INDIAN DATA CLUB

CODE BASICS

databricks

# 14 DAYS

## AI CHALLENGE

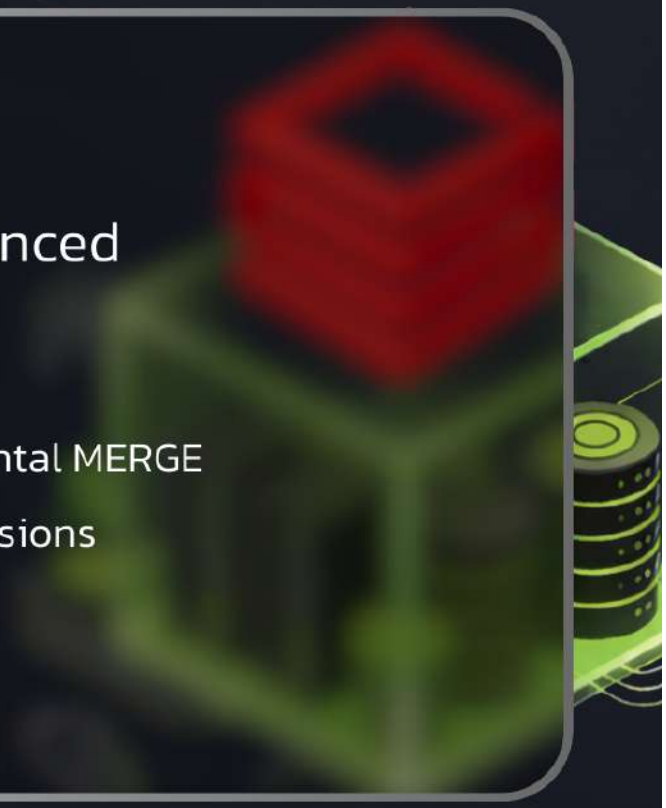
## DAY 05

**Topic:**

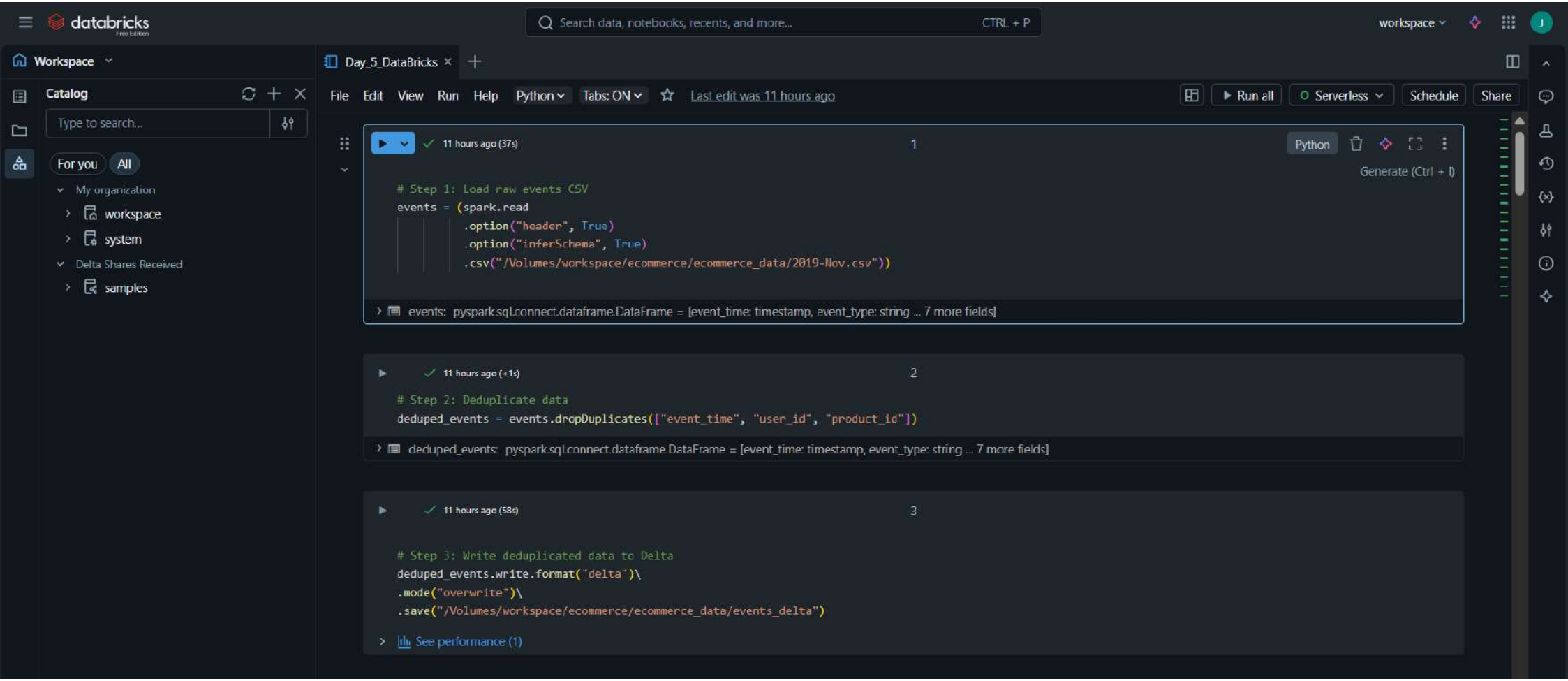
Delta Lake Advanced

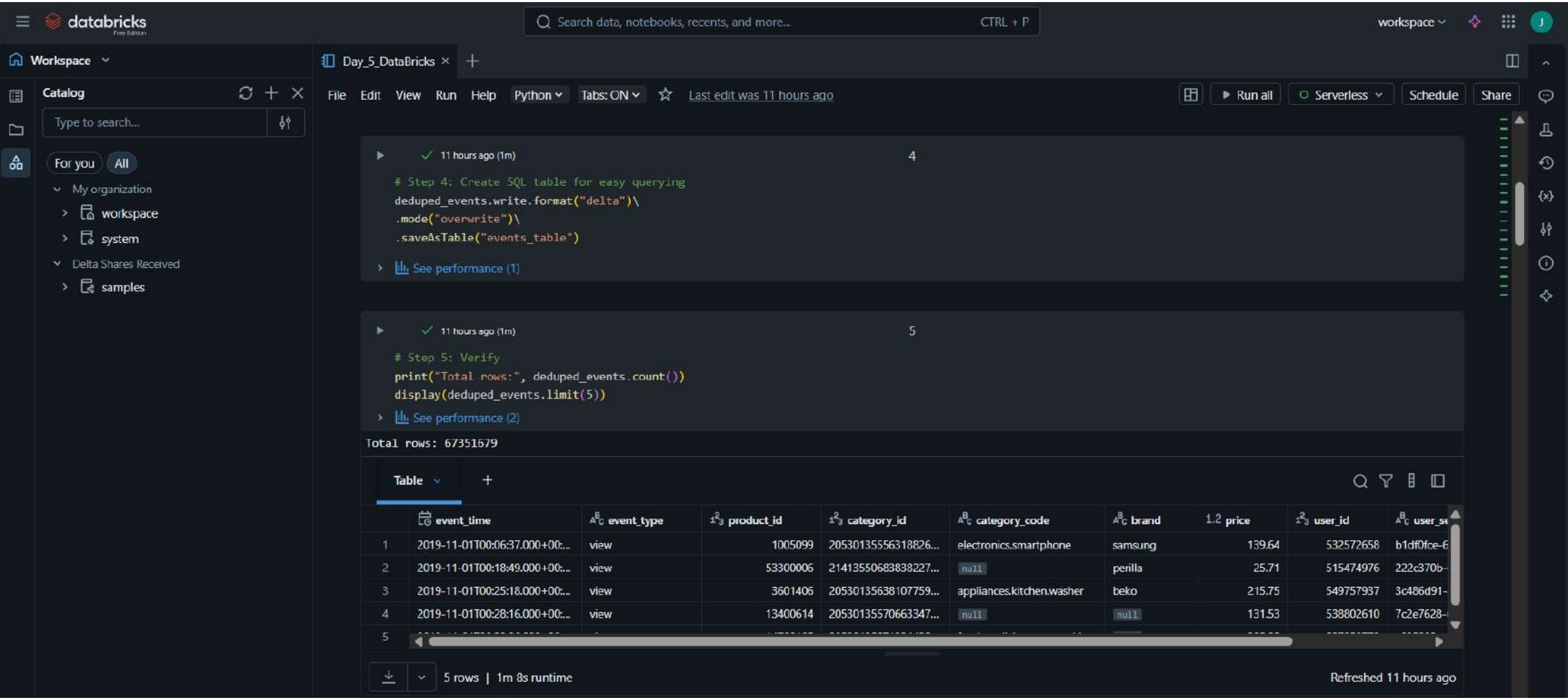
**Challenge:**

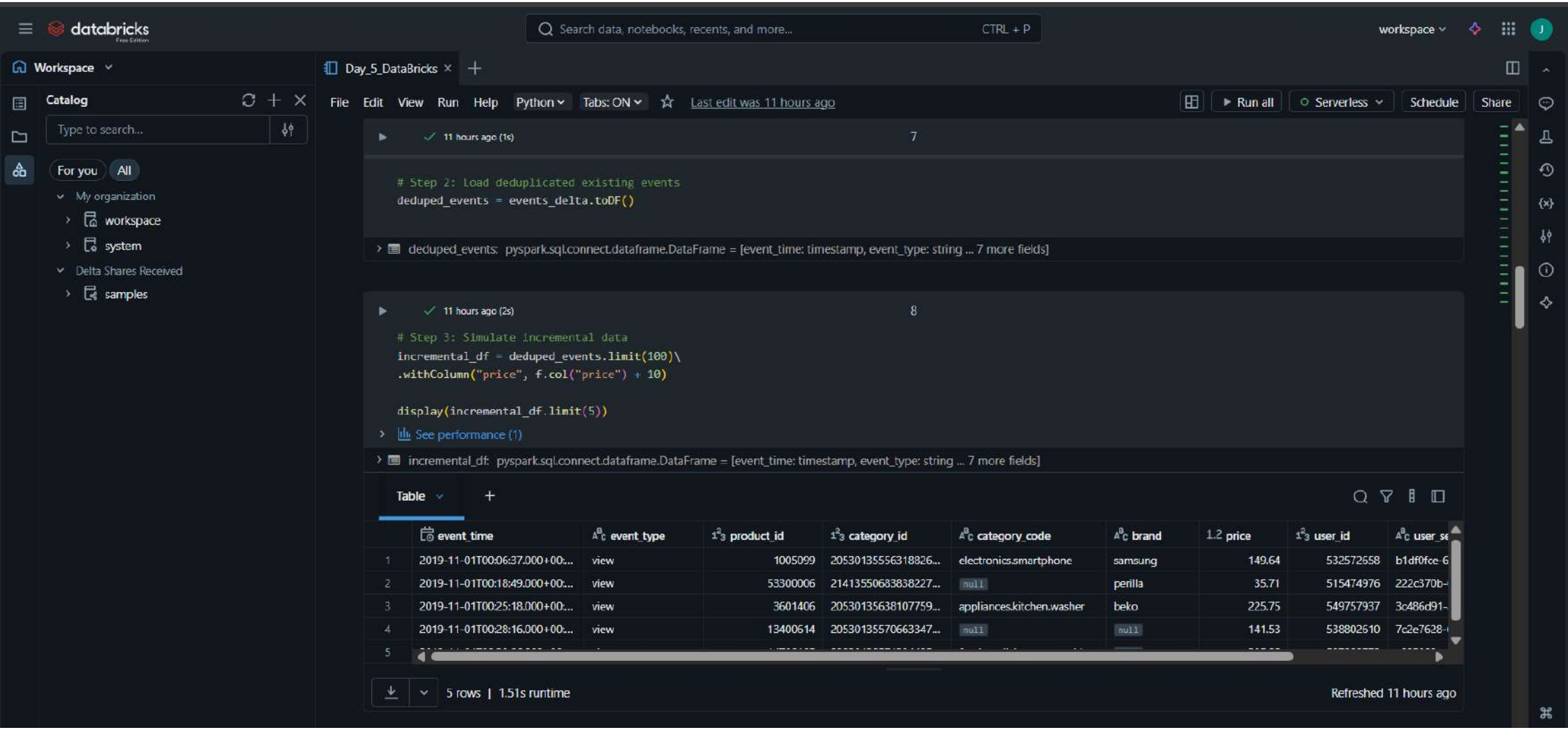
1. Implement incremental MERGE
2. Query historical versions
3. Optimize tables
4. Clean old files

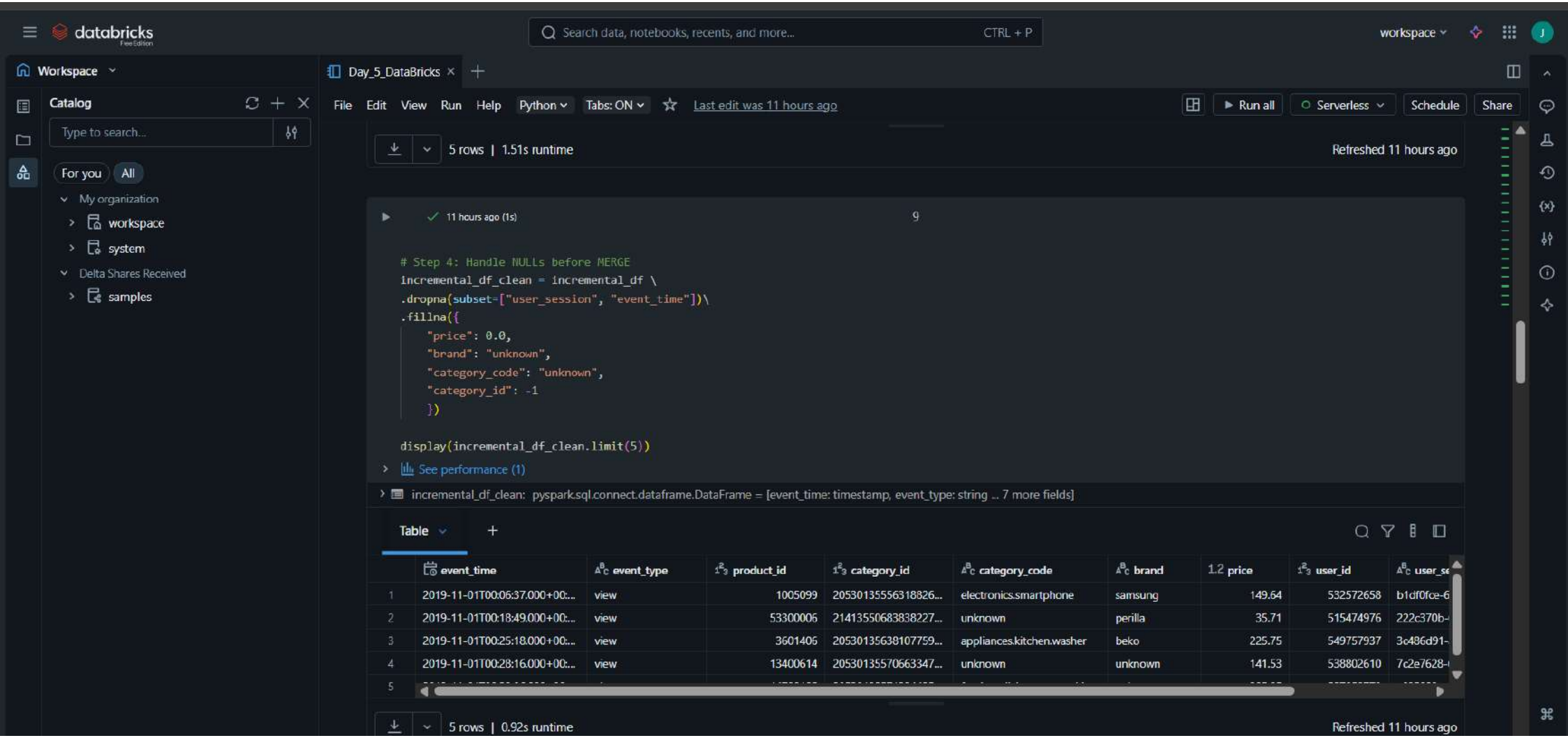


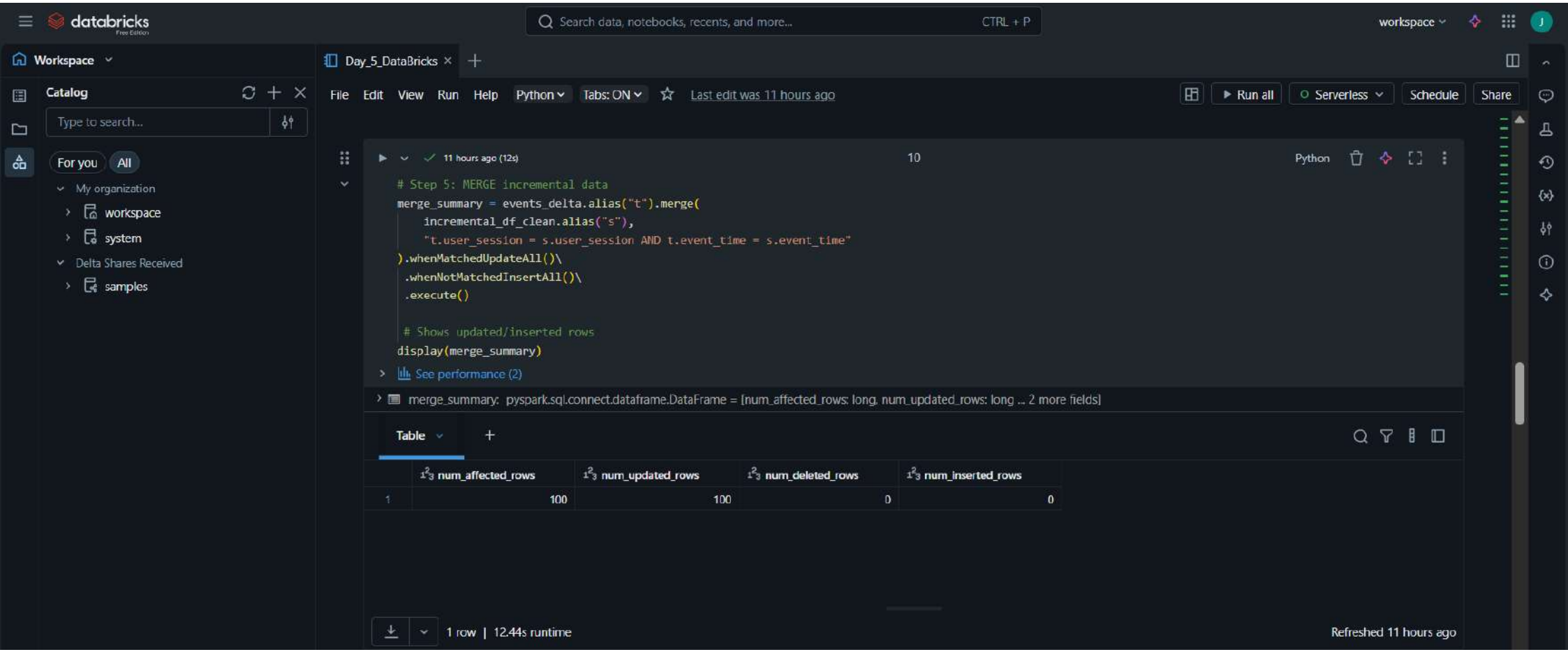
#DatabricksWithIDC

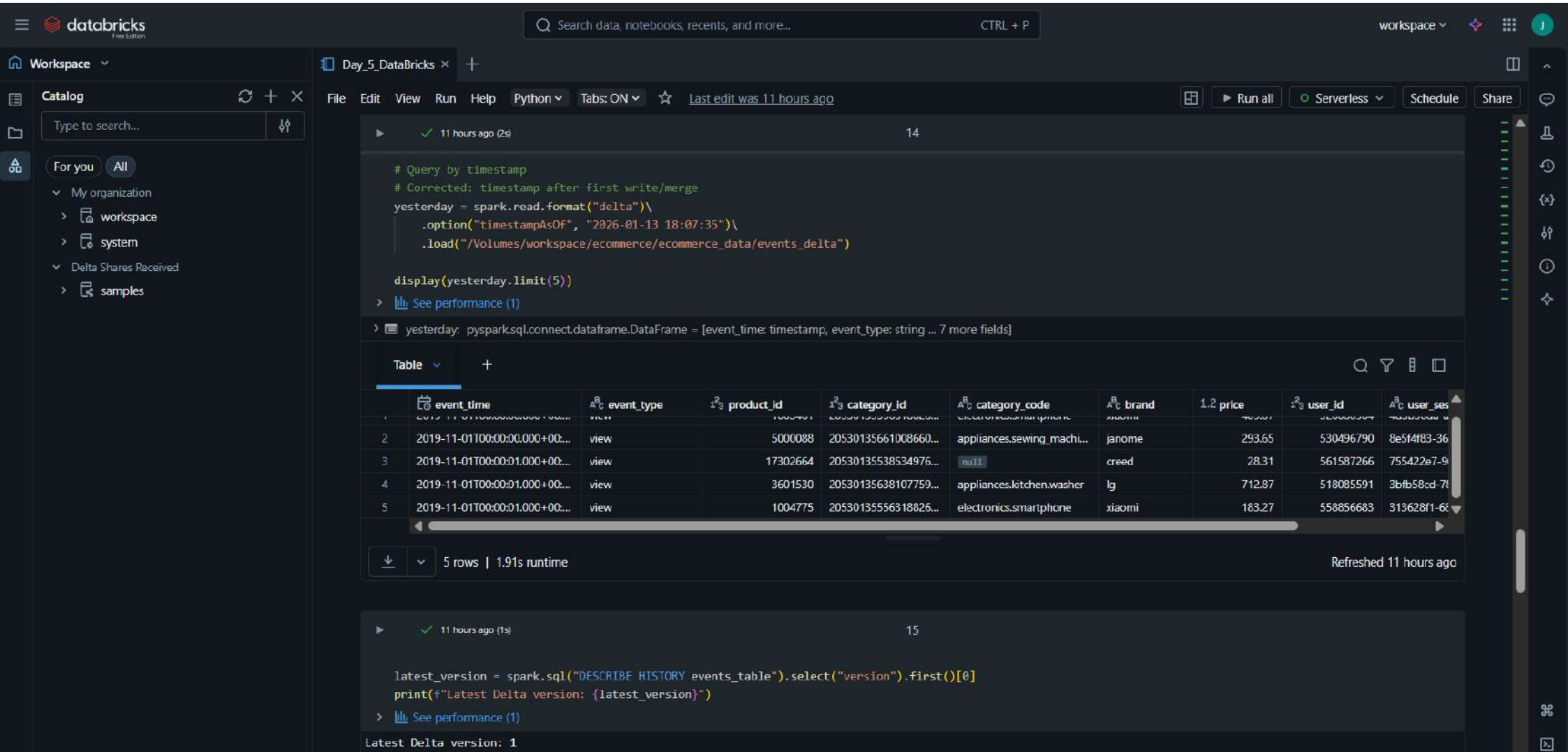

databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_5_DataBricks
Catalog
Type to search...
For you
All
My organization
workspace
system
Delta Shares Received
samples
File
Edit
View
Run
Help
Python
Tabs: ON
Last edit was 11 hours ago
Run all
Serverless
Schedule
Share
11 hours ago (37s)
1
Python
Generate (Ctrl + I)
# Step 1: Load raw events CSV
events = (spark.read
.option("header", True)
.option("inferSchema", True)
.csv("/Volumes/workspace/ecommerce/ecommerce_data/2019-Nov.csv"))
events: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 7 more fields]
11 hours ago (<1s)
2
# Step 2: Deduplicate data
deduped_events = events.dropDuplicates(["event_time", "user_id", "product_id"])
deduped_events: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 7 more fields]
11 hours ago (58s)
3
# Step 3: Write deduplicated data to Delta
deduped_events.write.format("delta")\
.mode("overwrite")\
.save("/Volumes/workspace/ecommerce/ecommerce_data/events_delta")
See performance (1)


databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_5_DataBricks
Catalog
Type to search...
For you
All
My organization
workspace
system
Delta Shares Received
samples
File
Edit
View
Run
Help
Python
Tabs: ON
Last edit was 11 hours ago
Run all
Serverless
Schedule
Share
11 hours ago (1m)
4
# Step 4: Create SQL table for easy querying
deduped_events.write.format("delta")\
.mode("overwrite")\
.saveAsTable("events_table")
See performance (1)
11 hours ago (1m)
5
# Step 5: Verify
print("Total rows:", deduped_events.count())
display(deduped_events.limit(5))
See performance (2)
Total rows: 67351679
Table
event_time
event_type
product_id
category_id
category_code
brand
price
user_id
user_st
1 2019-11-01T00:06:37.000+00:... view 1005099 20530135556318826... electronics.smartphone samsung 139.64 532572658 b1df0fce-6
2 2019-11-01T00:18:49.000+00:... view 53300006 21413550683838227... null perilla 25.71 515474976 222c370b-
3 2019-11-01T00:25:18.000+00:... view 3601406 20530135638107759... appliances.kitchen.washer beko 215.75 549757937 3c486d91-
4 2019-11-01T00:28:16.000+00:... view 13400614 20530135570663347... null null 131.53 538802610 7c2e7628-
5
5 rows | 1m 8s runtime
Refreshed 11 hours ago


databricks
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_5_DataBricks
Catalog
Type to search...
For you
All
My organization
workspace
system
Delta Shares Received
samples
File Edit View Run Help Python Tabs: ON
Last edit was 11 hours ago
Run all
Serverless
Schedule
Share
11 hours ago (1s)
7
# Step 2: Load deduplicated existing events
deduped_events = events_delta.toDF()
deduped_events: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 7 more fields]
11 hours ago (2s)
8
# Step 3: Simulate incremental data
incremental_df = deduped_events.limit(100)\
.withColumn("price", f.col("price") + 10)
display(incremental_df.limit(5))
See performance (1)
incremental_df: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 7 more fields]
Table
event_time
event_type
product_id
category_id
category_code
brand
price
user_id
user_se
1 2019-11-01T00:06:37.000+00... view 1005099 20530135556318826... electronics.smartphone samsung 149.64 532572558 b1df0fce-6
2 2019-11-01T00:18:49.000+00... view 53300006 21413550683838227... null perilla 35.71 515474976 222c370b-
3 2019-11-01T00:25:18.000+00... view 3601406 20530135638107759... appliances.kitchen.washer beko 225.75 549757937 3c486d91-
4 2019-11-01T00:28:16.000+00... view 13400514 20530135570663347... null null 141.53 538802510 7c2e7628-
5
5 rows | 1.51s runtime
Refreshed 11 hours ago


databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_5_DataBricks
Catalog
Type to search...
For you
All
My organization
workspace
system
Delta Shares Received
samples
File Edit View Run Help Python Tabs: ON
Last edit was 11 hours ago
Run all
Serverless
Schedule
Share
5 rows | 1.51s runtime
Refreshed 11 hours ago
11 hours ago (1s)
# Step 4: Handle NULLs before MERGE
incremental_df_clean = incremental_df \
.dropna(subset=["user_session", "event_time"])\
.fillna({
"price": 0.0,
"brand": "unknown",
"category_code": "unknown",
"category_id": -1
})
display(incremental_df_clean.limit(5))
See performance (1)
incremental_df_clean: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 7 more fields]
Table
event_time
event_type
product_id
category_id
category_code
brand
price
user_id
user_se
1 2019-11-01T00:06:37.000+00:... view 1005099 2053013555631882... electronics.smartphone samsung 149.64 532572658 b1df0fce-6
2 2019-11-01T00:18:49.000+00:... view 53300006 2141355068383822... unknown perilla 35.71 515474976 222c370b-
3 2019-11-01T00:25:18.000+00:... view 3601406 2053013563810775... appliances.kitchen.washer beko 225.75 549757937 3c486d91-
4 2019-11-01T00:28:16.000+00:... view 13400614 2053013557066334... unknown unknown 141.53 538802610 7c2e7628-
5
5 rows | 0.92s runtime
Refreshed 11 hours ago


databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_5_DataBricks
Catalog
Type to search...
For you
All
My organization
workspace
system
Delta Shares Received
samples
File
Edit
View
Run
Help
Python
Tabs: ON
Last edit was 11 hours ago
Run all
Serverless
Schedule
Share
11 hours ago (12s)
10
Python
# Step 5: MERGE incremental data
merge_summary = events_delta.alias("t").merge(
incremental_df_clean.alias("s"),
"t.user_session = s.user_session AND t.event_time = s.event_time"
).whenMatchedUpdateAll()\
.whenNotMatchedInsertAll()\
.execute()
# Shows updated/inserted rows
display(merge_summary)
See performance (2)
merge_summary: pyspark.sql.connect.dataframe.DataFrame = [num_affected_rows: long, num_updated_rows: long ... 2 more fields]
Table
num_affected_rows
num_updated_rows
num_deleted_rows
num_inserted_rows
1
100
100
0
0
1 row | 12.44s runtime
Refreshed 11 hours ago


databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_5_DataBricks
Catalog
Type to search...
For you
All
My organization
workspace
system
Delta Shares Received
samples
File Edit View Run Help Python Tabs: ON
Last edit was 11 hours ago
Run all
Serverless
Schedule
Share
11 hours ago (2s)
14
# Query by timestamp
# Corrected: timestamp after first write/merge
yesterday = spark.read.format("delta")\
.option("timestampAsOf", "2026-01-13 18:07:35")\
.load("/Volumes/workspace/ecommerce/ecommerce_data/events_delta")
display(yesterday.limit(5))
See performance (1)
yesterday: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 7 more fields]
Table
event_time
event_type
product_id
category_id
category_code
brand
price
user_id
user_ses
2 2019-11-01T00:00:00.000+00:... view 5000088 2053013566100866O... appliances.sewing_machi... janome 293.65 530496790 8e5f4f83-36
3 2019-11-01T00:00:01.000+00:... view 17302664 20530135538534976... null creed 28.31 561587266 755422e7-9
4 2019-11-01T00:00:01.000+00:... view 3601530 20530135638107759... appliances.kitchen.washer lg 712.87 518085591 3bfb58cd-7
5 2019-11-01T00:00:01.000+00:... view 1004775 20530135556318826... electronics.smartphone xiaomi 183.27 558856683 313628f1-6
5 rows | 1.91s runtime
Refreshed 11 hours ago
11 hours ago (1s)
15
latest_version = spark.sql("DESCRIBE HISTORY events_table").select("version").first()[0]
print(f"Latest Delta version: {latest_version}")
See performance (1)
Latest Delta version: 1