hubs

Event Hubs Namespace

Search

+ Event Hub Refresh Give feedback

Search to filter items by name...

Event Hubs 1

Name	Status	Message retention	Partition count
eh-demo	Active	1 hour	1

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Generate data (preview)

Events

Settings

Entities

Event Hubs

Monitoring

Automation

Help

Create Event Hub
Successfully created Event Hub eh-demo

- Generate a data in JSON format

Microsoft Azure | Search resources, services, and docs (G+/I) | Copilot | W0828878@myscc.ca | ST. CLAIR COLLEGE (STCLAIRCOL...)

Home > eventhub-vidya | Event Hubs > eh-demo (eventhub-vidya/eh-demo)

eh-demo (eventhub-vidya/eh-demo) | Generate data (preview)

Event Hubs Instance

Search

Generate Data in quick steps to stream into Azure Event Hubs

Settings Give feedback

Overview

Access control (IAM)

Diagnose and solve problems

Settings

Entities

Features

Capture

Generate data (preview)

Process data

Analyze data (preview)

Automation

Help

Select Dataset *

Custom payload

Select Content-Type *

JSON

Repeat send *

1

Enter payload

The content in the Enter payload section is treated as one event. For more information, [click here](#)

```
1 {
2   "temperature": 20,
3   "humidity": 60,
4   "windSpeed": 10,
5   "windDirection": "NW",
6   "precipitation": 0,
7   "conditions": "Partly Cloudy"
8 }
```

Send Cancel

- send it to events

Microsoft Azure | Search resources, services, and docs (G+/I) | Copilot | W0828878@myscc.ca | ST. CLAIR COLLEGE (STCLAIRCOL...)

Home > eventhub-vidya | Event Hubs > eh-demo (eventhub-vidya/eh-demo)

eh-demo (eventhub-vidya/eh-demo) | Generate data (preview)

Event Hubs Instance

Search

Generate Data in quick steps to stream into Azure Event Hubs

Settings Give feedback

Overview

Access control (IAM)

Diagnose and solve problems

Settings

Entities

Features

Capture

Generate data (preview)

Process data

Analyze data (preview)

Automation

Help

Select Dataset *

Custom payload

Select Content-Type *

JSON

Repeat send *

1

Enter payload

The content in the Enter payload section is treated as one event. For more information, [click here](#)

```
1 {
2   "temperature": 30,
3   "humidity": 60,
4   "windSpeed": 10,
5   "windDirection": "NW",
6   "precipitation": 0,
7   "conditions": "Partly Cloudy"
8 }
```

Send Cancel

added one more record to the table

Microsoft Azure | Search resources, services, and docs (G+/I) | Copilot | W0828878@myscc.ca | ST. CLAIR COLLEGE (STCLAIRCOL...)

Home > eventhub-vidya | Event Hubs > eh-demo (eventhub-vidya/eh-demo)

eh-demo (eventhub-vidya/eh-demo) | Generate data (preview)

Event Hubs Instance

Search

Generate Data in quick steps to stream into Azure Event Hubs

Settings Give feedback

Overview

Access control (IAM)

Diagnose and solve problems

Settings

Entities

Features

Capture

Generate data (preview)

Process data

Analyze data (preview)

Automation

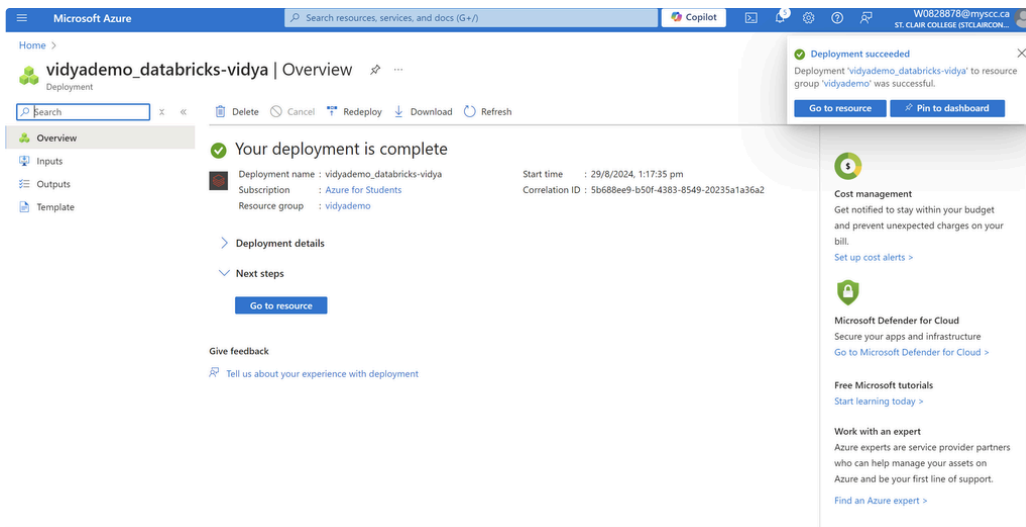
Help

View Events already sent to this Event Hub

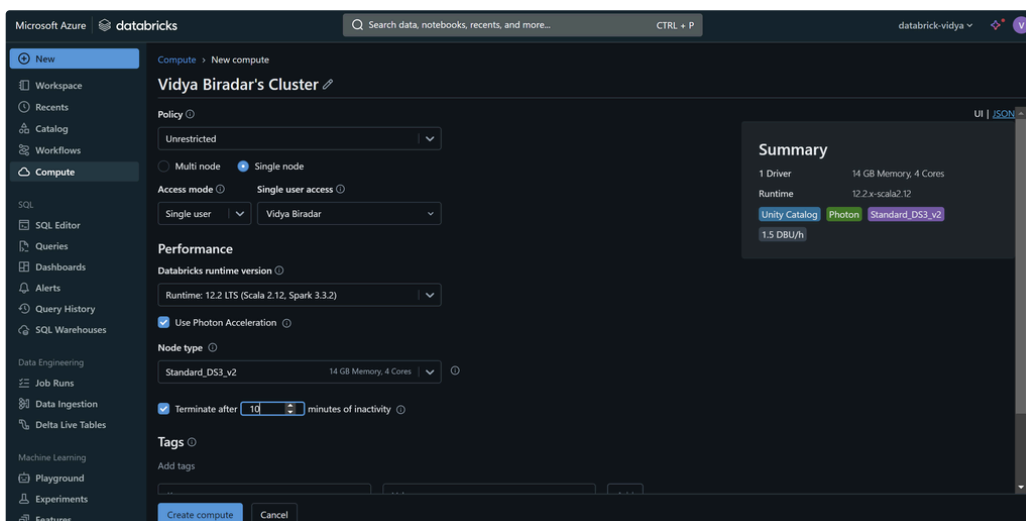
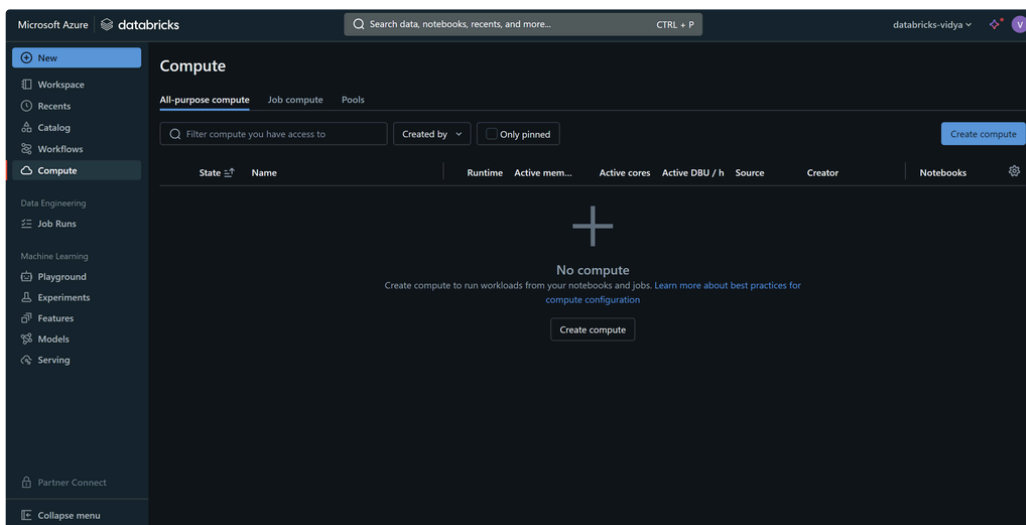
Showing the last 15 events starting from Thu, 29 Aug 2024 17:13:00 GMT

Event Body	Offset	Sequence Number	Partition ID	Enqueued Time	Content Type
{ "temperature": 30, "humidity": 60, "windSpeed": 10, "windDirection": "NW", "precipitation": 0, "conditions": "Partly Cloudy" }	272	1	0	Thu, 29 Aug 2024 17:14:13 GMT	application/json
{ "temperature": 20, "humidity": 60, "windSpeed": 10, "windDirection": "NW", "precipitation": 0, "conditions": "Partly Cloudy" }	0	0	0	Thu, 29 Aug 2024 17:13:00 GMT	application/json

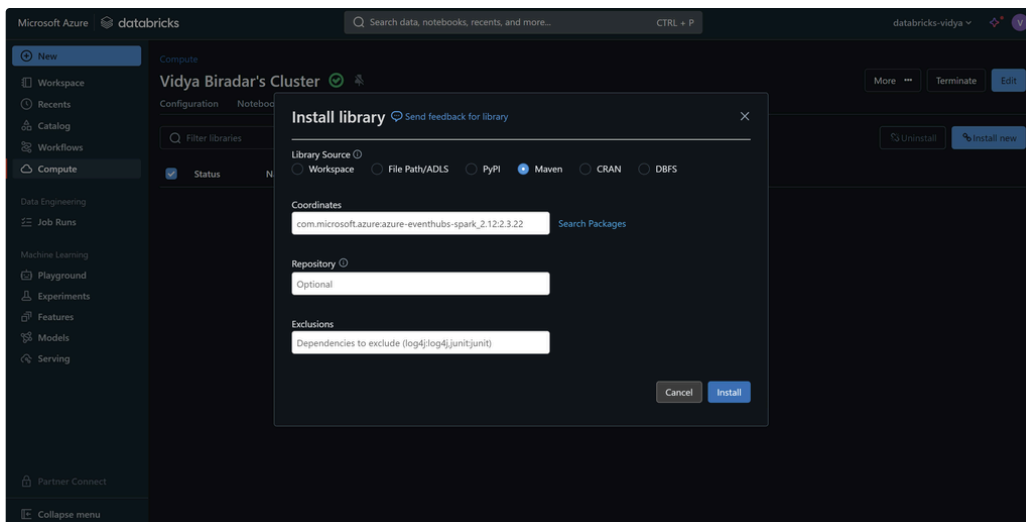
- create a databricks



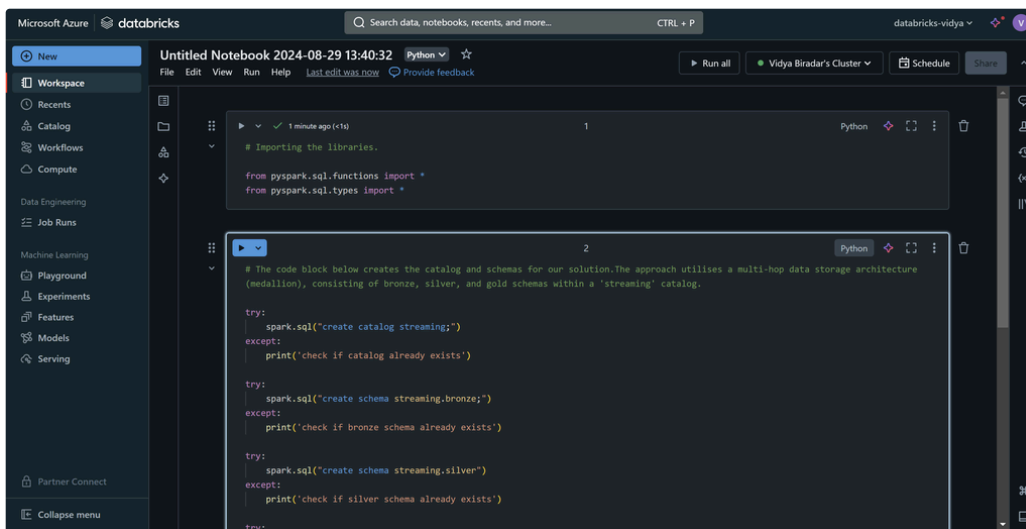
- Launch the data bricks and create a compute



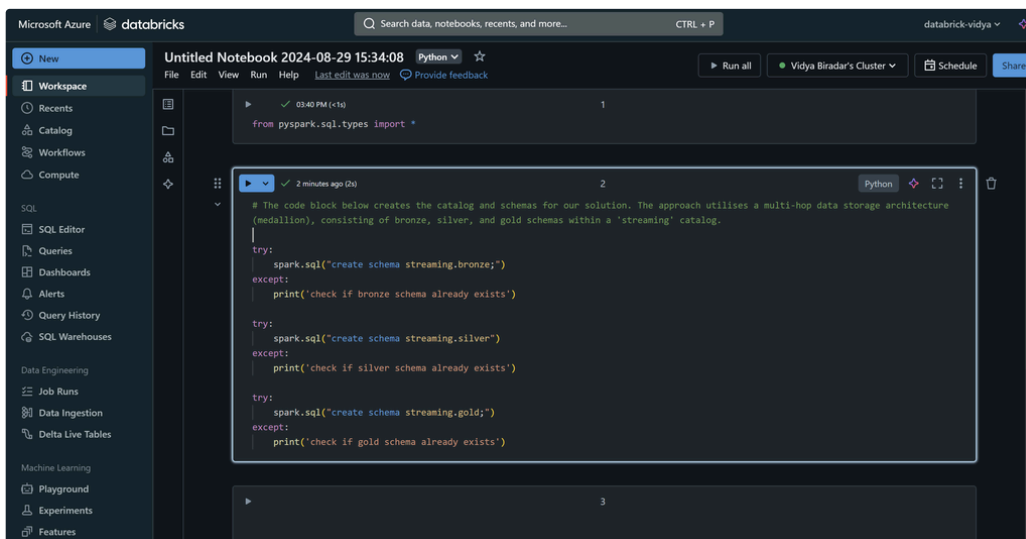
- After creating compute cluster navigate to library and install required library



- In the workspace create a notebook and write the python code



- import a library and create a catalog followed by creating schemas bronze,silver and gold folders



databricks

Untitled Notebook 2024-08-29 15:34:08 (Python)

Import Notebook

```
# Importing the libraries.
from pyspark.sql.functions import *
from pyspark.sql.types import *
```

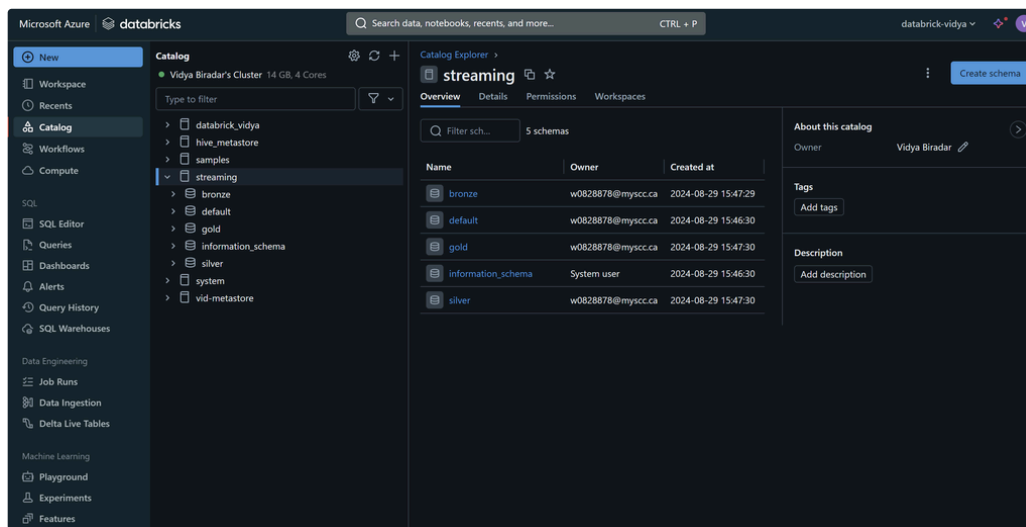
```
# The code block below creates the catalog and schemas for our solution. The approach utilises a multi-hop data storage architecture (medallion), consisting of bronze, silver, and gold schemas.

try:
    spark.sql("create schema streaming.bronze;")
except:
    print('check if bronze schema already exists')

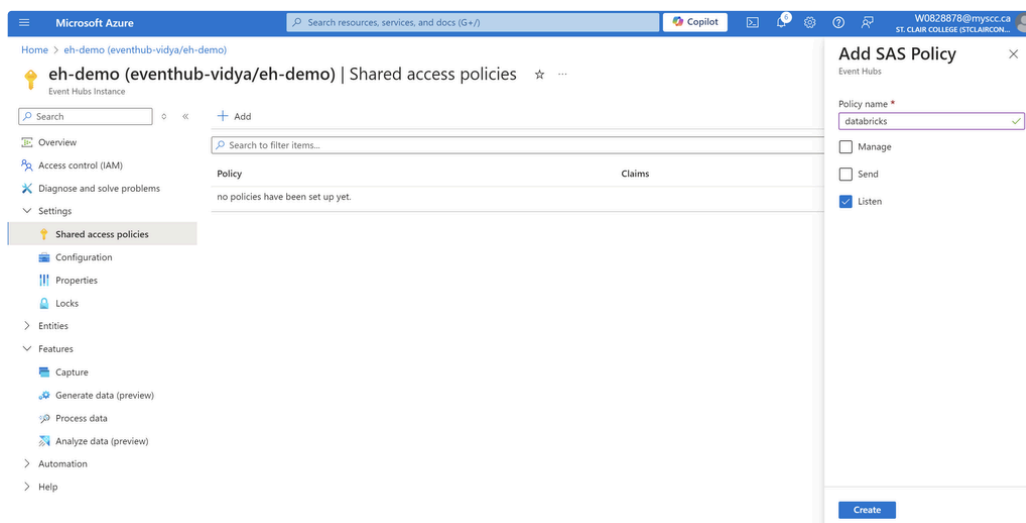
try:
    spark.sql("create schema streaming.silver;")
except:
    print('check if silver schema already exists')

try:
    spark.sql("create schema streaming.gold;")
except:
    print('check if gold schema already exists')
```

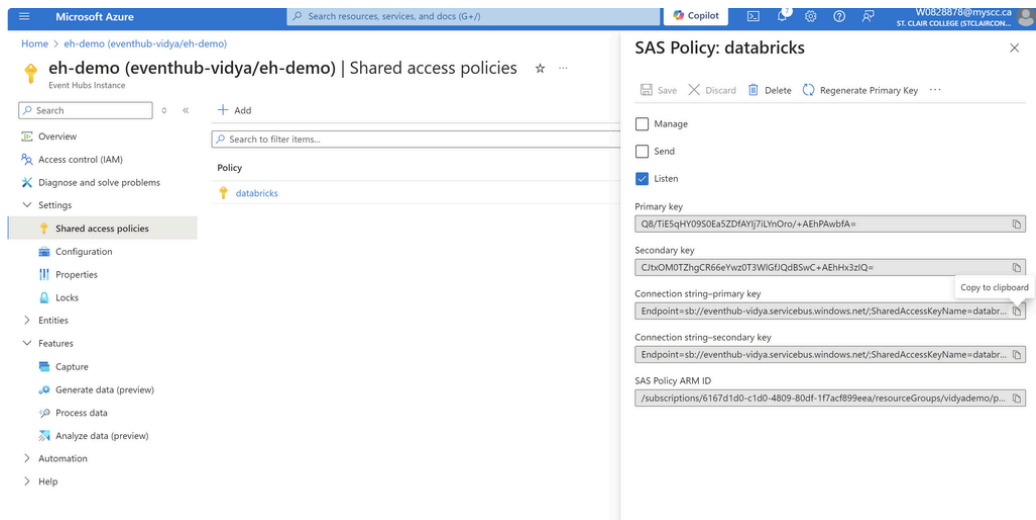
- We can see its added in catalog



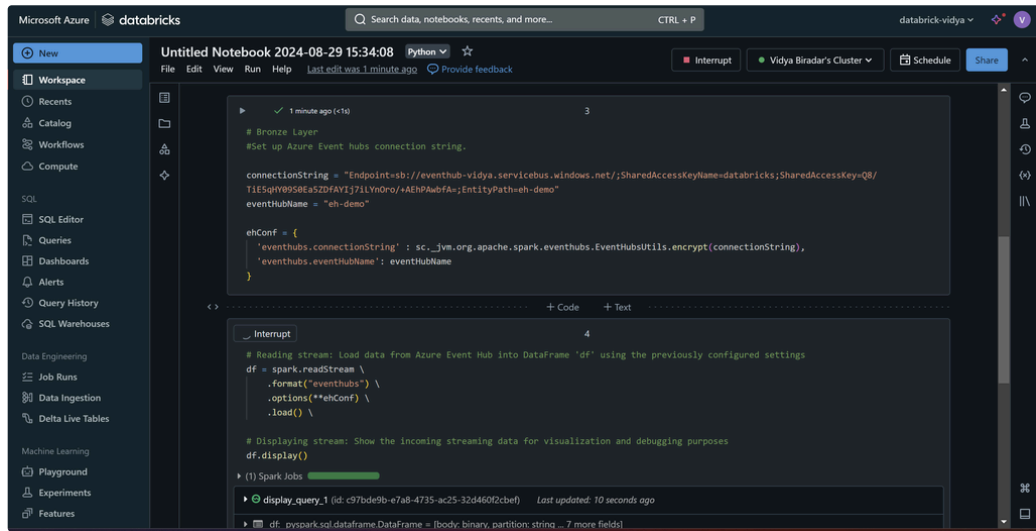
- Create a shared access policy



- Copy the primary key



- In the **Bronze layer** set up Azure event hub connecting to a string by adding a key and event hub name



- Generate one more data



Microsoft Azure databricks

Untitled Notebook 2024-08-29 15:34:08 (Python)

display_query_2 (id: 2fe859bf-5d56-4ba7-b798-fb9131f100ce3) Last updated: 2 hours ago

	body	partition	offset	sequenceNumber	enqueuedTime
1	ewogICAgbnRlbXBldXJpogMjAgICAgIAAhaVtaWRpdHkOIAZMcWkiAgICj3aW5kU3BZWQOIAAMCwkiAgICj3aW5kRGlzZWNO...	0	816	3	2024-08-29T20:08:11.663+00
2	ewogICAgbnRlbXBldXJpogMjAgICAgIAAhaVtaWRpdHkOIAZMcWkiAgICj3aW5kU3BZWQOIAAMCwkiAgICj3aW5kRGlzZWNO...	0	1088	4	2024-08-29T20:08:49.745+00
3	ewogICAgbnRlbXBldXJpogMjAgICAgIAAhaVtaWRpdHkOIAZMcWkiAgICj3aW5kU3BZWQOIAAMCwkiAgICj3aW5kRGlzZWNO...	0	1360	5	2024-08-29T20:19:08.478+00
4	ewogICAgbnRlbXBldXJpogMjAgICAgIAAhaVtaWRpdHkOIAZMcWkiAgICj3aW5kU3BZWQOIAAMCwkiAgICj3aW5kRGlzZWNO...	0	1632	6	2024-08-29T20:26:25.685+00
5	ewogICAgbnRlbXBldXJpogMjAgICAgIAAhaVtaWRpdHkOIAZMcWkiAgICj3aW5kU3BZWQOIAAMCwkiAgICj3aW5kRGlzZWNO...	0	1904	7	2024-08-29T20:51:22.725+00

6 rows

```
# Silver Layer

# Defining the schema for the JSON object

json_schema = StructType([
    StructField("temperature", IntegerType()),
    StructField("humidity", IntegerType()),
    StructField("windSpeed", IntegerType()),
    StructField("windDirection", StringType()),
    StructField("precipitation", IntegerType()),
    StructField("conditions", StringType())
])
```

- Reading, Transforming and writing the data into silver layer

Microsoft Azure databricks

Untitled Notebook 2024-08-29 15:34:08 (Python)

Interrupt

```
# Reading and Transforming: Load streaming data from the 'streaming.bronze.weather' Delta table, cast 'body' to string, parse JSON,
and select specific fields
df = spark.readStream\
    .format("delta")\
    .table("streaming.bronze.weather")\
    .withColumn("body", col("body").cast("string"))\
    .withColumn("body", from_json(col("body"), json_schema))\
    .select("body.temperature", "body.humidity", "body.windSpeed", "body.windDirection", "body.precipitation", "body.conditions", col
("enqueuedTime").alias("timestamp"))

# Displaying stream: Visualize the transformed data in the Dataframe for verification and analysis
df.display()

# Writing stream: Save the transformed data to the 'streaming.silver.weather' Delta table in 'append' mode with checkpointing for data
reliability
df.writeStream\
    .option("checkpointinterval", "/mnt/streaming/silver/weather")\
    .outputMode("append")\
    .format("delta")\
    .toTable("streaming.silver.weather")
```

3 Spark Jobs

display_query_3 (id: 4f2840cb-f12d-46af-87f7-f7f29dc7e8a) Last updated: 5 seconds ago

1a369c02-5895-4044-bcac-00e6d168e936 Last updated: 0 seconds ago

	body	partition	offset	sequenceNumber	enqueuedTime
--	------	-----------	--------	----------------	--------------

	temperature	humidity	windSpeed	windDirection	precipitation	conditions	timestamp
1	22	60	10	NW	0	Partly Cloudy	2024-08-29T20:08:49.7...
2	20	60	10	SW	0	Partly Cloudy	2024-08-29T20:19:08.4...
3	20	60	10	NW	0	Partly Cloudy	2024-08-29T20:08:11.6...

Untitled Notebook 2024-08-29 15:34:08 (Python)

Import Notebook

```

# Reading and Transforming: Load streaming data from the 'streaming.bronze.weather' Delta table, cast 'body' to string, parse JSON, and select specific fields
df = spark.readStream\
    .format("delta")\
    .table("streaming.bronze.weather")\
    .withColumn("body", col("body").cast("string"))\
    .withColumn("body", from_json(col("body"), json_schema))\
    .select("body.temperature", "body.humidity", "body.windSpeed", "body.windDirection", "body.precipitation", "body.conditions", col("enqueuedTime").alias("timestamp"))

# Displaying stream: Visualize the transformed data in the DataFrame for verification and analysis
df.display()

# Writing stream: Save the transformed data to the 'streaming.silver.weather' Delta table in 'append' mode with checkpointing for data reliability
df.writeStream\
    .option("checkpointLocation", "/mnt/streaming/silver/weather")\
    .outputMode("append")\
    .format("delta")\
    .toTable("streaming.silver.weather")

```

display_query_3 (id: 4f2840cb-f12d-46af-87f7-47f2f9dc7e8a) Last updated: 2 hours ago

1a369c02-5895-4044-bcac-00e6d168e936 Last updated: 2 hours ago

	temperature	humidity	windSpeed	windDirection	precipitation	conditions	timestamp
1	22	60	10	NW	0	Partly Cloudy	2024-08-29T20:08:49.745+00...
2	20	60	10	SW	0	Partly Cloudy	2024-08-29T20:19:08.478+00...
3	20	60	10	NW	0	Partly Cloudy	2024-08-29T20:08:11.663+00...
4	18	60	10	SW	0	Partly Cloudy	2024-08-29T20:26:25.685+00...
5	25	30	40	SW	0	Partly Cloudy	2024-08-29T20:51:22.725+00...

- In the **Gold layer** aggregating stream data and writing to the gold layer

New

Untitled Notebook 2024-08-29 15:34:08 Python

Interrupt Vidya Biradar's Cluster Schedule Share

Workspace

Reents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

Machine Learning

Playground

Experiments

Features

Interrupt

7

Python

```

# Gold Layer

# Aggregating Stream: Read from 'streaming.silver.weather', apply watermarking and windowing, and calculate average weather metrics
df = spark.readStream\
    .format("delta")\
    .table("streaming.silver.weather")\
    .withWatermark("timestamp", "5 minutes")\
    .groupBy(window("timestamp", "5 minutes"))\
    .agg(avg("temperature").alias("temperature"), avg("humidity").alias("humidity"), avg("windSpeed").alias("windSpeed"), avg("precipitation").alias("precipitation"))\
    .select("window.start", "window.end", "temperature", "humidity", "windSpeed", "precipitation")

# Displaying Aggregated Stream: Visualize aggregated data for insights into weather trends
df.display()

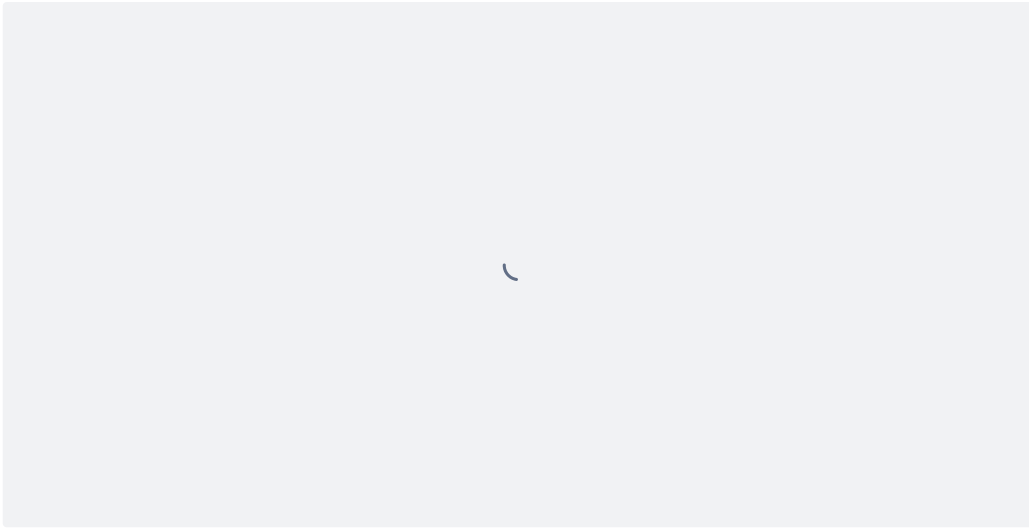
# Writing Aggregated Stream: Store the aggregated data in 'streaming.gold.weather_aggregated' with checkpointing for data integrity
df.writeStream\
    .option("checkpointLocation", "/mnt/streaming/weather_summary")\
    .outputMode("append")\
    .format("delta")\
    .toTable("streaming.gold.weather_summary")

```

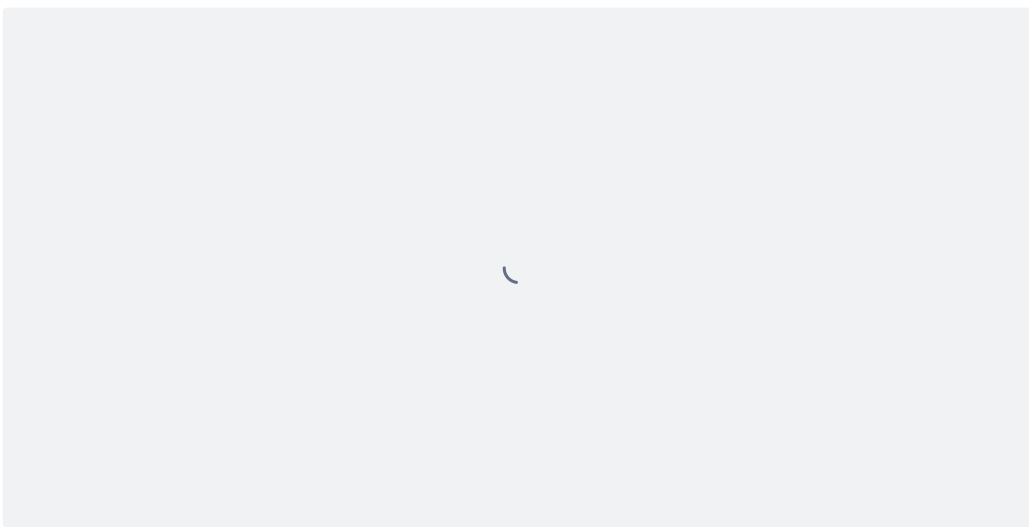
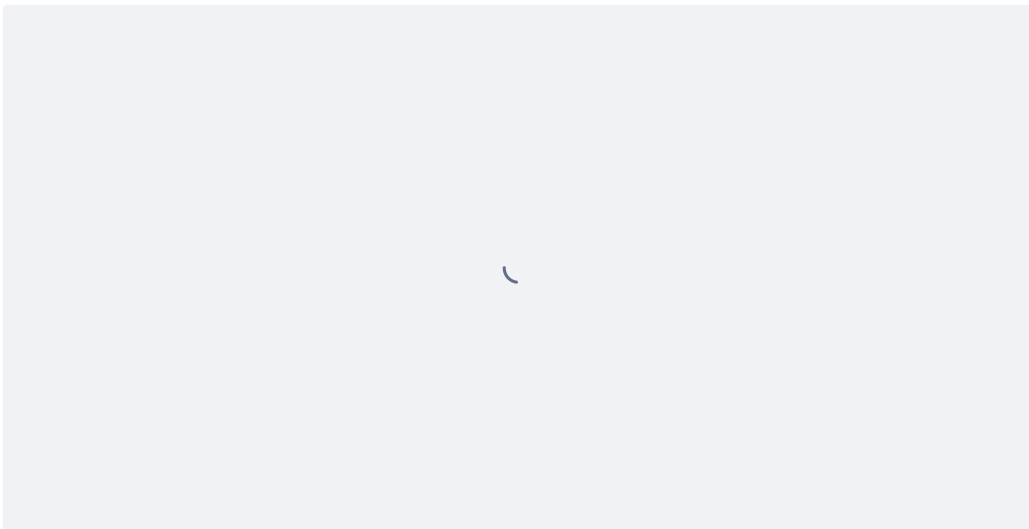
(3) Spark Jobs

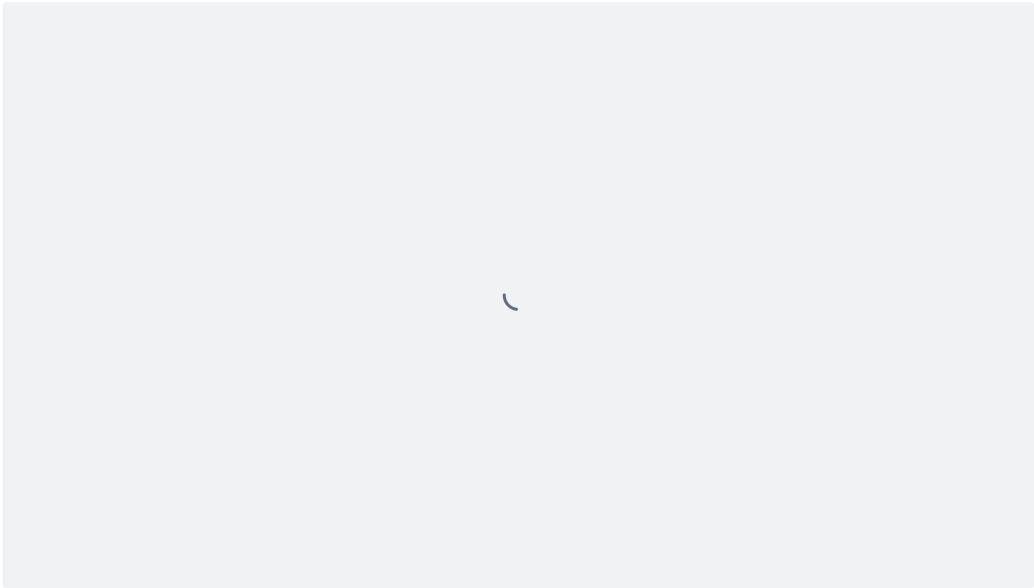
Saa605c2-44e5-4a78-b2b6-f2a7ff6d780b Last updated: 20 seconds ago

display_query_4 (id: ab6c7020-2c1e-4338-9274-406a01b02a7d) Last updated: 5 seconds ago

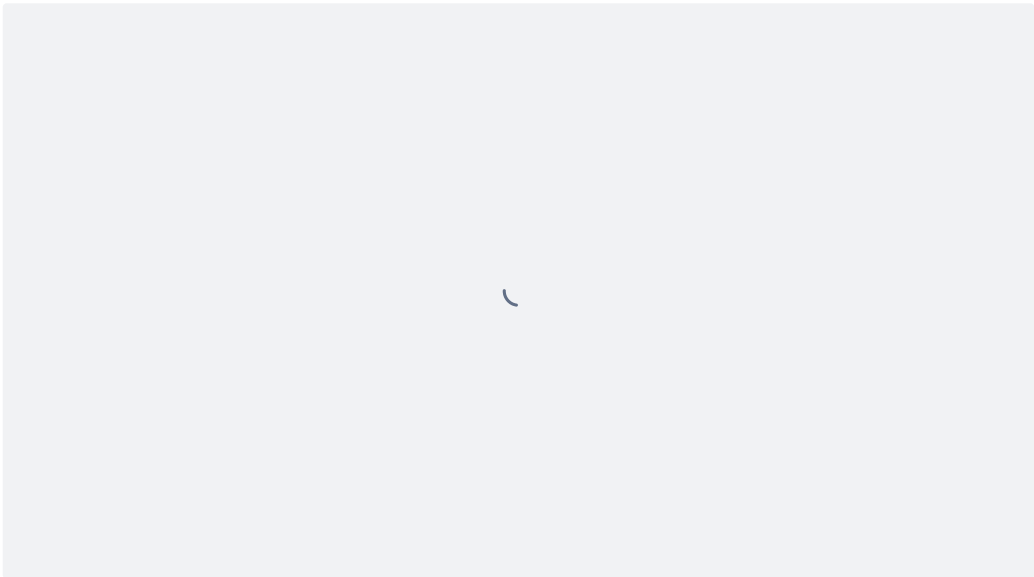


- Using the partner connect option select power BI option for visualization





- Dashboard



Error Encountered

To avoid high-cost	Turn of cluster after use or pause it
Unity cluster will not appear DB worspace	So while creating DB select premium
problem in create a catalog and schema	Enable a unity catalog and create a metastore