INDIAN DATA CLUB

CODE BASICS

databricks

14 DAYS

AI CHALLENGE

DAY 02

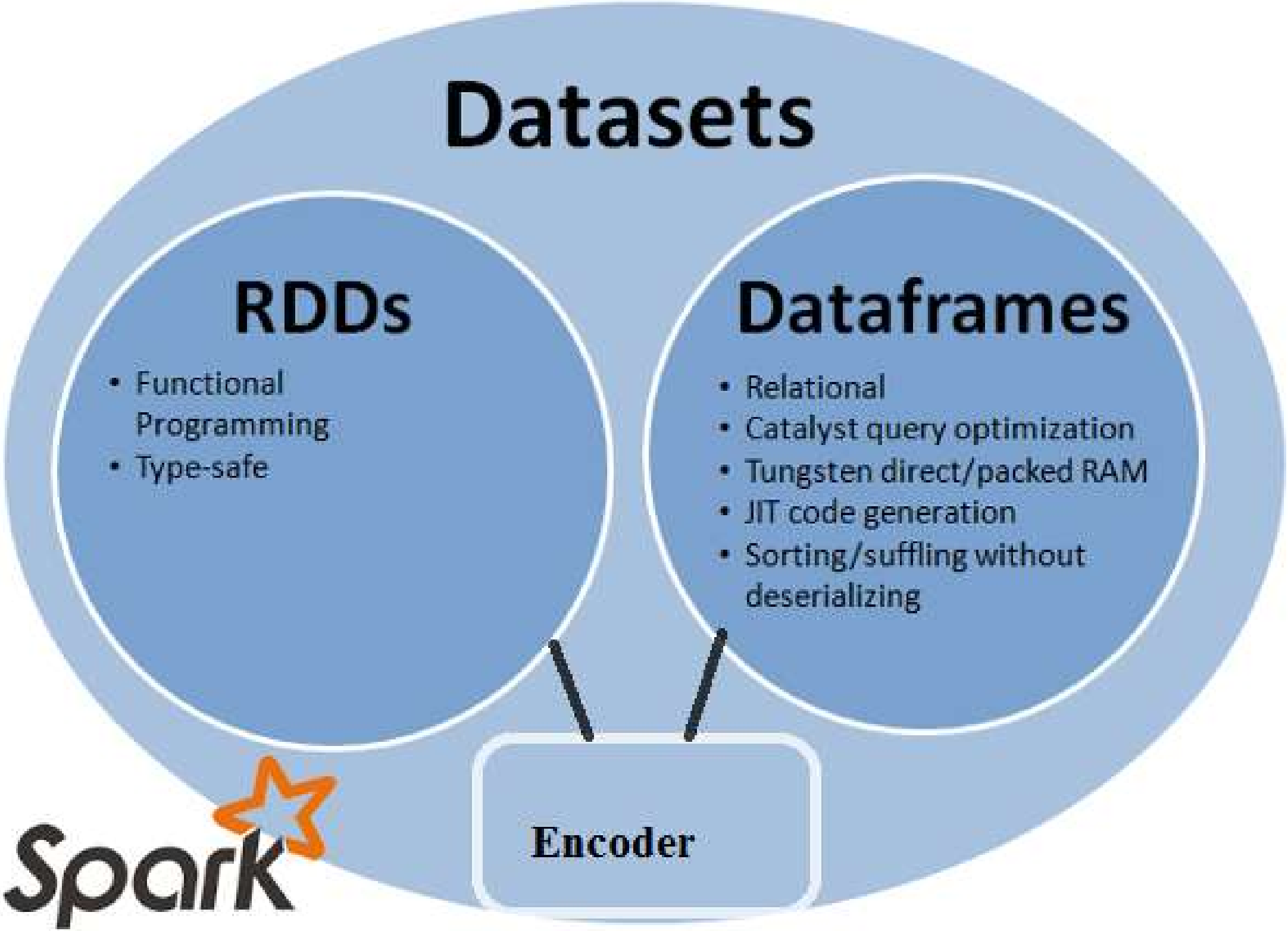
Topic:

Apache Spark Fundamentals

Challenge:

1. Upload sample e-commerce CSV
2. Read data into DataFrame
3. Basic operations: select, filter, groupBy, orderBy
4. Export results

#DatabricksWithIDC


Datasets
RDDs
• Functional Programming
• Type-safe
Dataframes
• Relational
• Catalyst query optimization
• Tungsten direct/packed RAM
• JIT code generation
• Sorting/suffling without deserializing
Encoder
Spark

databricks
Free Edition

Search data, notebooks, recents, and more... CTRL + P

Workspace

Day2 ×

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 1 minute ago

Read Data into Spark DataFrame

Just now (<1s) 2

```
df = spark.read.option("header", "true") \
                .option("inferSchema", "true") \
                .csv("workspace.default.ecommerce_data")
```

df: pyspark.sql.connect.dataframe.DataFrame

Basic Spark Operations

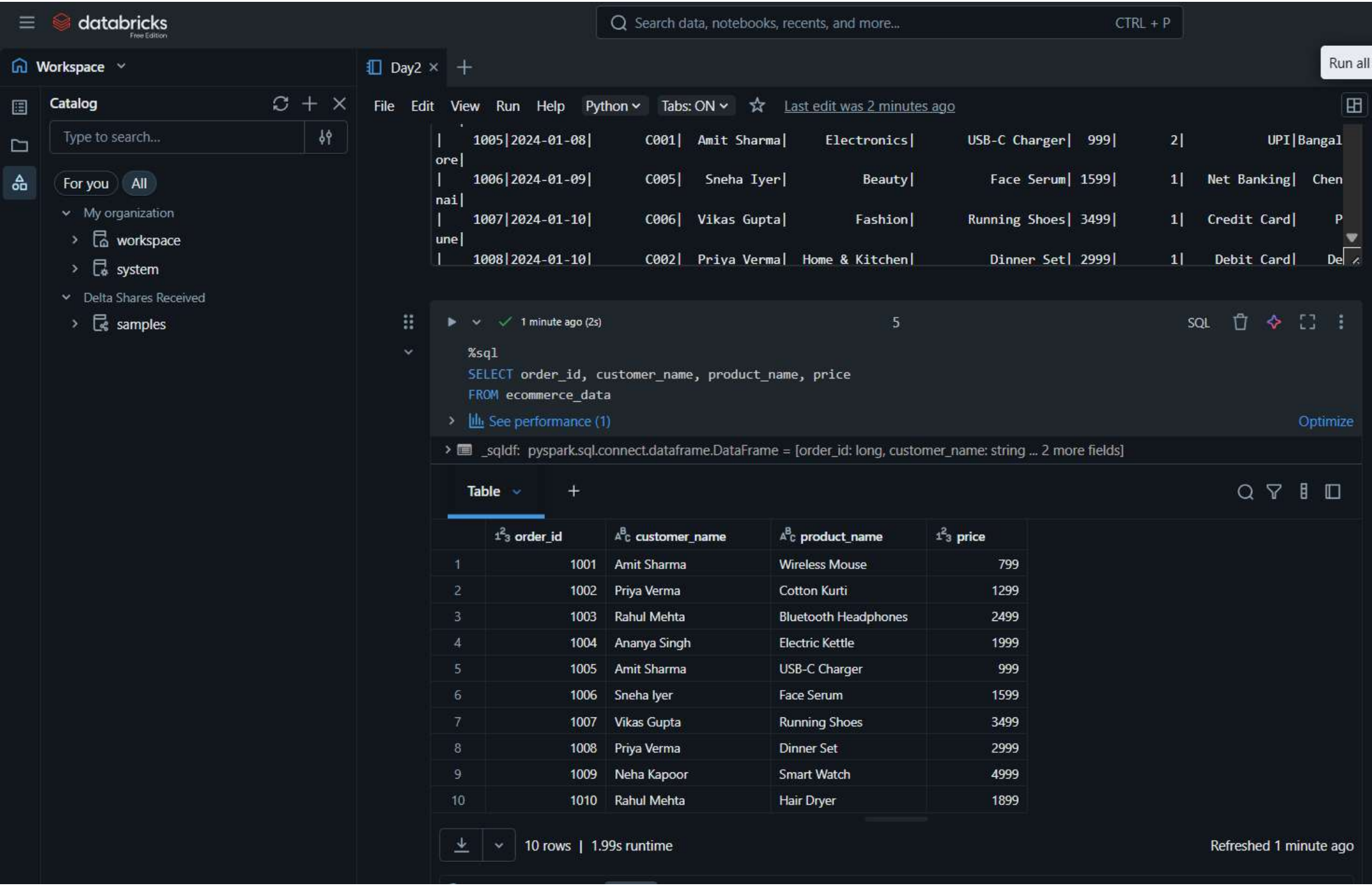
Just now (1s) 4

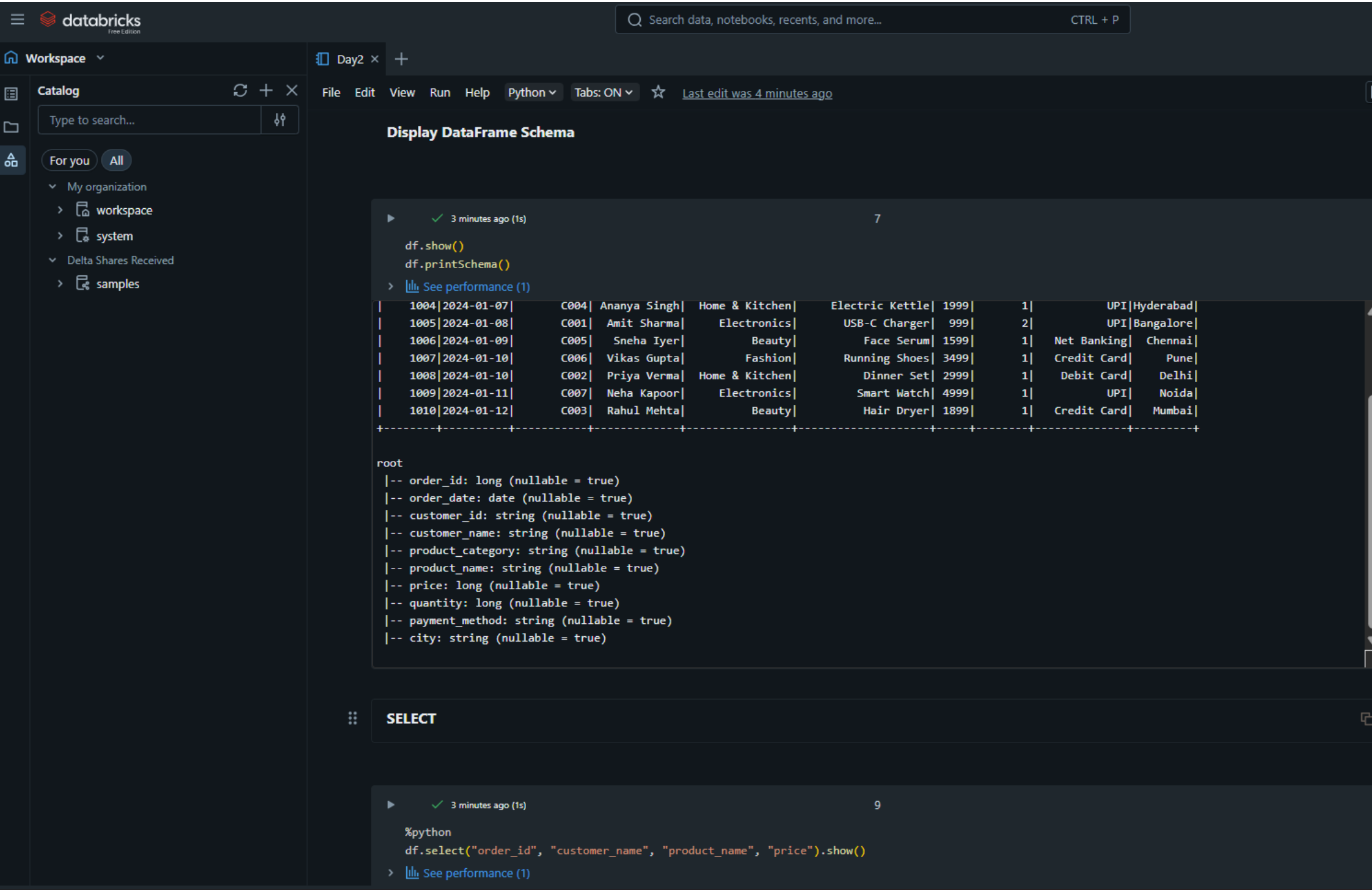
```
df = spark.table("workspace.default.ecommerce_data")
df.show()
```

See performance (1)

df: pyspark.sql.connect.dataframe.DataFrame = [order_id: long, order_date: date ... 8 more fields]

```
+--------+----------+-----------+-------------+----------------+--------------------+-----+--------+--------------+------
---+
|order_id|order_date|customer_id|customer_name|product_category|        product_name|price|quantity|payment_method|     c
ity|
+--------+----------+-----------+-------------+----------------+--------------------+-----+--------+--------------+------
---+
|    1001|2024-01-05|       C001|  Amit Sharma|     Electronics|      Wireless Mouse|  799|       1|          UPI|Bangal
ore|
|    1002|2024-01-06|       C002|  Priya Verma|         Fashion|       Cotton Kurti| 1299|       2|   Credit Card|    De
lhi|
```


databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
Workspace
Day2
Run all
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
| 1005|2024-01-08| C001| Amit Sharma| Electronics| USB-C Charger| 999| 2| UPI|Bangal
ore|
| 1006|2024-01-09| C005| Sneha Iyer| Beauty| Face Serum| 1599| 1| Net Banking| Chen
nai|
| 1007|2024-01-10| C006| Vikas Gupta| Fashion| Running Shoes| 3499| 1| Credit Card| P
une|
| 1008|2024-01-10| C002| Priya Verma| Home & Kitchen| Dinner Set| 2999| 1| Debit Card| De
1 minute ago (2s)
5
SQL
%sql
SELECT order_id, customer_name, product_name, price
FROM ecommerce_data
See performance (1)
Optimize
_sqldf: pyspark.sql.connect.dataframe.DataFrame = [order_id: long, customer_name: string ... 2 more fields]
Table
order_id
customer_name
product_name
price
1 1001 Amit Sharma Wireless Mouse 799
2 1002 Priya Verma Cotton Kurti 1299
3 1003 Rahul Mehta Bluetooth Headphones 2499
4 1004 Ananya Singh Electric Kettle 1999
5 1005 Amit Sharma USB-C Charger 999
6 1006 Sneha Iyer Face Serum 1599
7 1007 Vikas Gupta Running Shoes 3499
8 1008 Priya Verma Dinner Set 2999
9 1009 Neha Kapoor Smart Watch 4999
10 1010 Rahul Mehta Hair Dryer 1899
10 rows | 1.99s runtime
Refreshed 1 minute ago


databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
Workspace
Day2
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
Display DataFrame Schema
3 minutes ago (1s)
7
df.show()
df.printSchema()
See performance (1)
1004	2024-01-07	C004	Ananya Singh	Home & Kitchen	Electric Kettle	1999	1	UPI	Hyderabad
1005	2024-01-08	C001	Amit Sharma	Electronics	USB-C Charger	999	2	UPI	Bangalore
1006	2024-01-09	C005	Sneha Iyer	Beauty	Face Serum	1599	1	Net Banking	Chennai
1007	2024-01-10	C006	Vikas Gupta	Fashion	Running Shoes	3499	1	Credit Card	Pune
1008	2024-01-10	C002	Priya Verma	Home & Kitchen	Dinner Set	2999	1	Debit Card	Delhi
1009	2024-01-11	C007	Neha Kapoor	Electronics	Smart Watch	4999	1	UPI	Noida
1010	2024-01-12	C003	Rahul Mehta	Beauty	Hair Dryer	1899	1	Credit Card	Mumbai
root
-- order_id: long (nullable = true)
-- order_date: date (nullable = true)
-- customer_id: string (nullable = true)
-- customer_name: string (nullable = true)
-- product_category: string (nullable = true)
-- product_name: string (nullable = true)
-- price: long (nullable = true)
-- quantity: long (nullable = true)
-- payment_method: string (nullable = true)
-- city: string (nullable = true)
SELECT
3 minutes ago (1s)
9
%python
df.select("order_id", "customer_name", "product_name", "price").show()
See performance (1)

databricks
Free Edition

Search data, notebooks, recents, and more... CTRL + P

Workspace

Day2

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

3 minutes ago (1s) 9 Python

```
%python
df.select("order_id", "customer_name", "product_name", "price").show()
```

See performance (1)

```
+--------+-------------+--------------------+-----+
|order_id|customer_name|        product_name|price|
+--------+-------------+--------------------+-----+
|    1001|  Amit Sharma|      Wireless Mouse|  799|
|    1002|  Priya Verma|        Cotton Kurti| 1299|
|    1003|  Rahul Mehta|Bluetooth Headphones| 2499|
|    1004| Ananya Singh|     Electric Kettle| 1999|
|    1005|  Amit Sharma|       USB-C Charger|  999|
|    1006|   Sneha Iyer|          Face Serum| 1599|
|    1007|  Vikas Gupta|       Running Shoes| 3499|
|    1008|  Priya Verma|          Dinner Set| 2999|
|    1009|  Neha Kapoor|         Smart Watch| 4999|
|    1010|  Rahul Mehta|          Hair Dryer| 1899|
+--------+-------------+--------------------+-----+
```

FILTER

4 minutes ago (1s) 11

```
df.filter(df.price > 2000).show()
```

See performance (1)

```
+--------+----------+-----------+-------------+----------------+--------------------+-----+--------+--------------+------+
|order_id|order_date|customer_id|customer_name|product_category|        product_name|price|quantity|payment_method|  city|
+--------+----------+-----------+-------------+----------------+--------------------+-----+--------+--------------+------+
|    1003|2024-01-06|       C003|  Rahul Mehta|     Electronics|Bluetooth Headphones| 2499|       1|    Debit Card|Mumbai|
|    1007|2024-01-10|       C006|  Vikas Gupta|         Fashion|       Running Shoes| 3499|       1|   Credit Card|  Pune|
|    1008|2024-01-10|       C002|  Priya Verma|  Home & Kitchen|          Dinner Set| 2999|       1|    Debit Card| Delhi|
|    1009|2024-01-11|       C007|  Neha Kapoor|     Electronics|         Smart Watch| 4999|       1|           UPI| Noida|
+--------+----------+-----------+-------------+----------------+--------------------+-----+--------+--------------+------+
```

databricks
Free Edition

Search data, notebooks, recents, and more... CTRL + P

Workspace

Day2

File Edit View Run Help Python Tabs: ON Last edit was 6 minutes ago

Catalog

Type to search...

For you All

My organization
- workspace
- system

Delta Shares Received
- samples

GROUP BY

4 minutes ago (1s) 13

```
from pyspark.sql.functions import sum

df.groupBy("product_category") \
  .agg(sum("price").alias("total_sales")) \
  .show()
```

See performance (1)

```
+----------------+-----------+
|product_category|total_sales|
+----------------+-----------+
|     Electronics|       9296|
|         Fashion|       4798|
|  Home & Kitchen|       4998|
|          Beauty|       3498|
+----------------+-----------+
```

+ Code + Text Assistant

ORDER BY

5 minutes ago (1s) 15

```
df.orderBy(df.price.desc()).show()
```

See performance (1)

```
+--------+----------+-----------+-------------+----------------+--------------------+-----+--------+--------------+---------+
|order_id|order_date|customer_id|customer_name|product_category|        product_name|price|quantity|payment_method|     city|
+--------+----------+-----------+-------------+----------------+--------------------+-----+--------+--------------+---------+
|    1009|2024-01-11|       C007|  Neha Kapoor|     Electronics|         Smart Watch| 4999|       1|           UPI|    Noida|
|    1007|2024-01-10|       C006|  Vikas Gupta|         Fashion|       Running Shoes| 3499|       1|   Credit Card|     Pune|
|    1008|2024-01-10|       C002|  Priya Verma|  Home & Kitchen|          Dinner Set| 2999|       1|    Debit Card|    Delhi|
|    1003|2024-01-06|       C003|  Rahul Mehta|     Electronics|Bluetooth Headphones| 2499|       1|    Debit Card|   Mumbai|
|    1004|2024-01-07|       C004| Ananya Singh|  Home & Kitchen|      Electric Kettle| 1999|       1|           UPI|Hyderabad|
|    1010|2024-01-12|       C003|  Rahul Mehta|          Beauty|          Hair Dryer| 1899|       1|   Credit Card|   Mumbai|
|    1006|2024-01-09|       C005|   Sneha Iyer|          Beauty|          Face Serum| 1599|       1|   Net Banking|  Chennai|
```