

# SPARK STREAMING

terraform apply:

```
MINGW64:/c/Users/YuriiiHordiichuk/Desktop/m13_sparkstreaming_python_az... ━ ━ X
azurerm_databricks_workspace.bdcc: Still creating... [11m10s elapsed]
azurerm_databricks_workspace.bdcc: Still creating... [11m20s elapsed]
azurerm_databricks_workspace.bdcc: Still creating... [11m30s elapsed]
azurerm_databricks_workspace.bdcc: Still creating... [11m40s elapsed]
azurerm_databricks_workspace.bdcc: Still creating... [11m50s elapsed]
azurerm_databricks_workspace.bdcc: Still creating... [12m00s elapsed]
azurerm_databricks_workspace.bdcc: Still creating... [12m10s elapsed]
azurerm_databricks_workspace.bdcc: Still creating... [12m20s elapsed]
azurerm_databricks_workspace.bdcc: Still creating... [12m30s elapsed]
azurerm_databricks_workspace.bdcc: Still creating... [12m40s elapsed]
azurerm_databricks_workspace.bdcc: Creation complete after 12m45s [id=/subscriptions/316ab800-22b5-40ab-be41-8595de4fb2d4/resourceGroups/rg-dev-westeurope-ertk/providers/Microsoft.Databricks/workspaces/dbw-dev-westeurope-ertk]
Releasing state lock. This may take a few moments...
Apply complete! Resources: 5 added, 0 changed, 0 destroyed.

Outputs:
```

resource\_group\_name = "rg-dev-westeurope-ertk"

AzureAD+YuriiiHordiichuk@EPPLWROW0218 MINGW64 ~/Desktop/m13\_sparkstreaming\_python\_azure-master/terraform

\$ 1.

I loaded only 2016 year:

Home > stdevwesteuropeertk | Containers >

**data** Container

Search Add Directory Upload Refresh Delete Copy Paste Rename Acquire lease Break lease Edit columns

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Authentication method: Access key (Switch to Microsoft Entra user account)

Search blobs by prefix (case-sensitive)

Showing all 2 items

Name	Last modified
year=2016	12/25/2025, 8:17:22 PM
.DS_Store	12/25/2025, 8:17:22 PM

2.I ran all the notebooks:

Detach

<input type="checkbox"/>	Name	Status	Last Command Run	Location
<input type="checkbox"/>	04_silver_to_gold	<span>Running</span>	Thu, Dec 25, 2025, 21:05:19 GMT+1 by ffwsw@gmail.com	/Users/ffwsw@gmail.com/04_silver_to_gold
<input type="checkbox"/>	01_create_metadata	<span>Idle</span>	Thu, Dec 25, 2025, 20:53:07 GMT+1 by ffwsw@gmail.com	/Users/ffwsw@gmail.com/01_create_metadata
<input type="checkbox"/>	02_load_bronze_data	<span>Running</span>	Thu, Dec 25, 2025, 21:03:41 GMT+1 by ffwsw@gmail.com	/Users/ffwsw@gmail.com/02_load_bronze_data
<input type="checkbox"/>	03_bronze_to_silver	<span>Running</span>	Thu, Dec 25, 2025, 21:05:08 GMT+1 by ffwsw@gmail.com	/Users/ffwsw@gmail.com/03_bronze_to_silver
<input type="checkbox"/>	test	<span>Idle</span>	Thu, Dec 25, 2025, 21:03:51 GMT+1 by ffwsw@gmail.com	/Users/ffwsw@gmail.com/test

### 3. Results for 2016 year:

```

▶ ✓ Just now (4s) 1

%sql
select 'count bronze.hotel_weather_raw' tbl_name, count(*) cnt from bronze.hotel_weather_raw
union
select 'count silver.hotel_weather_processed', count(*) from silver.hotel_weather_processed
union
select 'count gold.hotel_weather_metrics', count(*) from gold.hotel_weather_metrics

▶ (14) Spark Jobs

> _sqldf: pyspark.sql.DataFrame = [tbl_name: string, cnt: long]

Table ▾ +
```

A	B	C	tbl_name	D	E	cnt
1	count	bronze.hotel_weather_raw		4980		
2	count	silver.hotel_weather_process...		4980		
3	count	gold.hotel_weather_metrics		1512		

Preview gold table:

► ✓ Just now (4s)

2

```
%sql  
select * from gold.hotel_weather_metrics
```

► (3) Spark Jobs

> \_sql: pyspark.sql.dataframe.DataFrame = [country: string, city: string ... 6 more fields]

Table +

Q Y E

	A <sup>B</sup> country	A <sup>B</sup> city	wthr_date	1 <sup>2</sup> num_distinct_hotels	1.2 avg_temp_c	1.2 max_temp_c	1.2 min_temp_c	1.
1	FR	Paris	2016-10-31	237	10.699999999999994	10.7	10.7	10.7
2	US	Maumee	2016-10-20	1	12.8	12.8	12.8	12.8
3	US	Milford	2016-10-10	1	13.8	13.8	13.8	13.8
4	US	Ashland	2016-10-25	1	6.8	6.8	6.8	6.8
5	US	Burdett	2016-10-26	1	2.3	2.3	2.3	2.3
6	US	Aberdeen	2016-10-21	1	8.8	8.8	8.8	8.8
7	US	Junction	2016-10-27	1	20.9	20.9	20.9	20.9
8	US	Lumberton	2016-10-20	1	22.4	22.4	22.4	22.4
9	US	Millbrook	2016-10-25	1	19.2	19.2	19.2	19.2
10	US	Millington	2016-10-10	1	18.3	18.3	18.3	18.3
11	US	Sweetwater	2016-10-20	1	22.7	22.7	22.7	22.7
12	US	Washington	2016-10-22	1	7.4	7.4	7.4	7.4
13	US	Palm Harbor	2016-10-04	2	25.9	25.9	25.9	25.9
14	US	Grand Prairie	2016-10-13	1	18.1	18.1	18.1	18.1
15								

## 4. Loading 2017 year:

Interrupt

```
load_encrypt_stream_write(  
    source_path=hotel_weather_source_path,  
    schema=hotel_weather_schema,  
    pii_columns=hotel_weather_pii_columns,  
    encryptor=encryptor,  
    fmt="parquet",  
    checkpoint_path="/checkpoints/bronze/hotel_weather_raw",  
    target_table="bronze.hotel_weather_raw"  
)  
  
print("Streaming Bronze ingestion started")
```

► (1) Spark Jobs

Job 873 View (2 stages)

29426e59-7a8b-4407-82a9-f389dcff6fe82 Last updated: 5 seconds ago

Dashboard Raw Data

Input vs. Processing Rate

0 rec/s 0 rec/s Input rate Processing rate

Batch Duration in milliseconds

722.7 ms 1 ms Average Latest

# write streaming silver  
(  
 hotel\_clean.writeStream  
 .format("delta")  
 .outputMode("append")  
 .option("checkpointLocation", "/mnt/checkpoints/silver\_hotel\_weather\_n")  
 .table("silver.hotel\_weather\_processed")  
)

► (1) Spark Jobs

57f250b4-7834-4b17-8a9e-88feb808d2e5 Last updated: 1 minute ago

Dashboard Raw Data

Input vs. Processing Rate

0 rec/s 0 rec/s Input rate Processing rate

Batch Duration in milliseconds

7666.4 ms 31.9k ms Average Latest

Aggregation State

13.3k Distinct keys

> hotel\_bronze: pyspark.sql.dataframe.DataFrame = [address: string, avg\_tmpr\_c: double ... 13 more fields]  
> hotel\_clean: pyspark.sql.dataframe.DataFrame = [address: string, avg\_tmpr\_c: string ... 13 more fields]

```
# Write Gold Delta table
(
    hotel_metrics.writeStream
        .format("delta")
        .outputMode("complete")
        .option("checkpointLocation", "/mnt/checkpoints/gold_hotel_weather_metrics_n")
        .table("gold.hotel_weather_metrics")
)
```



## Result after loading 2017 year:

```
%sql
select 'count bronze.hotel_weather_raw' tbl_name, count(*) cnt from bronze.hotel_weather_raw
union
select 'count silver.hotel_weather_processed', count(*) from silver.hotel_weather_processed
union
select 'count gold.hotel_weather_metrics', count(*) from gold.hotel_weather_metrics
```

▶ (15) Spark Jobs

> [SQL] \_sqldf: pyspark.sql.dataframe.DataFrame = [tbl\_name: string, cnt: long]

	tbl_name	cnt
1	count bronze.hotel_weather_raw	13330
2	count silver.hotel_weather_processed	13330
3	count gold.hotel_weather_metrics	4324

## 5. Data Visualization:

I can't use Databricks Dashboards:

The screenshot shows the Microsoft Azure Databricks interface. The search bar at the top contains the text 'das'. Below it, there are three tabs: 'Notebooks', 'Jobs', and 'My assets'. A red circle highlights the 'Dashboards' tab under the 'Products and Pages' section. A large red oval encloses a message: 'For access to Databricks SQL, contact your Databricks representative.' followed by 'To retry, refresh the page.' with an exclamation mark icon.

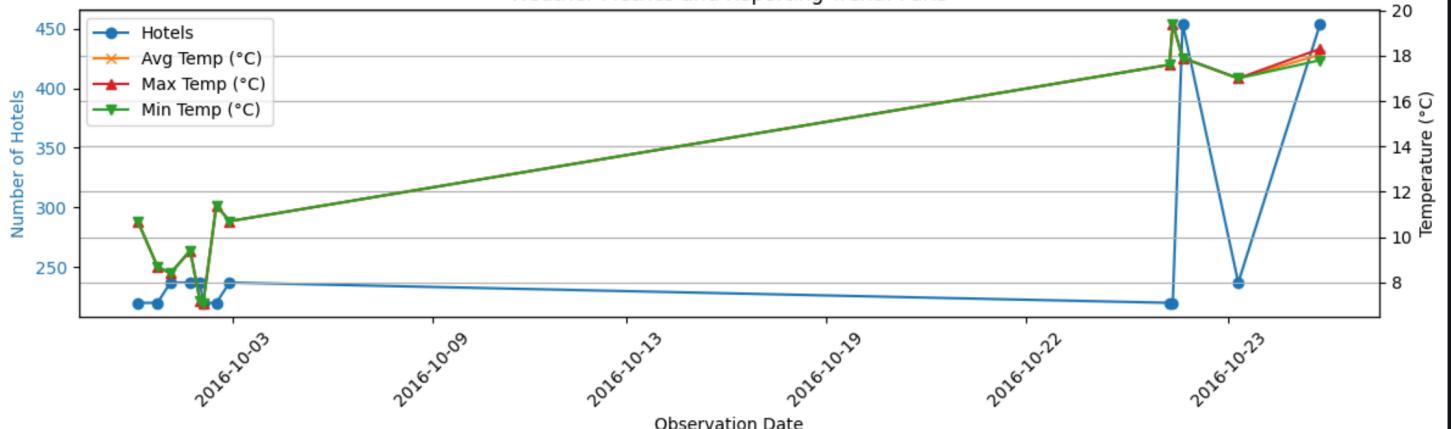
Databricks Dashboards is not available for me:

The screenshot shows the Microsoft Azure Databricks interface with the 'Workspace' tab selected in the sidebar. A modal window titled 'Add or upload data' is open, listing various options: Notebook, Job, Experiment, Model, App, More, Git folder, Cluster, and Serving endpoint. The 'More' option is expanded, showing the other items. The right side of the screen shows a dark sidebar with various icons and names like 'ffwsdwo', 'Search', 'Na', and 'Git'.

So I decided to visualize it using matplotlib (at least something):

```
/root/.ipykernel/23944/command-4968921652930149-2866295632:30: UserWarning: set_ticklabels() should only be used with a fixed number of ticks, i.e. after set_ticks() or u  
ax1.set_xticklabels(df["wthr_date"], rotation=45)
```

### Weather Metrics and Reporting Trend: Paris



```
/root/.ipykernel/23944/command-4968921652930149-2866295632:30: UserWarning: set_ticklabels() should only be used with a fixed number of ticks, i.e. after set_ticks() or u  
ax1.set_xticklabels(df["wthr_date"], rotation=45)
```

### Weather Metrics and Reporting Trend: London

