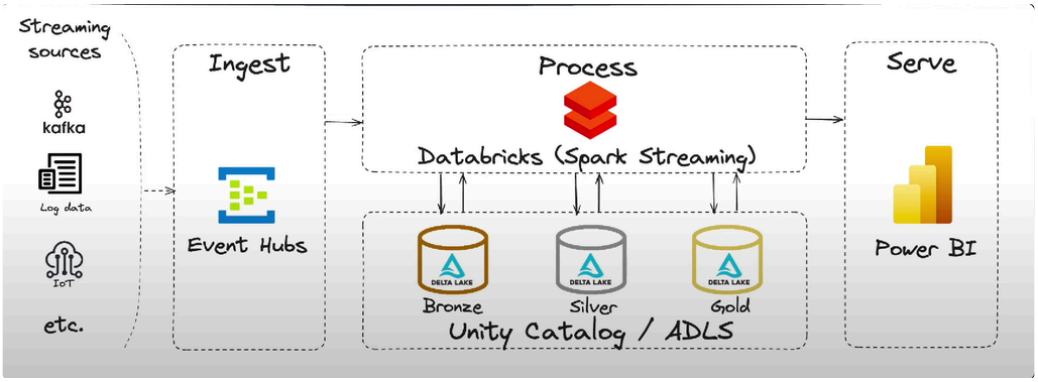
Real-time Data Processing with Azure Databricks (and Event Hubs)



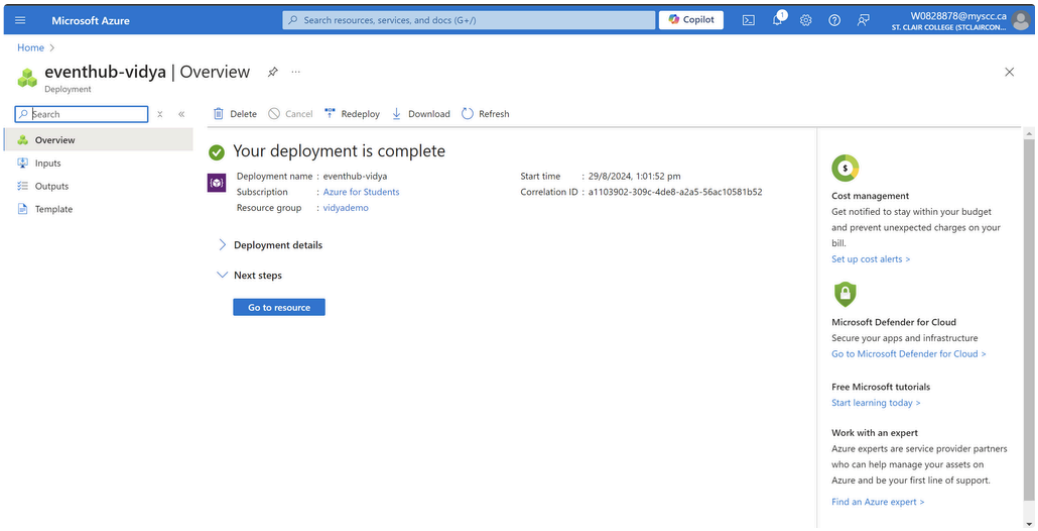
- Data Sources: Streaming data from IoT devices or social media feeds. (Simulated in Event Hubs)
- Ingestion: Azure Event Hubs for capturing real-time data.
- Processing: Azure Databricks for stream processing using Structured Streaming.
- Storage: Processed data stored Azure Data Lake (Delta Format).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

Azure Services Required

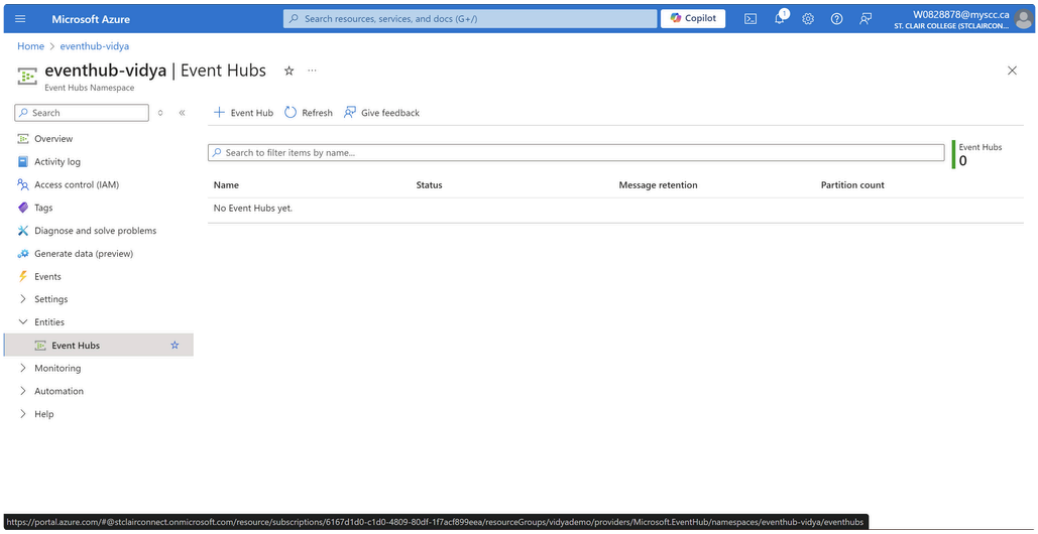
- Databricks Workspace (Unity Catalog enabled)
- Azure Data Lake Storage (Premium)
- Azure Event Hub (Basic Tier)

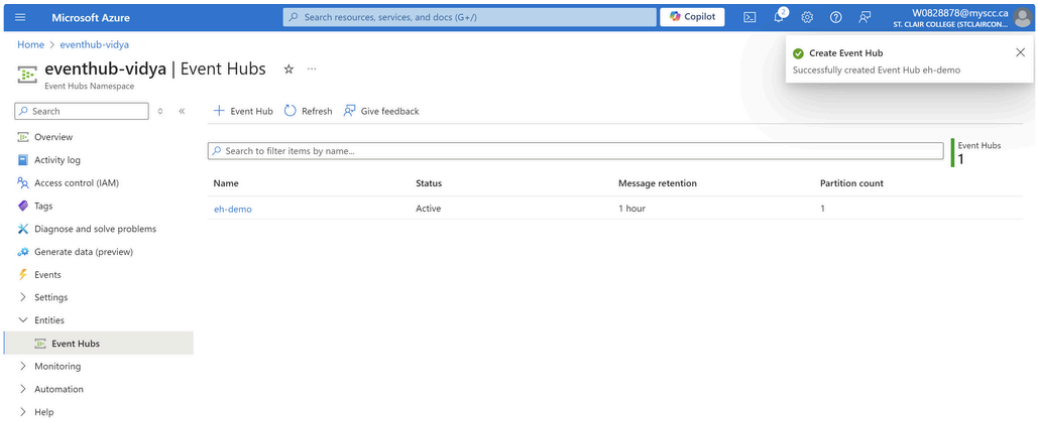
Event hubs namespace	eventhub-vidya
Event hubs Instance	eh-demo
Resource group	vidyademo
Azure Databricks service	databrick-vidya
Storage Account	evstorageaccountvidya

- Create an event hubs namespace

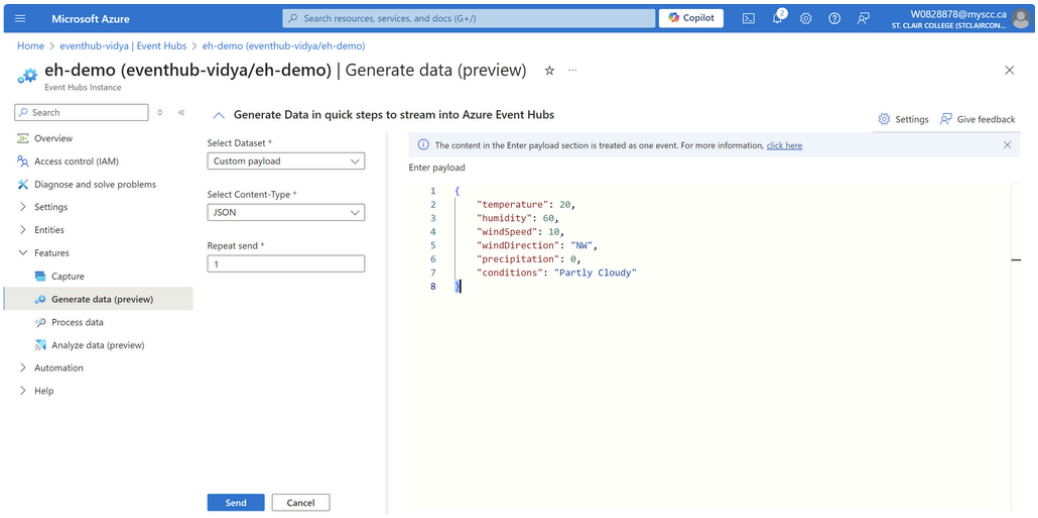


- Create an event hub instance

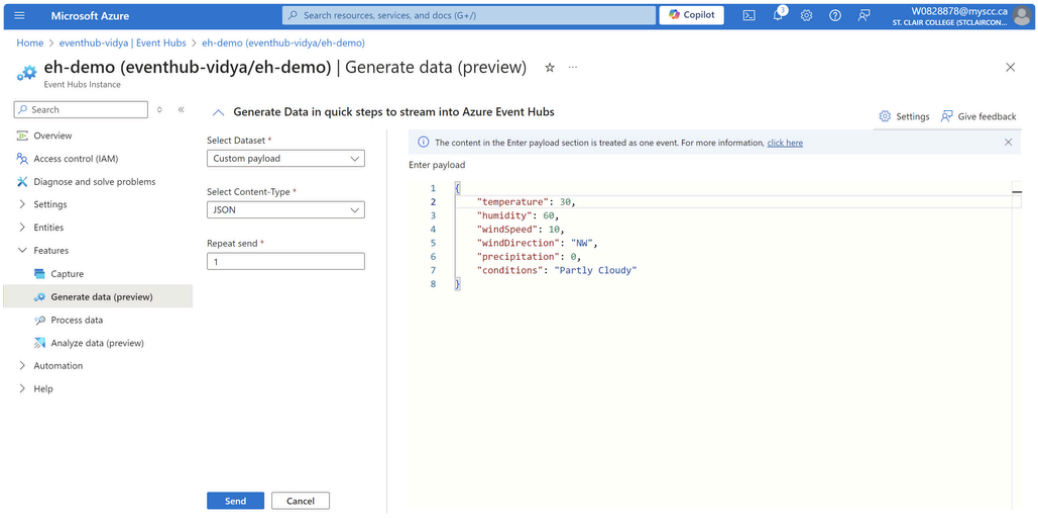




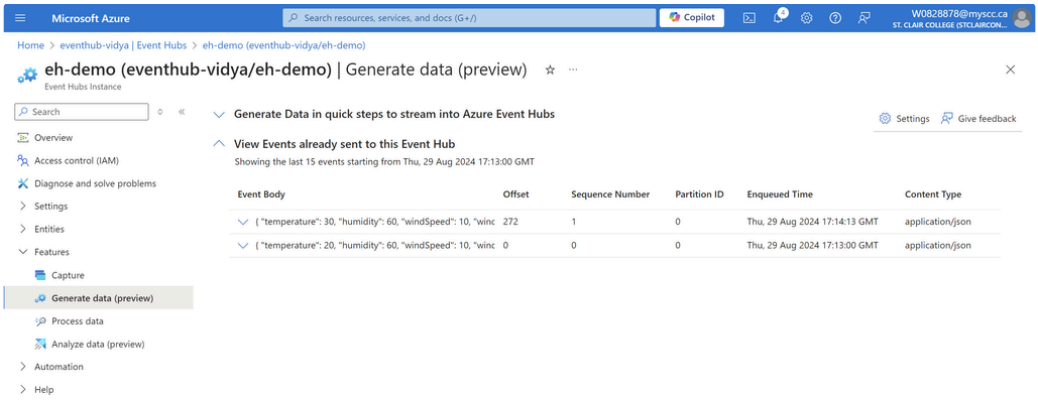
- Generate a data in JSON format



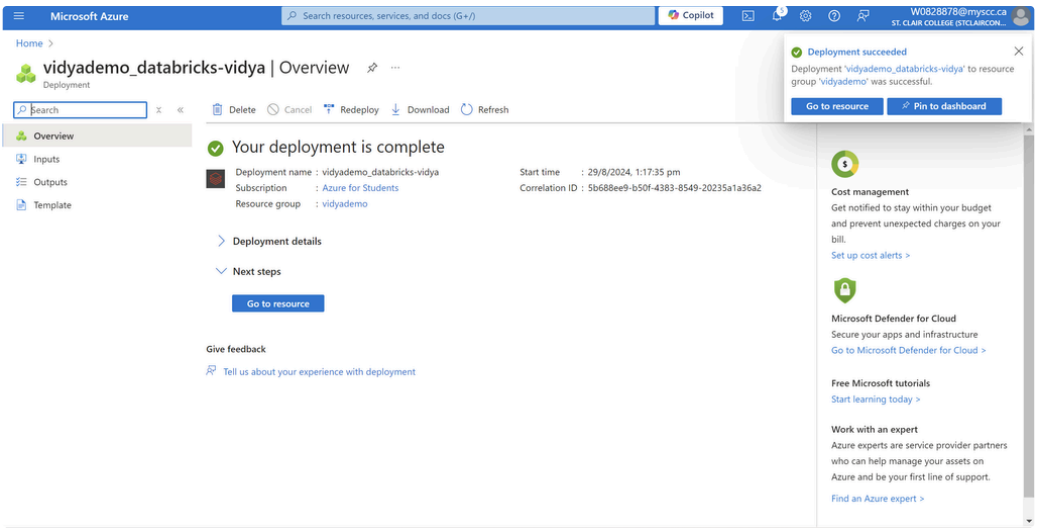
- send it to events



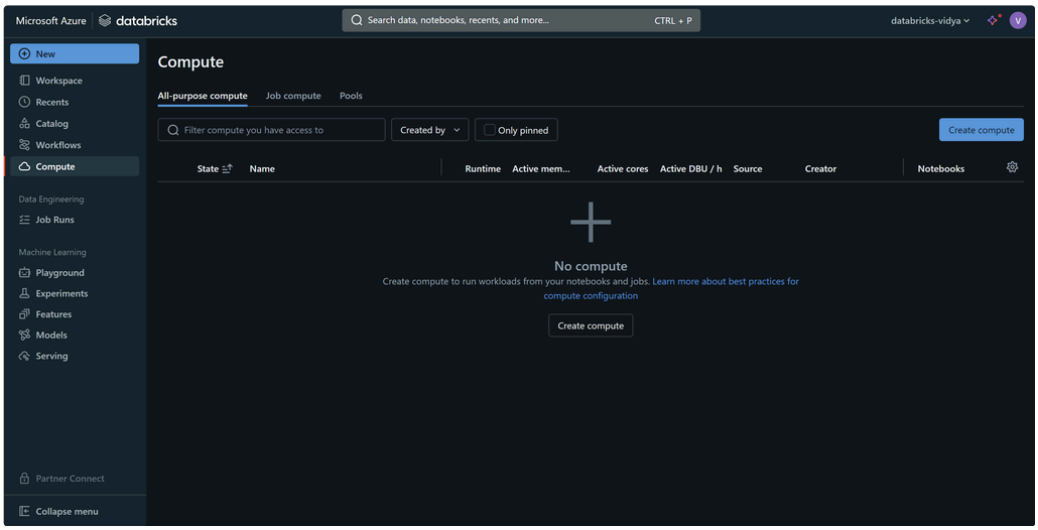
added one more record to the table

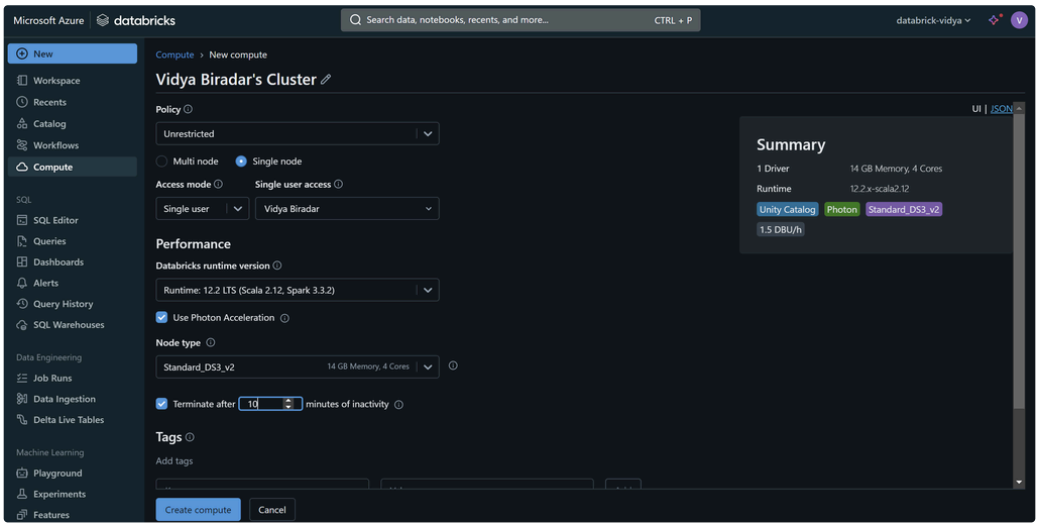


- create a databricks

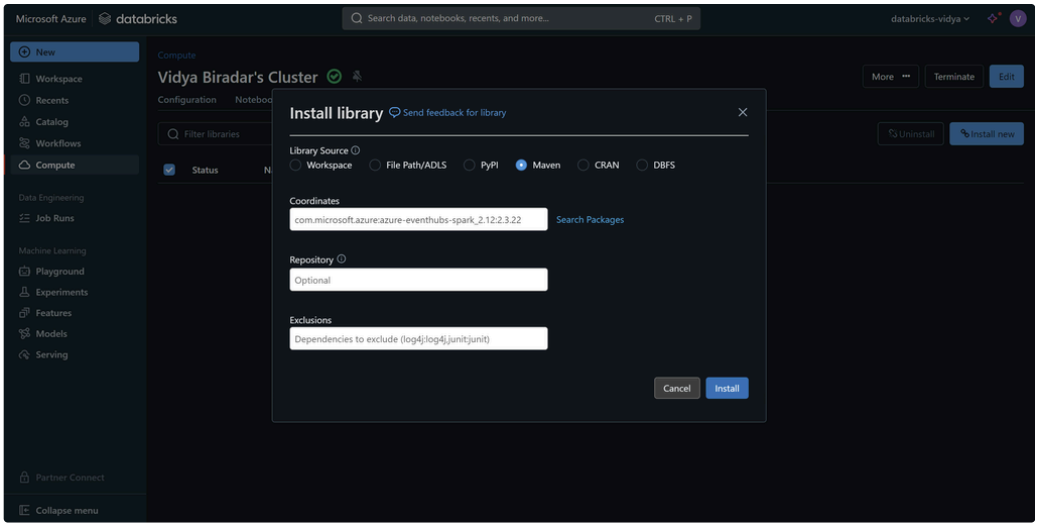


- Launch the data bricks and create a compute

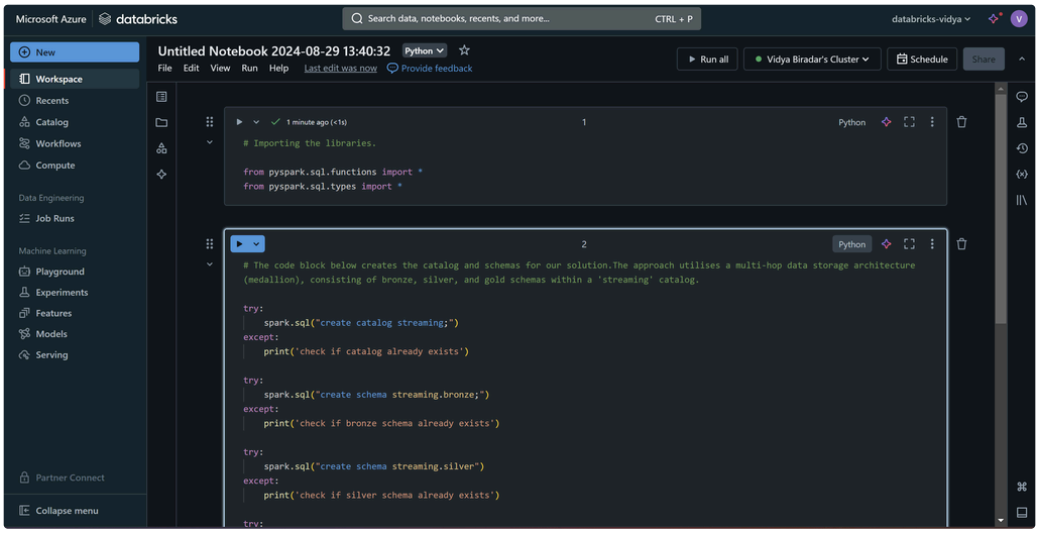




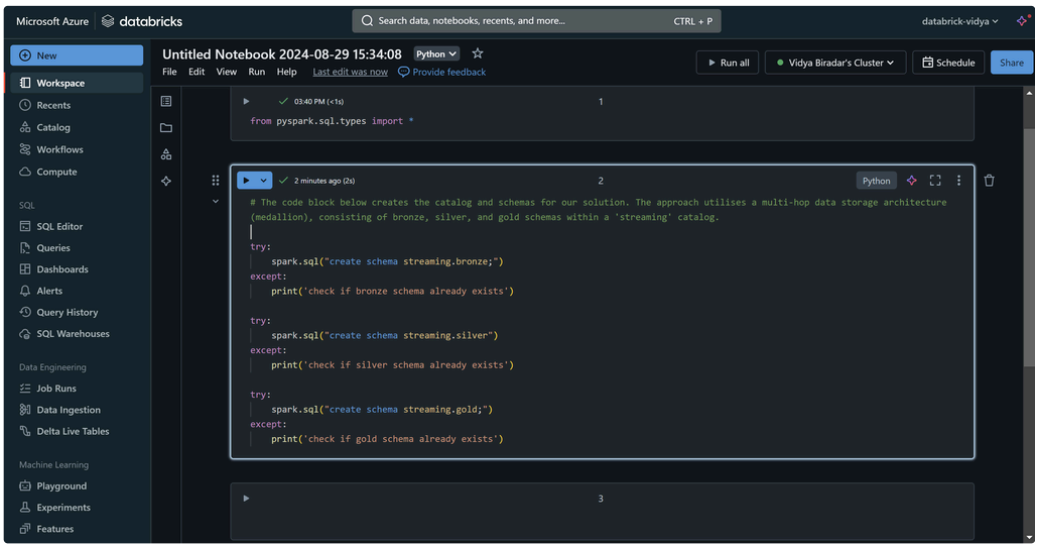
- After creating compute cluster navigate to library and install required library

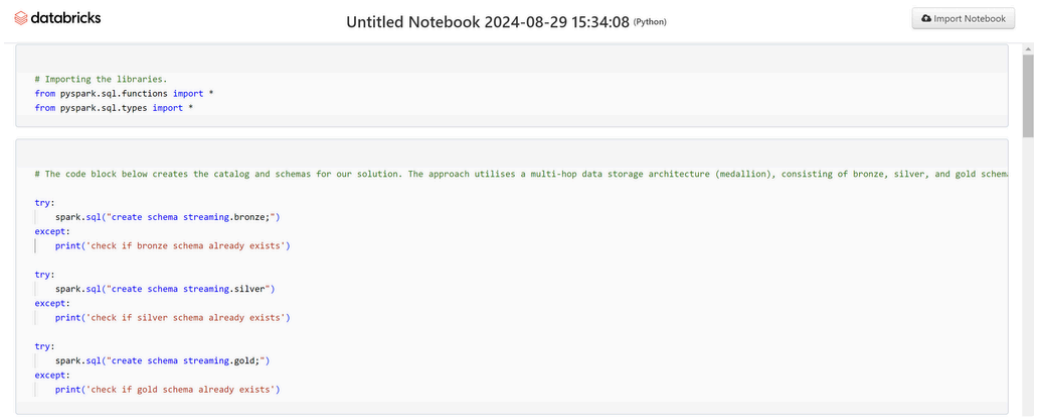


- In the workspace create a notebook and write the python code

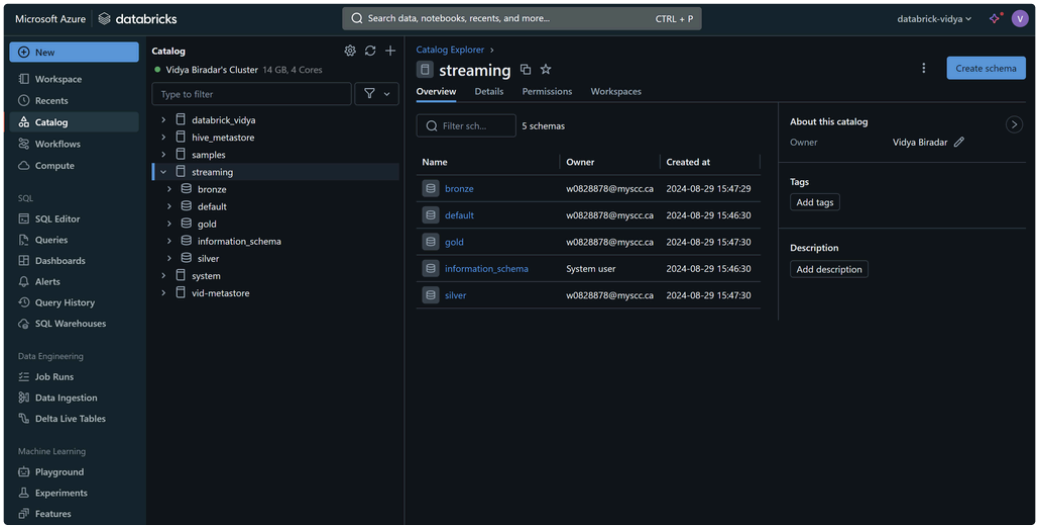


- import a library and create a catalog followed by creating schemas bronze,silver and gold folders

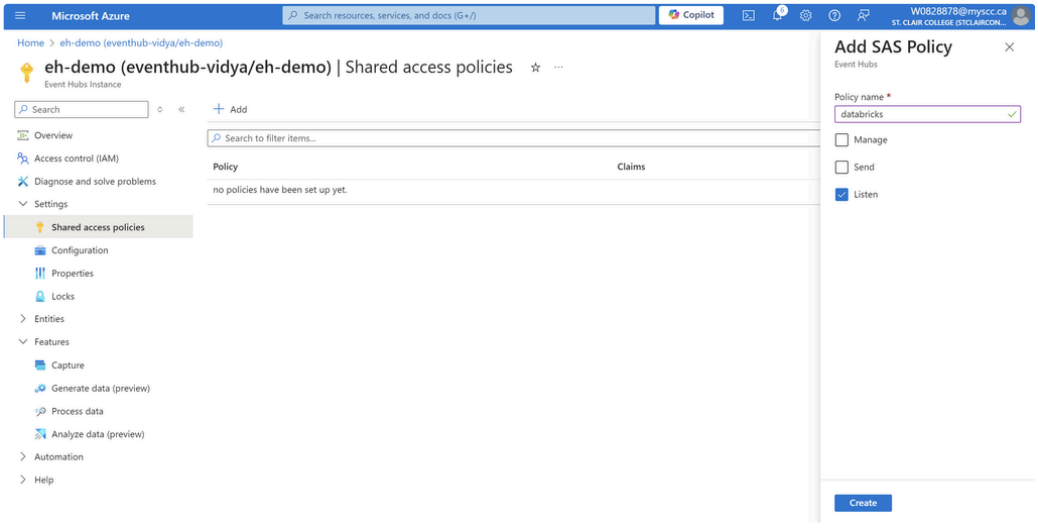




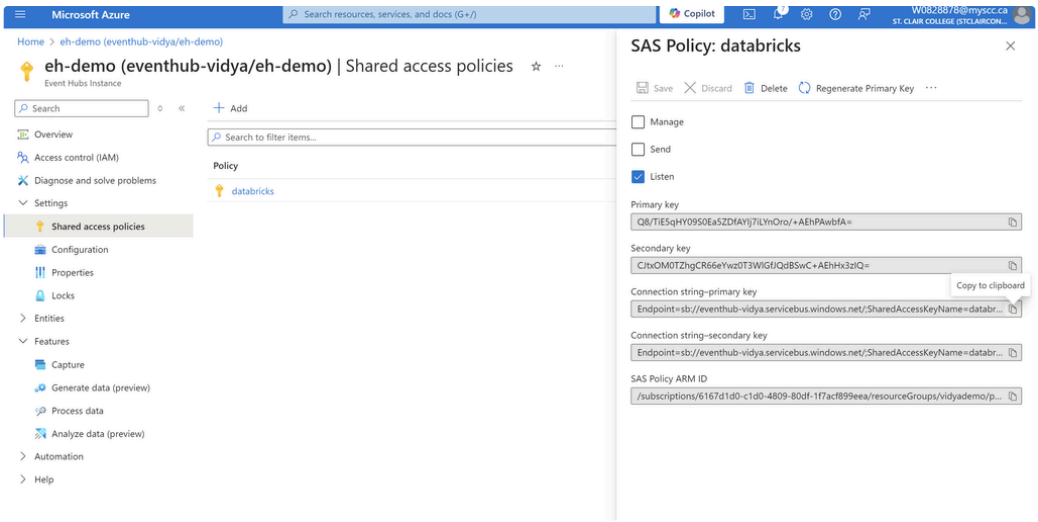
- We can see its added in catalog



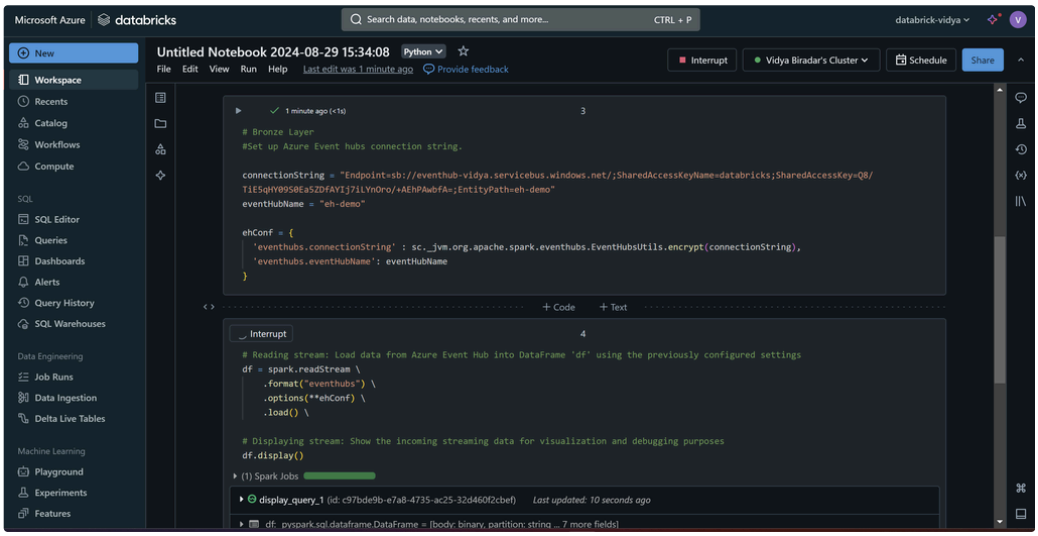
- Create a shared access policy



- 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

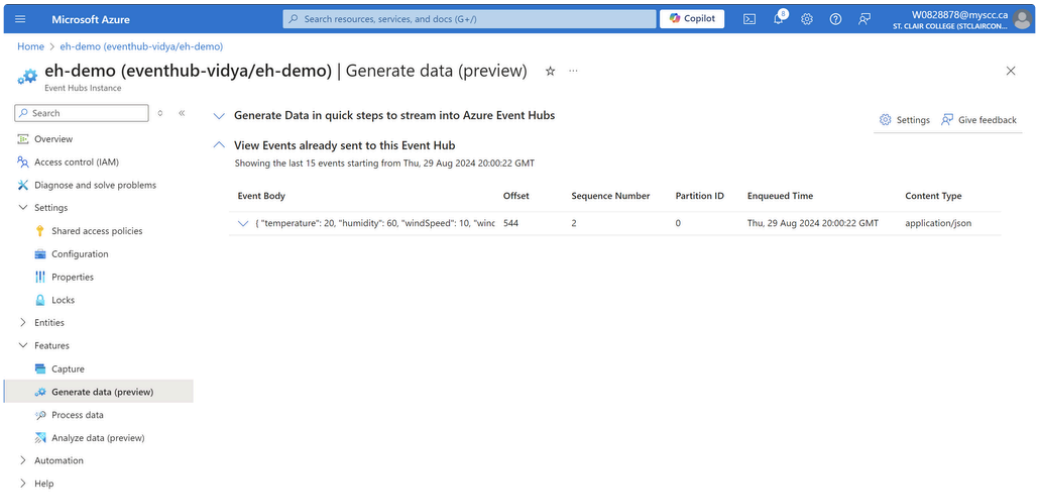


```
# Bronze Layer
#Set up Azure Event hubs connection string.

connectionString = "Endpoint=sb://eventhub-vidya.servicebus.windows.net/;SharedAccessKeyName=databricks;SharedAccessKey=Q8/TiE5qHY09S0Ea5ZDfAYIj7iLYnOro/+AEhPAwbfA=;EntityPath=eh-demo"
eventHubName = "eh-demo"

ehConf = {
  'eventhubs.connectionString' : sc._jvm.org.apache.spark.eventhubs.EventHubsUtils.encrypt(connectionString),
  'eventhubs.eventHubName': eventHubName
}
```

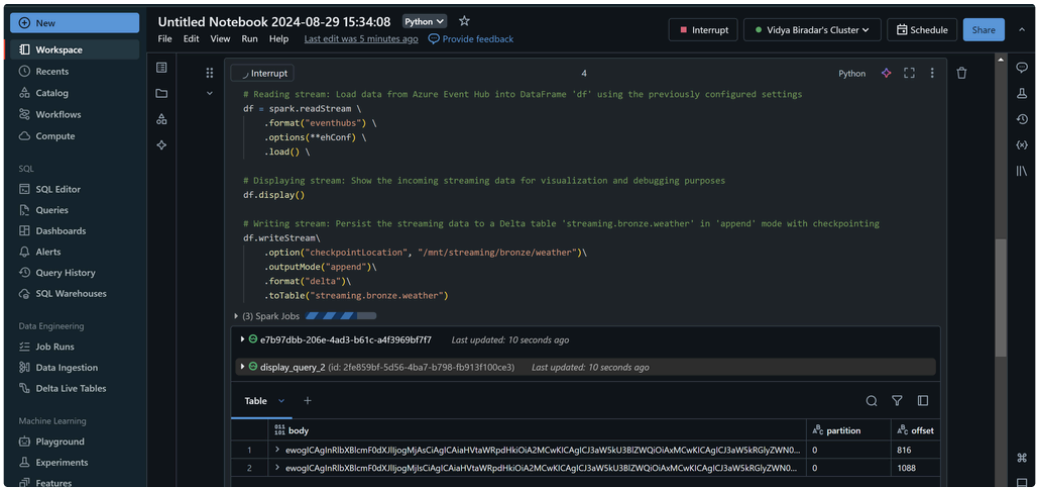
```
# Reading stream: Load data from Azure Event Hub into DataFrame 'df' using the previously configured settings
df = spark.readStream \
    .format("eventhubs") \
    .options(**ehConf) \
    .load() \

# Displaying stream: Show the incoming streaming data for visualization and debugging purposes
df.display()

# Writing stream: Persist the streaming data to a Delta table 'streaming.bronze.weather' in 'append' mode with checkpointing
df.writeStream\
    .option("checkpointLocation", "/mnt/streaming/bronze/weather")\
    .outputMode("append")\
    .format("delta")\
    .toTable("streaming.bronze.weather")
```

e7b97dbb-206e-4ad3-b61c-a4f3969bf7f7 *Last updated: 2 hours ago*

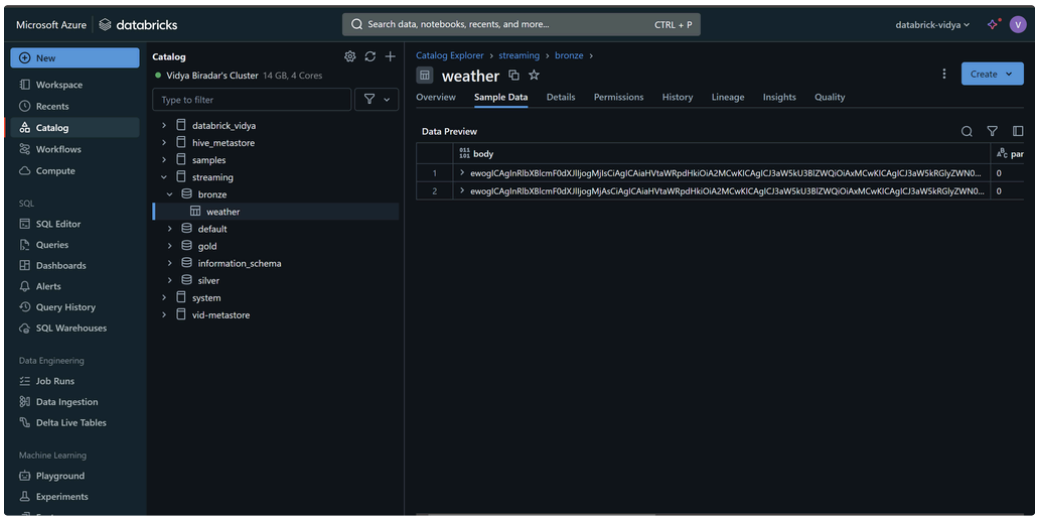
- Generate one more data



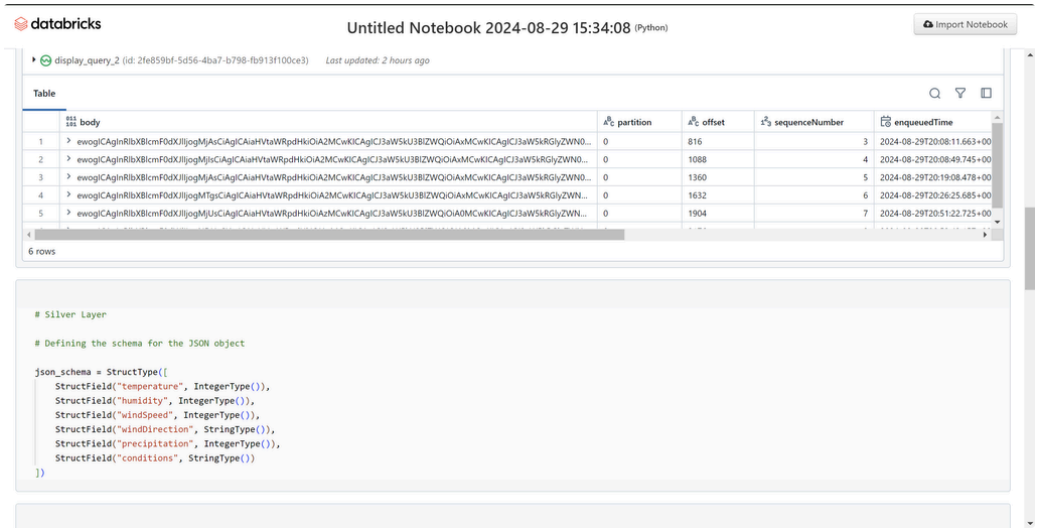
- Reading and writing stream data into bronze layer

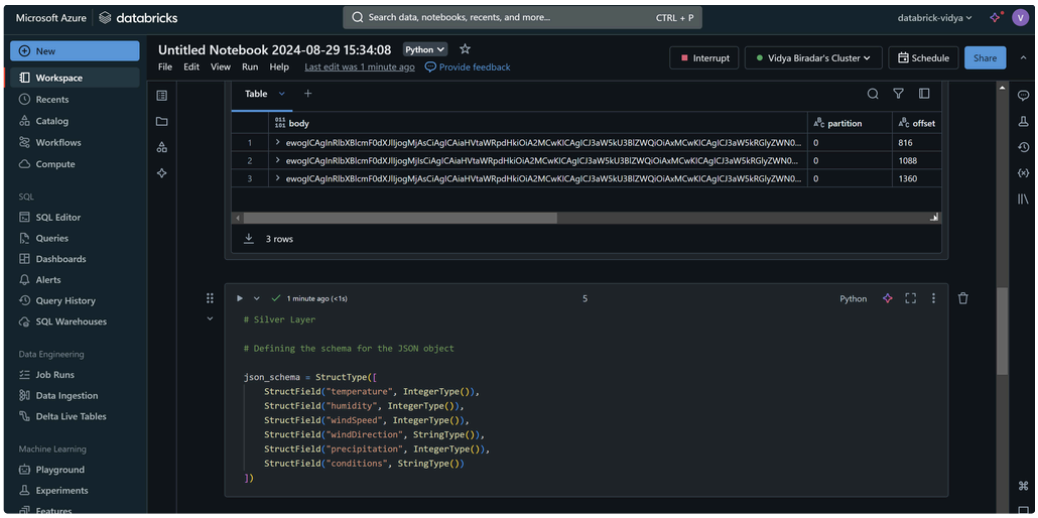


- The data added into bronze layer

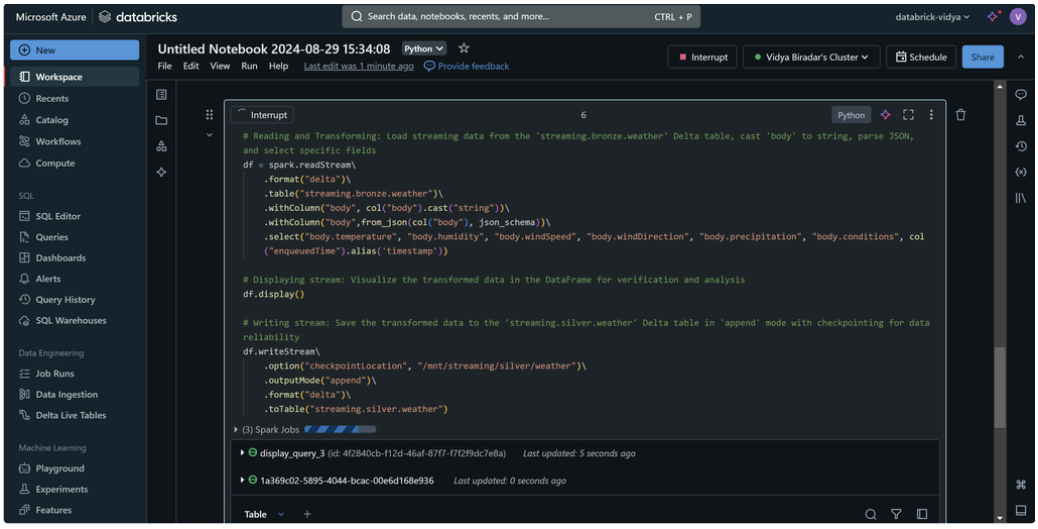


- In the **Silver layer** defining the schema





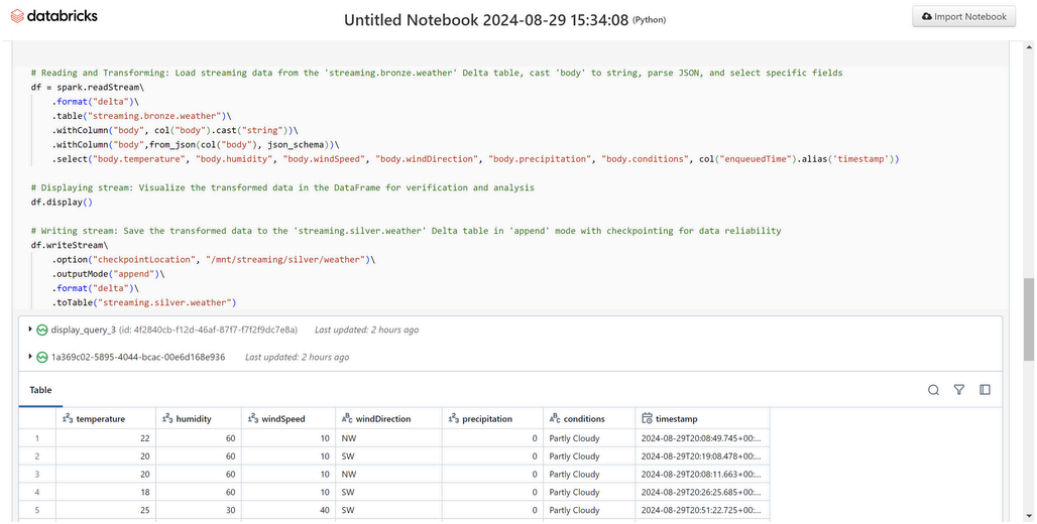
- Reading, Transforming and writing the data into silver layer



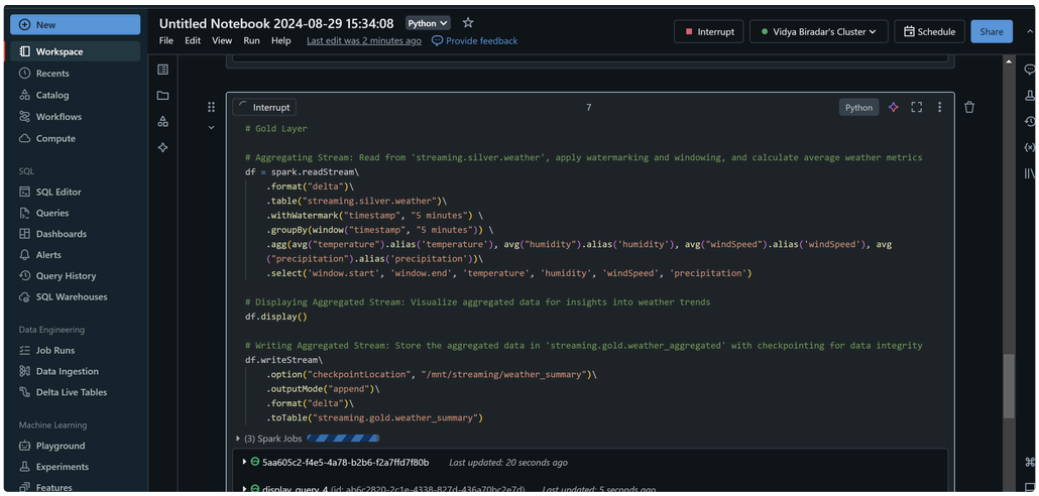
Table

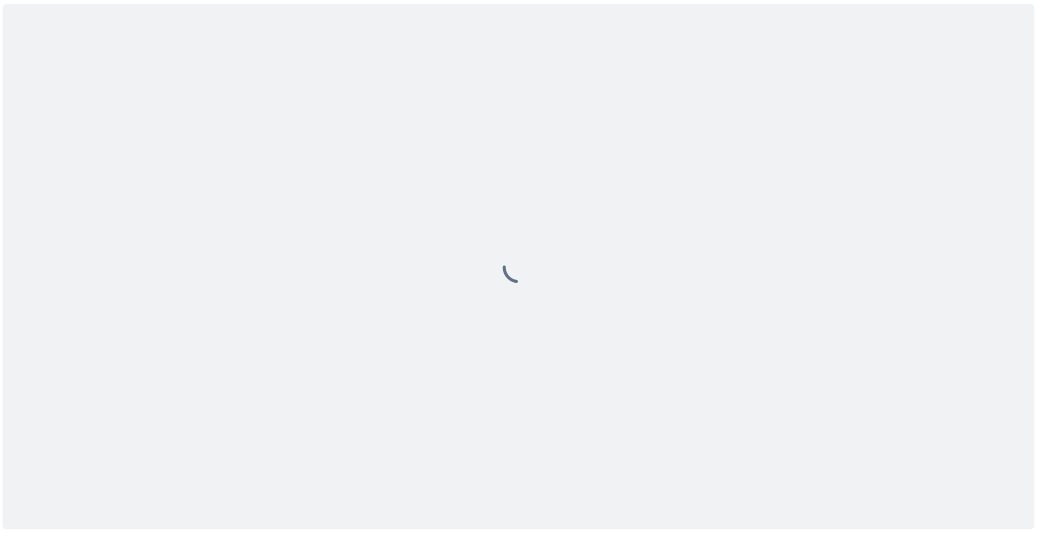
	temperature	humidity	windSpeed	windDirection	precipitation	conditions	timestamp
1	22	60	10	NW	0	Partly Cloudy	2024-08-29T20:08:49.7
2	20	60	10	SW	0	Partly Cloudy	2024-08-29T20:19:08.4
3	20	60	10	NW	0	Partly Cloudy	2024-08-29T20:08:11.6

3 rows

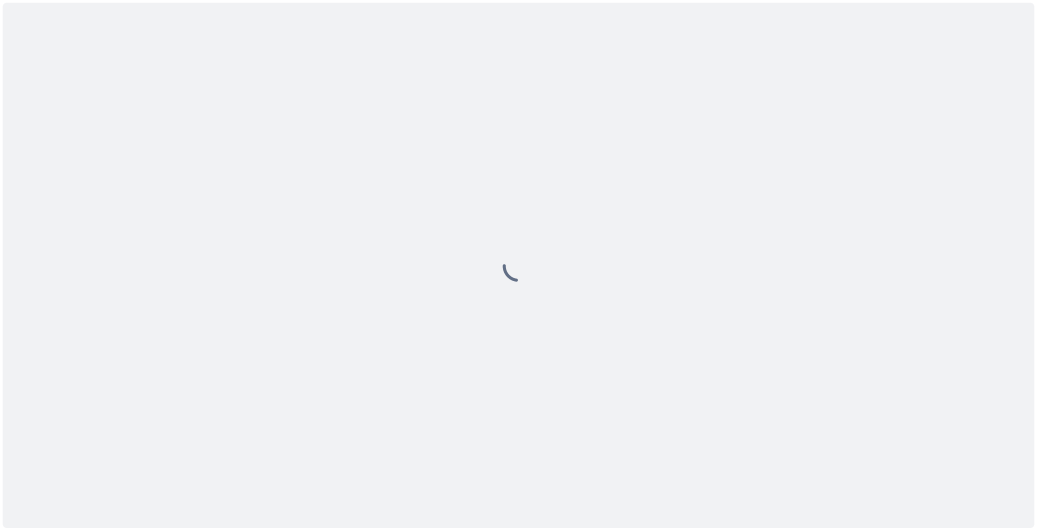


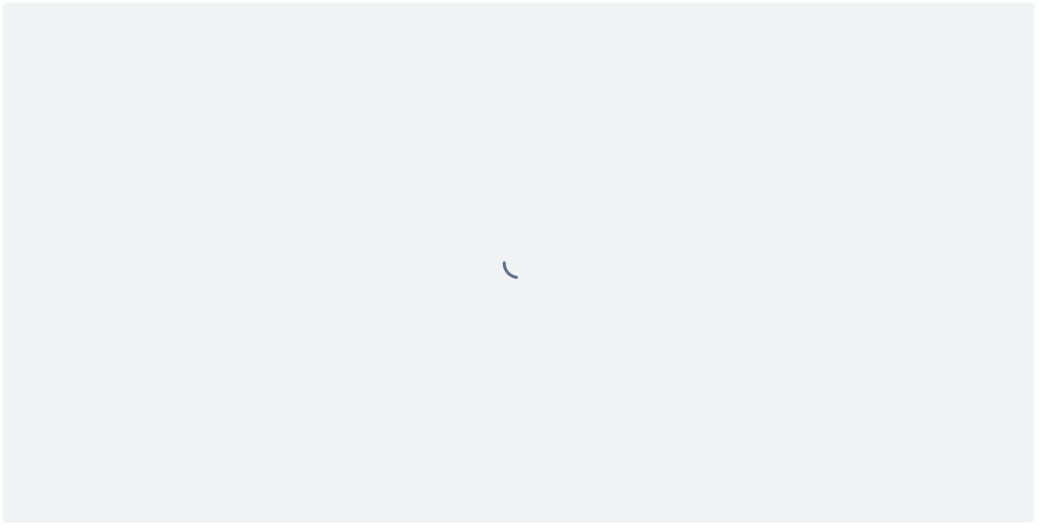
- In the **Gold layer** aggregating stream data and writing to the gold layer

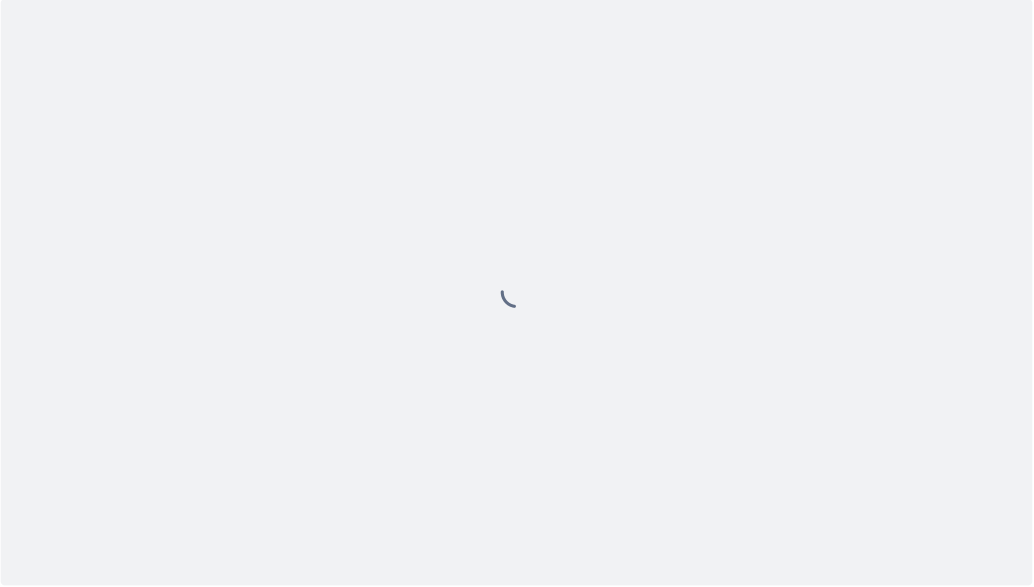




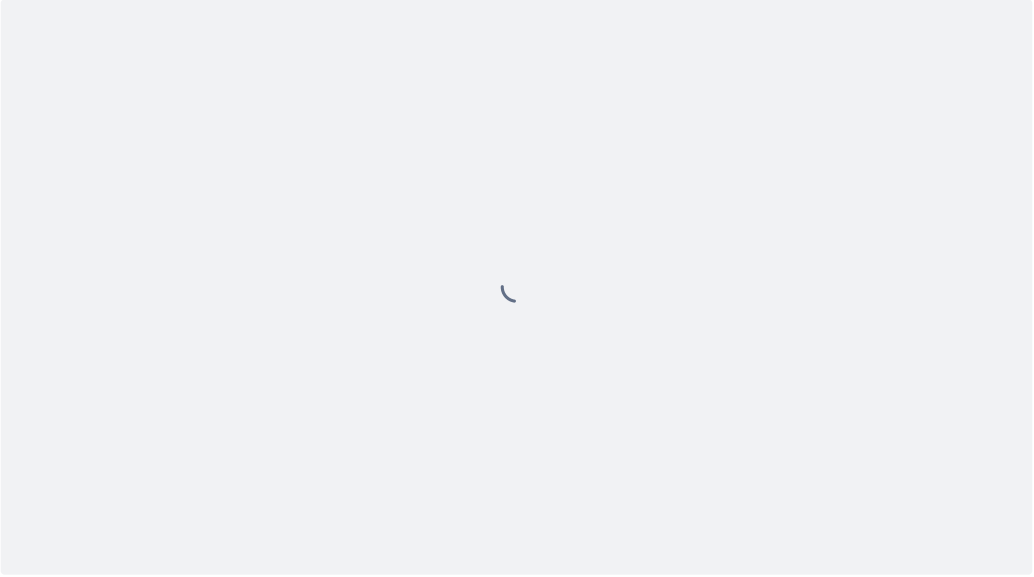
- Using the partner connect option select power BI option for visualization







- 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