

# SPARK STREAMING

terraform apply:

```
MINGW64:/c:/Users/YuriiHordiichuk/Desktop/m13_sparkstreaming_python_az...
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-west europe-ertk/providers/Microsoft.Databricks/workspaces/dbw-dev-west europe-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-west europe-ertk"

AzureAD+YuriiHordiichuk@EPPLWROW0218 MINGW64 ~/Desktop/m13_sparkstreaming_python_azure-master/terraform
$
```

1.

I loaded only 2016 year:

Home > stdevwest europeertk | 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

data > hotel-weather

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

Search blobs by prefix (case-sensitive)

Showing all 2 items

<input type="checkbox"/>	Name	Last modified
<input type="checkbox"/>	[-]	
<input type="checkbox"/>	year=2016	12/25/2025, 8:17:22 PM
<input type="checkbox"/>	.DS_Store	12/25/2025, 8:17:22 PM

2.I ran all the notebooks:

ffwsdw@gmail.com's Cluster 2025-12-25 20:10:30

Configuration

Notebooks (5)

Libraries

Event log

Spark UI

Driver logs

Metrics

Apps

Spark compute UI - Master

Detach

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

Table ▾

+

	A <sub>C</sub> <sup>B</sup> tbl_name	1 <sub>3</sub> <sup>2</sup> 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

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

Table

+

country

city

wthr\_date

num\_distinct\_hotels

avg\_temp\_c

max\_temp\_c

min\_temp\_c

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

#### 4. Loading 2017 year:

Interrupt

5

Python

```
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-f389dcf6fe82 Last updated: 5 seconds ago

Dashboard Raw Data

Input vs. Processing Rate

records per second

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")
)

> 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]
```

(1) Spark Jobs

577250b4-7834-4b17-8a9e-88feb808d2e5 Last updated: 1 minute ago

Dashboard Raw Data

Input vs. Processing Rate

records per second

0 rec/s 0 rec/s

Input rate Processing rate

Batch Duration

in milliseconds

7666.4 ms 31.9k ms

Average Latest

Aggregation State

Distinct keys

13.3k



Result after loading 2017 year:



5.Data Visualization:



## Datasets:

### Visualize TOP 5 cities by number of di...

Data | Untitled page

Datasets

Paris

London

Milan

Amsterdam

Barcelona

Create another dataset

Create from SQL

Add data source

Upload file

Run 2 minutes ago Default (dbw\_dev\_westeurope.j2a...)

SELECT wthr\_date, num\_distinct\_hotels, avg\_temp\_c, max\_temp\_c, min\_temp\_c  
FROM gold.hotel\_weather\_metrics  
WHERE city = 'Paris'  
ORDER BY wthr\_date

Add parameter

Result Table Schema

	wthr_date	num_distinct_hotels	avg_temp_c	max_temp_c	min_temp_c
1	2016-10-03	220	10.699999999999996	10.7	10.7
2	2016-10-09	220	8.699999999999996	8.7	8.7
3	2016-10-13	237	8.399999999999999	8.4	8.4
4	2016-10-19	237	9.399999999999997	9.4	9.4
5	2016-10-22	237	7.199999999999996	7.2	7.2
6	2016-10-23	220	7.099999999999996	7.1	7.1
7	2016-10-27	220	11.399999999999997	11.4	11.4
8	2016-10-31	237	10.699999999999994	10.7	10.7
9	2017-08-14	220	17.6000000000000012	17.6	17.6
10	2017-08-15	220	19.3999999999999984	19.4	19.4
11	2017-08-18	454	17.9000000000000006	17.9	17.9
12	2017-09-04	237	17	17	17
13	2017-09-29	454	18.043929359823377	18.3	17.8