INDIAN DATA CLUB

CODE BASICS

databricks

14 DAYS

AI CHALLENGE

DAY 06

Topic:

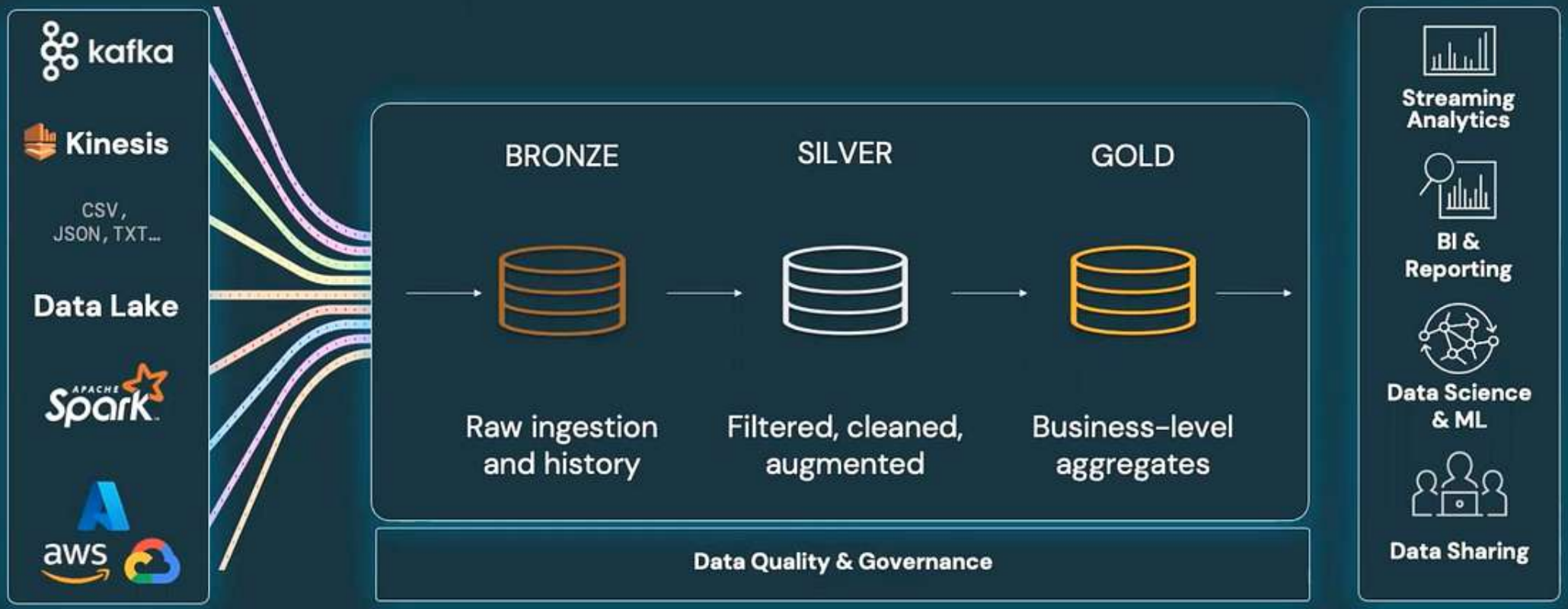
Medallion Architecture

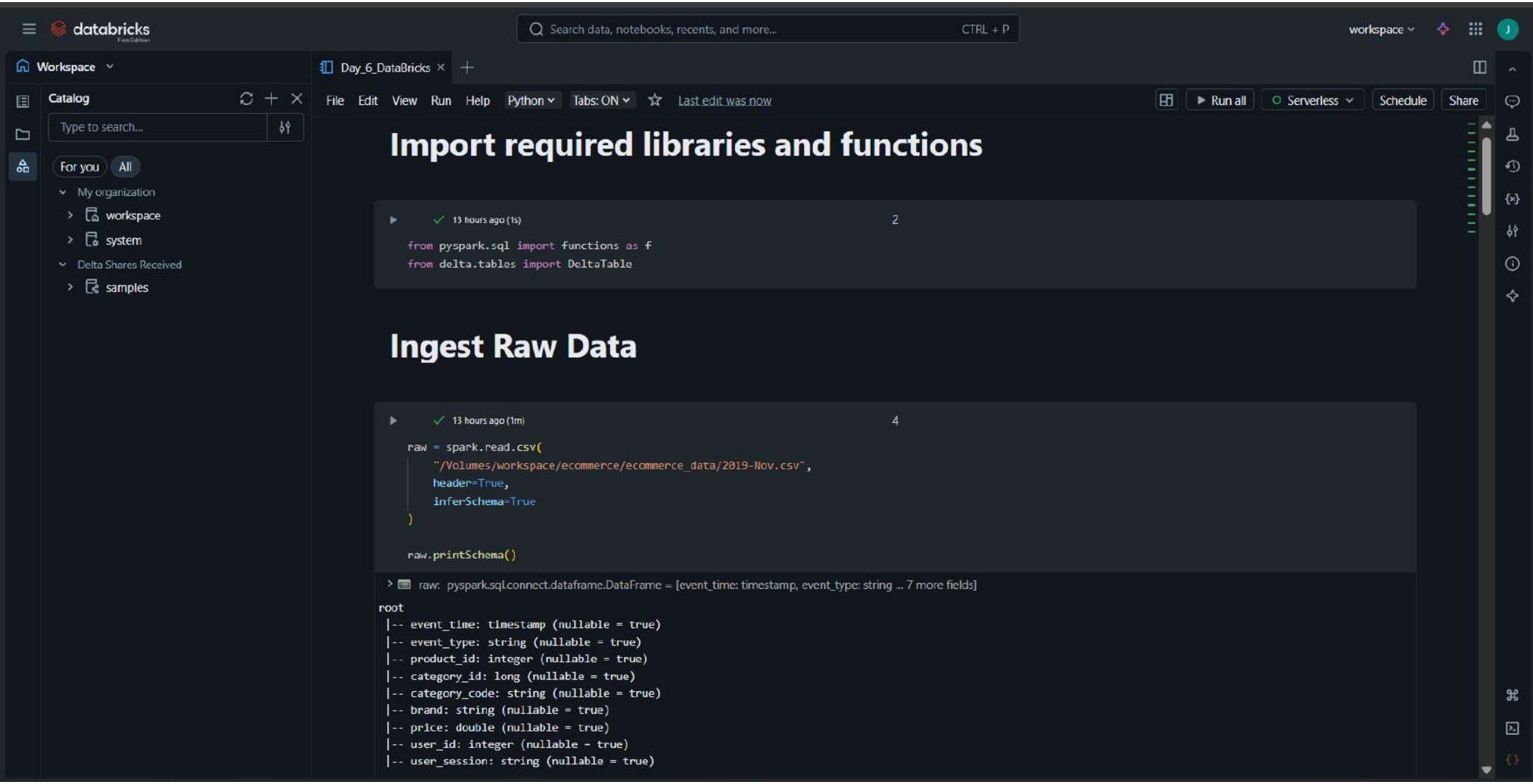
Challenge:

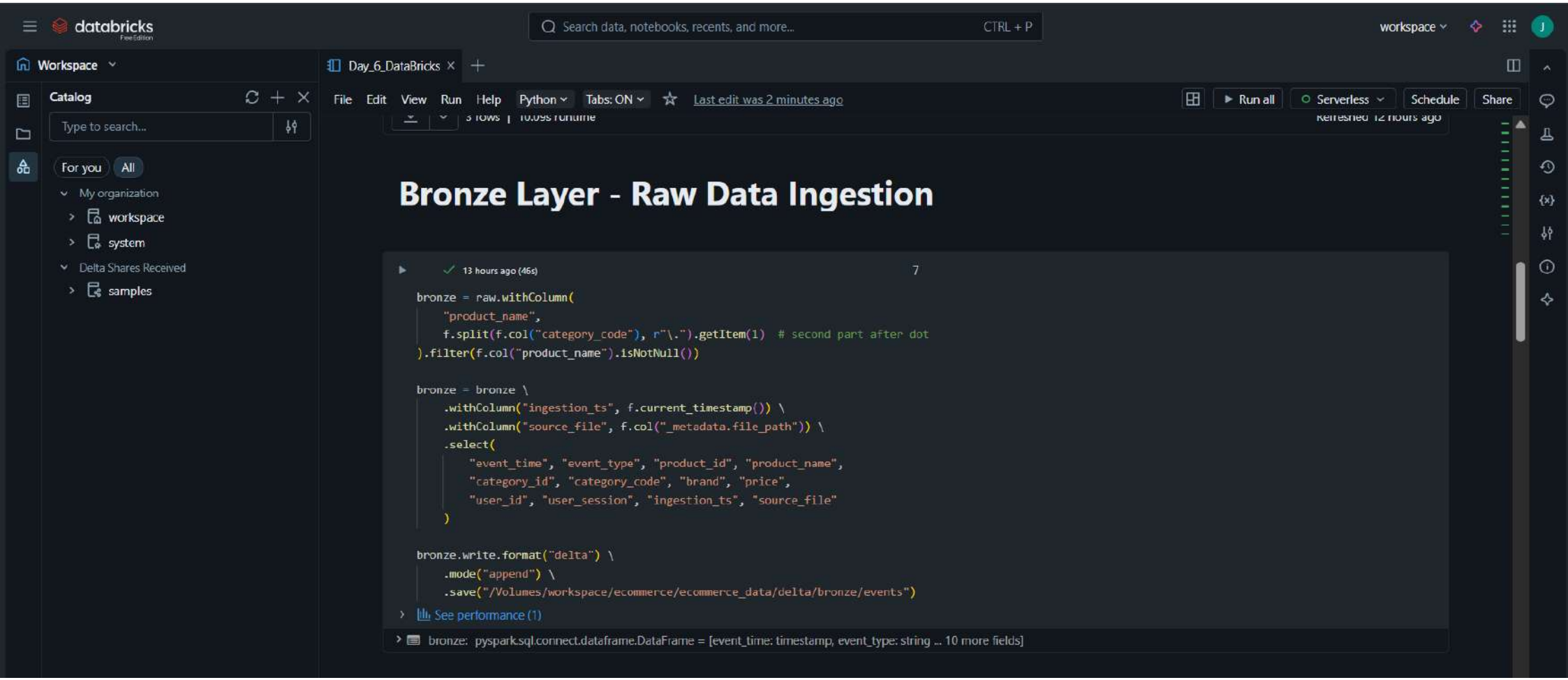
1. Design 3-layer architecture
2. Build Bronze: raw ingestion
3. Build Silver: cleaning & validation
4. Build Gold: business aggregates

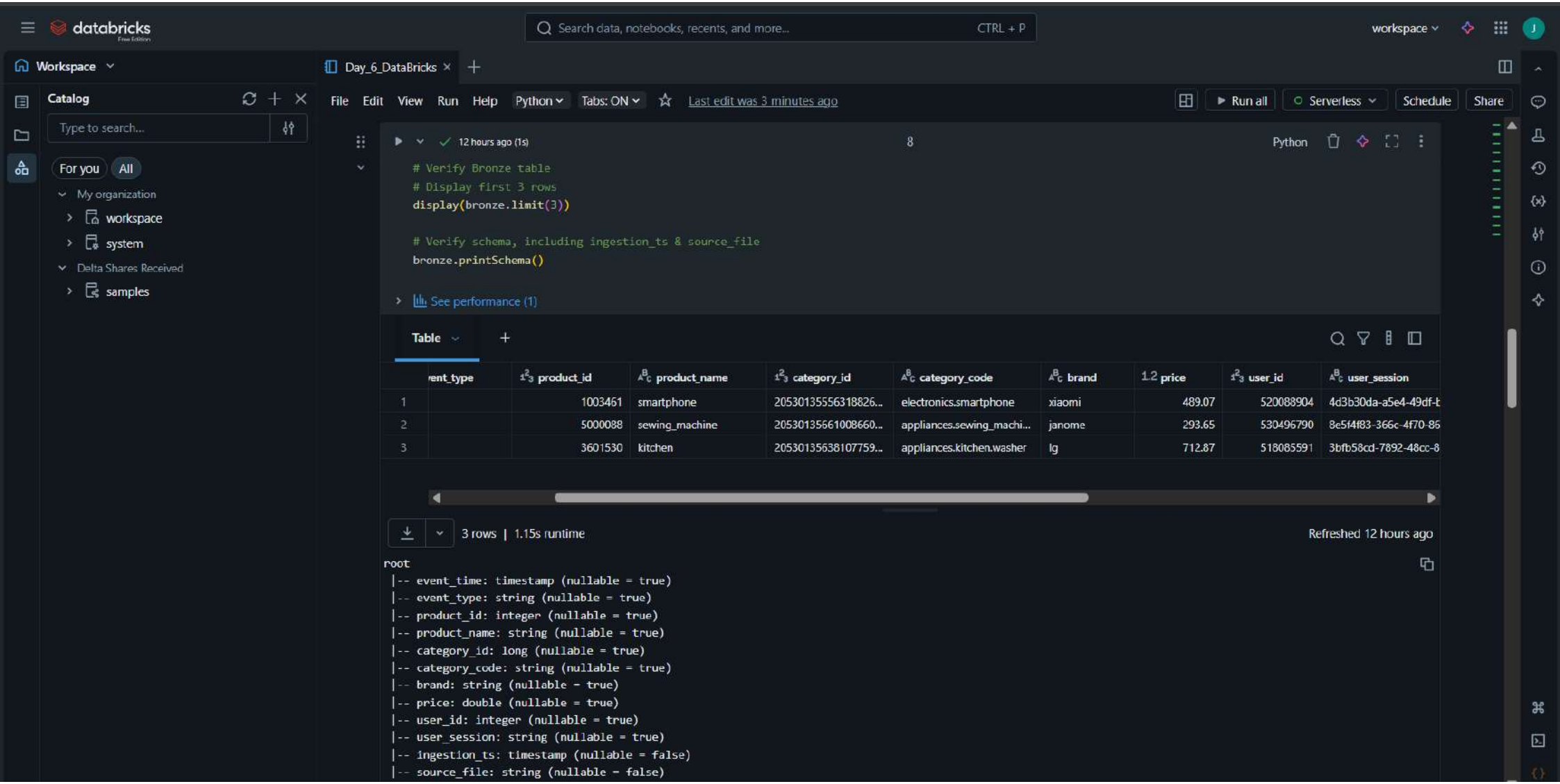
#DatabricksWithIDC

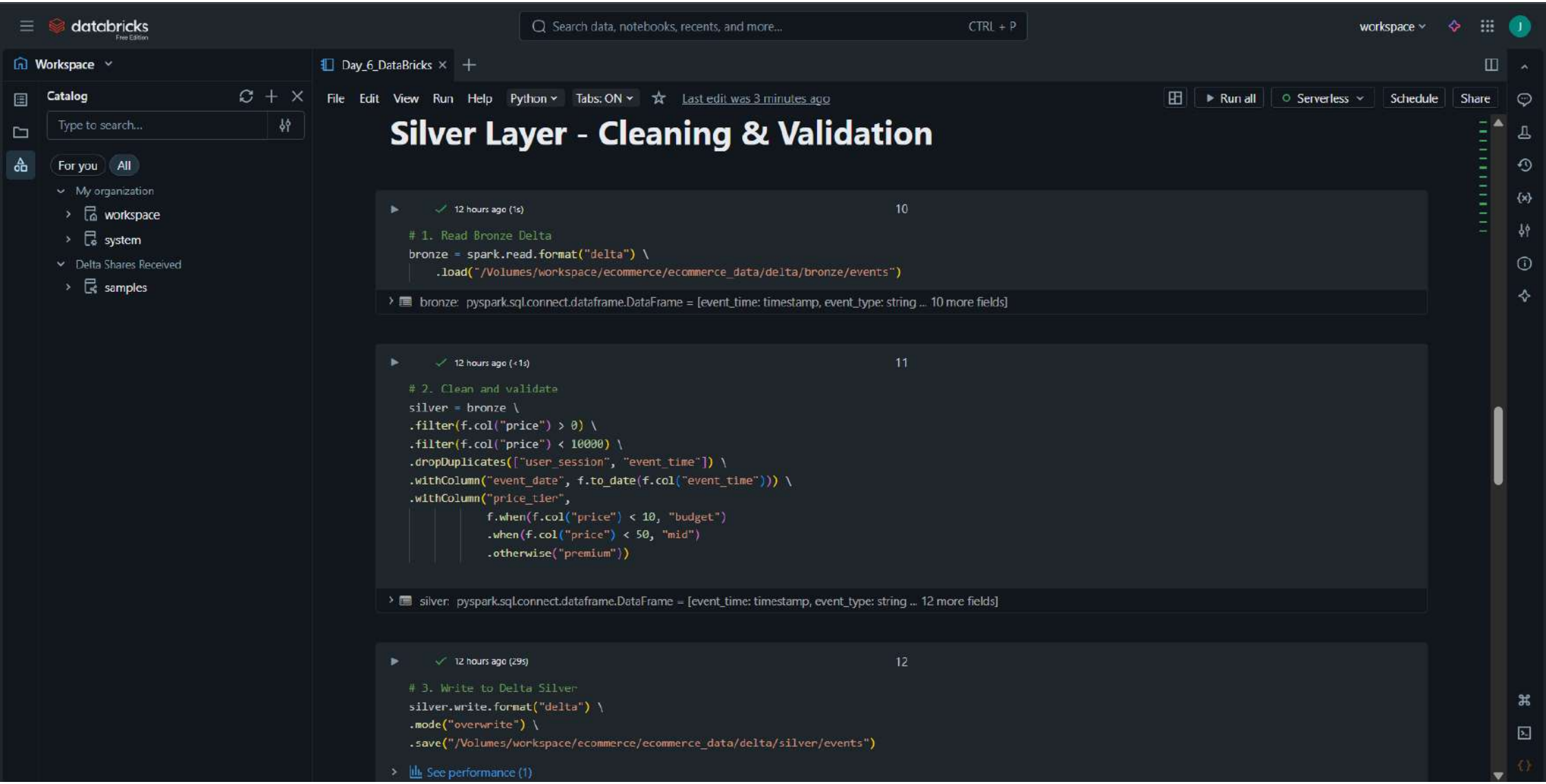
Medallion Architecture in the Lakehouse

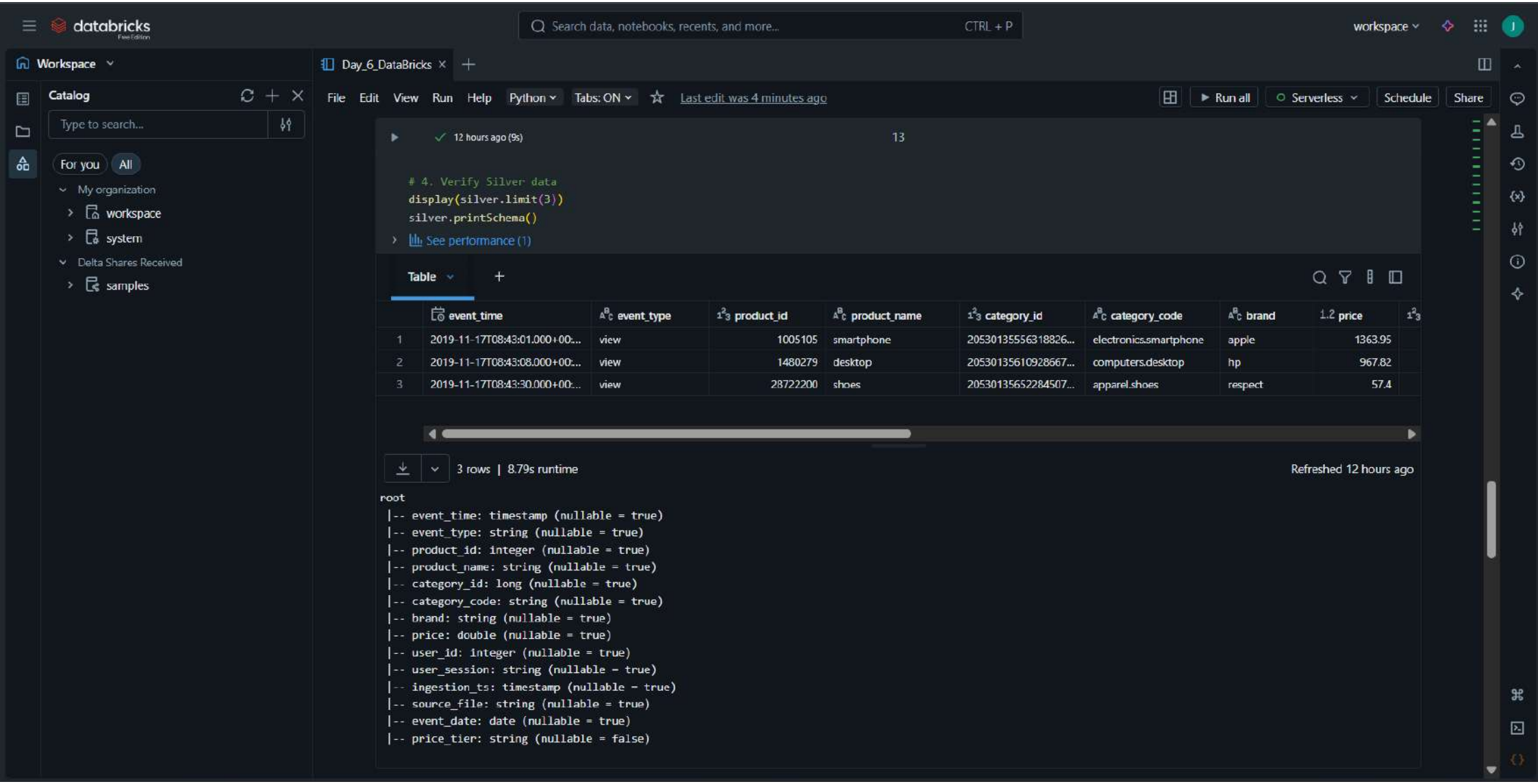


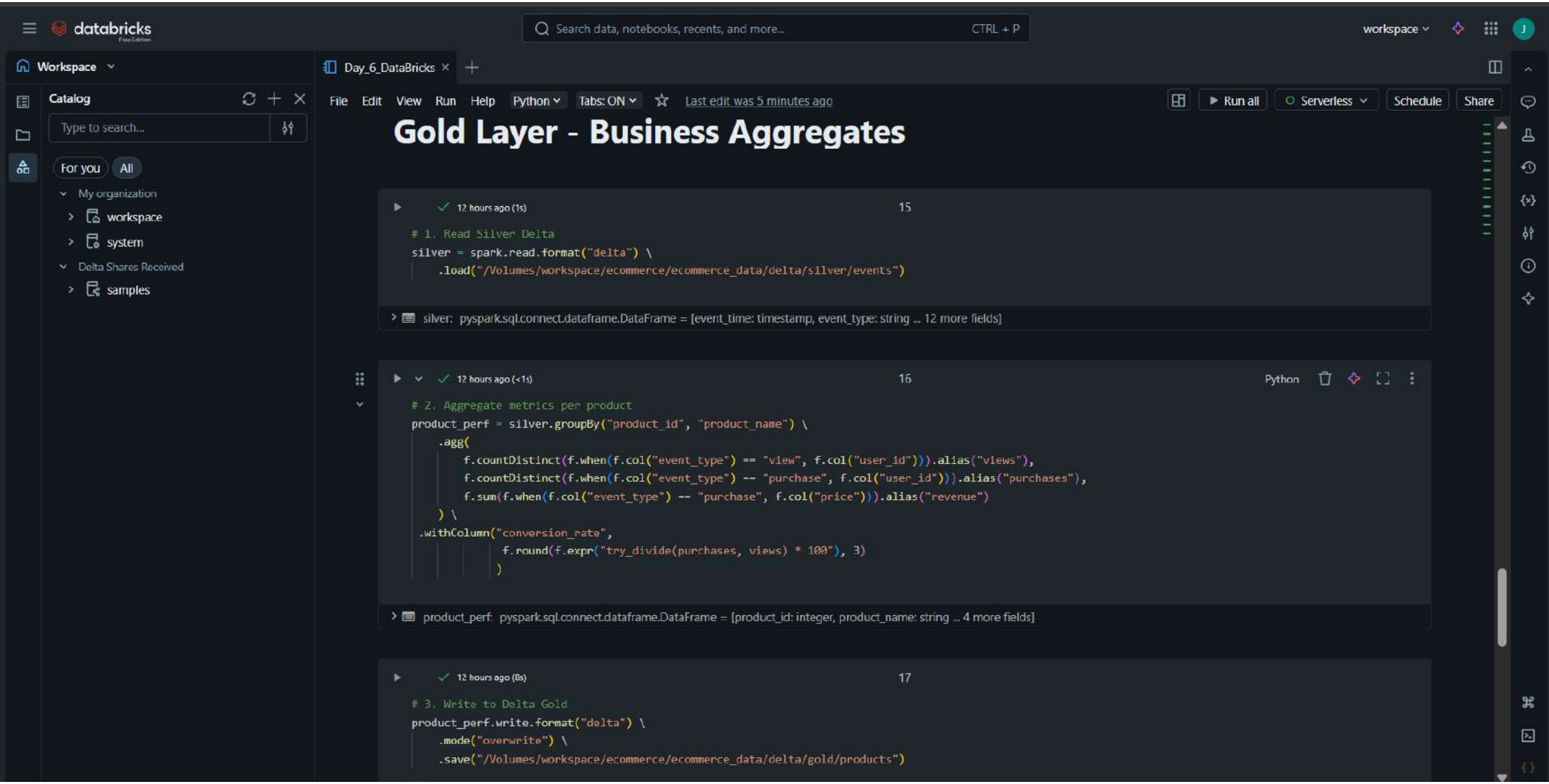

databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
Workspace
Day_6_DataBricks
File Edit View Run Help Python Tabs: ON Last edit was now
Run all Serverless Schedule Share
Catalog
Type to search...
For you All
My organization
workspace
system
Delta Shares Received
samples
Import required libraries and functions
13 hours ago (1s) 2
from pyspark.sql import functions as f
from delta.tables import DeltaTable
Ingest Raw Data
13 hours ago (1m) 4
raw = spark.read.csv(
"/Volumes/workspace/ecommerce/ecommerce_data/2019-Nov.csv",
header=True,
inferSchema=True
)
raw.printSchema()
raw: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 7 more fields]
root
|-- event_time: timestamp (nullable = true)
|-- event_type: string (nullable = true)
|-- product_id: integer (nullable = true)
|-- category_id: long (nullable = true)
|-- category_code: string (nullable = true)
|-- brand: string (nullable = true)
|-- price: double (nullable = true)
|-- user_id: integer (nullable = true)
|-- user_session: string (nullable = true)

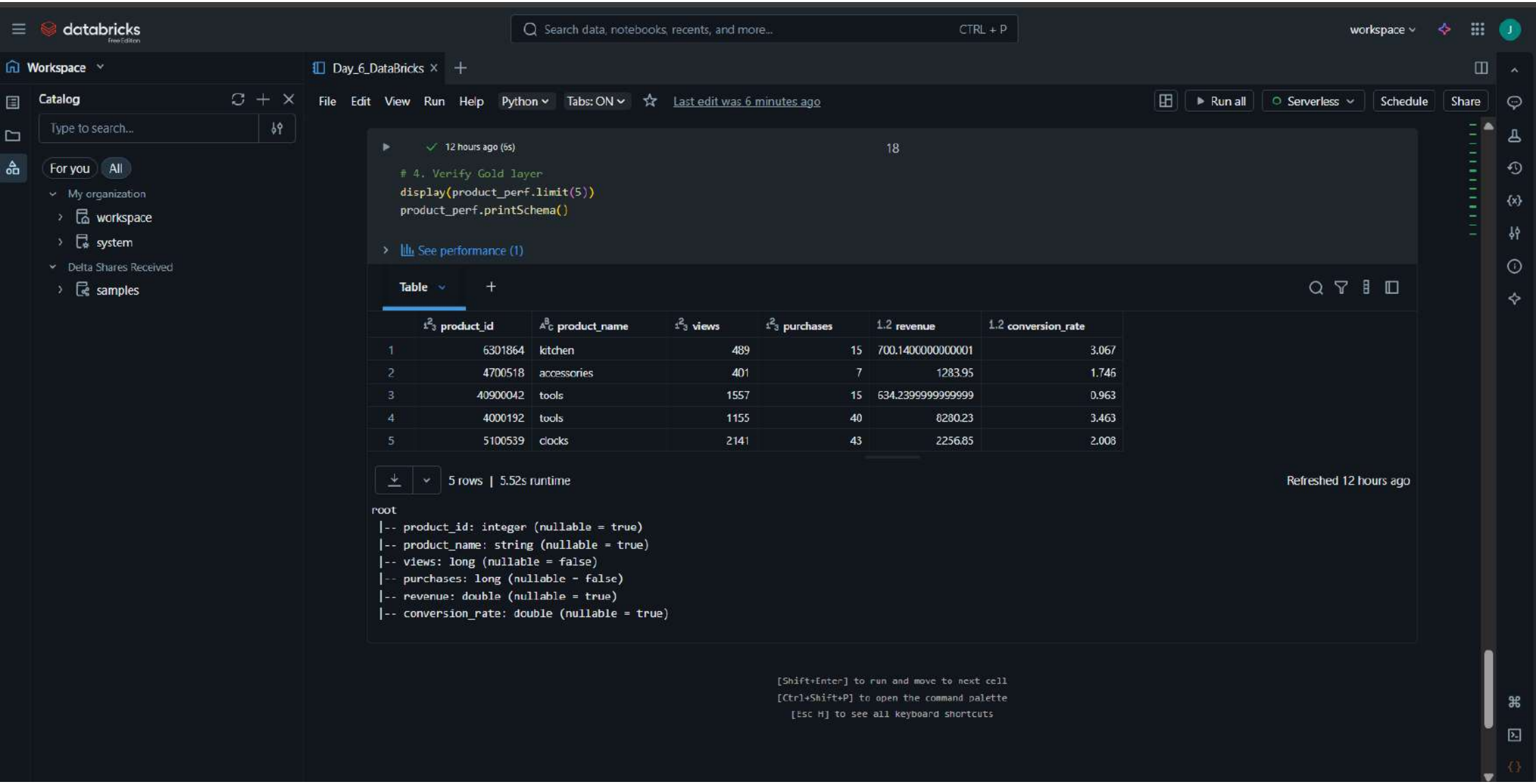

databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_6_DataBricks
Catalog
Type to search...
For you
All
My organization
workspace
system
Delta Shares Received
samples
File
Edit
View
Run
Help
Python
Tabs: ON
Last edit was 2 minutes ago
Run all
Serverless
Schedule
Share
Bronze Layer - Raw Data Ingestion
13 hours ago (46s)
bronze = raw.withColumn(
"product_name",
f.split(f.col("category_code"), r"\.").getItem(1) # second part after dot
).filter(f.col("product_name").isNotNull())
bronze = bronze \
.withColumn("ingestion_ts", f.current_timestamp()) \
.withColumn("source_file", f.col("_metadata.file_path")) \
.select(
"event_time", "event_type", "product_id", "product_name",
"category_id", "category_code", "brand", "price",
"user_id", "user_session", "ingestion_ts", "source_file"
)
bronze.write.format("delta") \
.mode("append") \
.save("/Volumes/workspace/ecommerce/ecommerce_data/delta/bronze/events")
See performance (1)
bronze: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 10 more fields]


databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
Workspace
Day_6_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 3 minutes ago
Run all
Serverless
Schedule
Share
12 hours ago (1s)
8
Python
# Verify Bronze table
# Display first 3 rows
display(bronze.limit(3))
# Verify schema, including ingestion_ts & source_file
bronze.printSchema()
See performance (1)
Table
ent_type
product_id
product_name
category_id
category_code
brand
price
user_id
user_session
1 1003461 smartphone 20530135556318826... electronics.smartphone xiaomi 489.07 520088904 4d3b30da-a5e4-49df-b
2 5000088 sewing_machine 20530135661008660... appliances.sewing_machi... janome 293.65 530496790 8e5f4f83-366c-4f70-86
3 3601530 kitchen 20530135638107759... appliances.kitchen.washer lg 712.87 518085591 3bfb58cd-7892-48cc-8
3 rows | 1.15s runtime
Refreshed 12 hours ago
root
|-- event_time: timestamp (nullable = true)
|-- event_type: string (nullable = true)
|-- product_id: integer (nullable = true)
|-- product_name: string (nullable = true)
|-- category_id: long (nullable = true)
|-- category_code: string (nullable = true)
|-- brand: string (nullable = true)
|-- price: double (nullable = true)
|-- user_id: integer (nullable = true)
|-- user_session: string (nullable = true)
|-- ingestion_ts: timestamp (nullable = false)
|-- source_file: string (nullable = false)


databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_6_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 3 minutes ago
Run all
Serverless
Schedule
Share
Silver Layer - Cleaning & Validation
12 hours ago (1s)
10
# 1. Read Bronze Delta
bronze = spark.read.format("delta") \
.load("/Volumes/workspace/ecommerce/ecommerce_data/delta/bronze/events")
bronze: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 10 more fields]
12 hours ago (<1s)
11
# 2. Clean and validate
silver = bronze \
.filter(f.col("price") > 0) \
.filter(f.col("price") < 10000) \
.dropDuplicates(["user_session", "event_time"]) \
.withColumn("event_date", f.to_date(f.col("event_time"))) \
.withColumn("price_tier",
f.when(f.col("price") < 10, "budget")
.when(f.col("price") < 50, "mid")
.otherwise("premium"))
silver: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 12 more fields]
12 hours ago (29s)
12
3. Write to Delta Silver
silver.write.format("delta") \
.mode("overwrite") \
.save("/Volumes/workspace/ecommerce/ecommerce_data/delta/silver/events")
See performance (1)


databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_6_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 4 minutes ago
Run all
Serverless
Schedule
Share
12 hours ago (9s)
13
# 4. Verify Silver data
display(silver.limit(3))
silver.printSchema()
See performance (1)
Table
event_time
event_type
product_id
product_name
category_id
category_code
brand
price
1 2019-11-17T08:43:01.000+00:... view 1005105 smartphone 20530135556318826... electronics.smartphone apple 1363.95
2 2019-11-17T08:43:08.000+00:... view 1480279 desktop 2053013561092866 7... computers.desktop hp 967.82
3 2019-11-17T08:43:30.000+00:... view 28722200 shoes 20530135652284507... apparel.shoes respect 57.4
3 rows | 8.79s runtime
Refreshed 12 hours ago
root
|-- event_time: timestamp (nullable = true)
|-- event_type: string (nullable = true)
|-- product_id: integer (nullable = true)
|-- product_name: string (nullable = true)
|-- category_id: long (nullable = true)
|-- category_code: string (nullable = true)
|-- brand: string (nullable = true)
|-- price: double (nullable = true)
|-- user_id: integer (nullable = true)
|-- user_session: string (nullable = true)
|-- ingestion_ts: timestamp (nullable = true)
|-- source_file: string (nullable = true)
|-- event_date: date (nullable = true)
|-- price_tier: string (nullable = false)


databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
Workspace
Day_6_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 5 minutes ago
Run all
Serverless
Schedule
Share
Gold Layer - Business Aggregates
12 hours ago (1s)
15
# 1. Read Silver Delta
silver = spark.read.format("delta") \
.load("/Volumes/workspace/ecommerce/ecommerce_data/delta/silver/events")
silver: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 12 more fields]
12 hours ago (<1s)
16
Python
# 2. Aggregate metrics per product
product_perf = silver.groupBy("product_id", "product_name") \
.agg(
f.countDistinct(f.when(f.col("event_type") == "view", f.col("user_id"))).alias("views"),
f.countDistinct(f.when(f.col("event_type") == "purchase", f.col("user_id"))).alias("purchases"),
f.sum(f.when(f.col("event_type") == "purchase", f.col("price"))).alias("revenue")
) \
.withColumn("conversion_rate",
f.round(f.expr("try_divide(purchases, views) * 100"), 3)
)
product_perf: pyspark.sql.connect.dataframe.DataFrame = [product_id: integer, product_name: string ... 4 more fields]
12 hours ago (0s)
17
3. Write to Delta Gold
product_perf.write.format("delta") \
.mode("overwrite") \
.save("/Volumes/workspace/ecommerce/ecommerce_data/delta/gold/products")


databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_6_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 6 minutes ago
Run all
Serverless
Schedule
Share
12 hours ago (6s)
18
# 4. Verify Gold layer
display(product_perf.limit(5))
product_perf.printSchema()
See performance (1)
Table
product_id
product_name
views
purchases
revenue
conversion_rate
1 6301864 kitchen 489 15 700.1400000000001 3.067
2 4700518 accessories 401 7 1283.95 1.746
3 40900042 tools 1557 15 634.2399999999999 0.963
4 4000192 tools 1155 40 8280.23 3.463
5 5100539 clocks 2141 43 2256.85 2.008
5 rows | 5.52s runtime
Refreshed 12 hours ago
root
|-- product_id: integer (nullable = true)
|-- product_name: string (nullable = true)
|-- views: long (nullable = false)
|-- purchases: long (nullable = false)
|-- revenue: double (nullable = true)
|-- conversion_rate: double (nullable = true)
[Shift+Enter] to run and move to next cell
[Ctrl+Shift+P] to open the command palette
[Esc H] to see all keyboard shortcuts