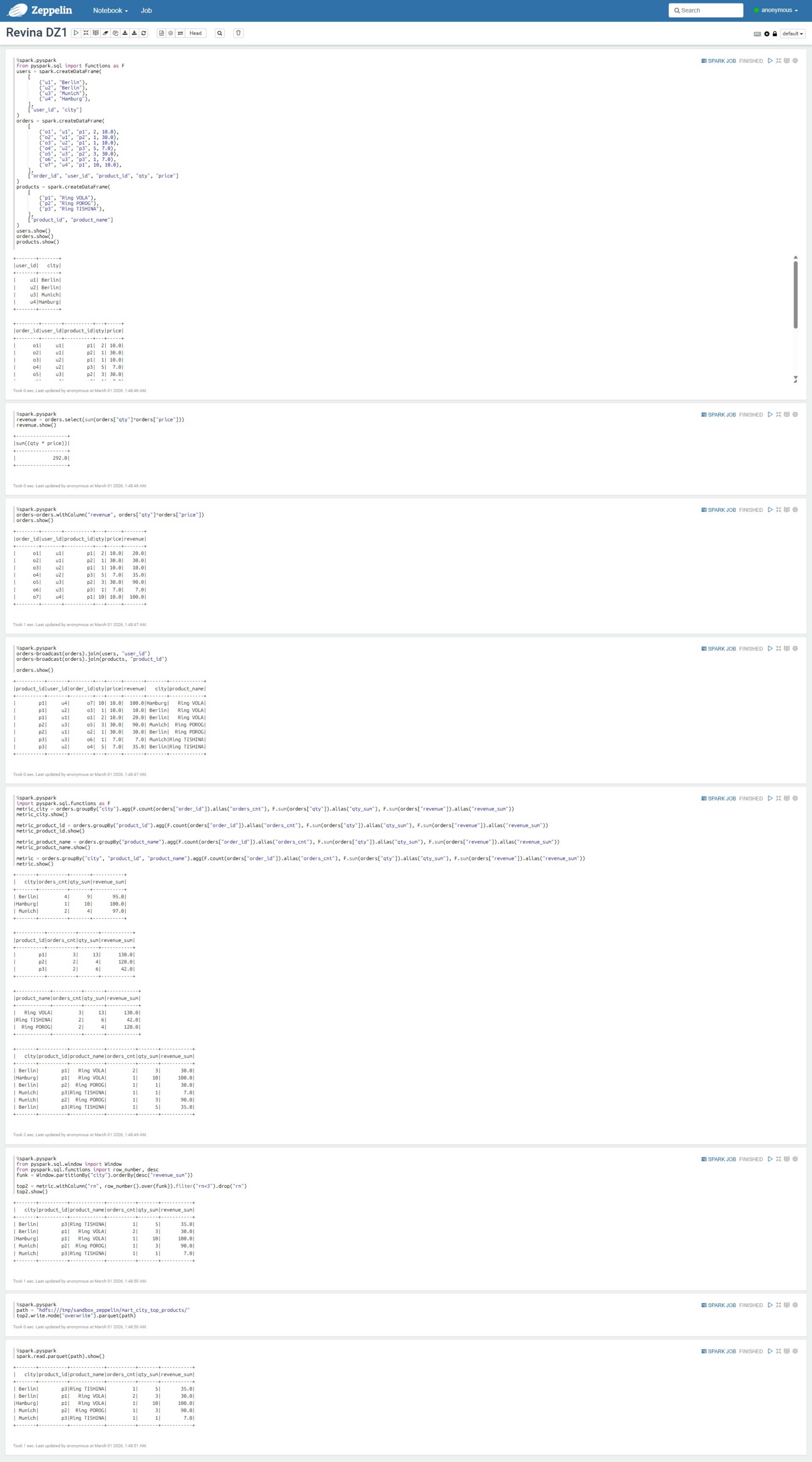

Revina DZ1

```
%spark.pyspark
from pyspark.sql import functions as F
users = spark.createDataFrame(
    [
        ("u1", "Berlin"),
        ("u2", "Berlin"),
        ("u3", "Munich"),
        ("u4", "Hamburg"),
    ],
    ["user_id", "city"]
)
orders = spark.createDataFrame(
    [
        ("o1", "u1", "p1", 2, 10.0),
        ("o2", "u1", "p2", 1, 30.0),
        ("o3", "u2", "p1", 1, 10.0),
        ("o4", "u2", "p3", 5, 7.0),
        ("o5", "u3", "p2", 3, 30.0),
        ("o6", "u3", "p3", 1, 7.0),
        ("o7", "u4", "p1", 10, 10.0),
    ],
    ["order_id", "user_id", "product_id", "qty", "price"]
)
products = spark.createDataFrame(
    [
        ("p1", "Ring VOLA"),
        ("p2", "Ring POROG"),
        ("p3", "Ring TISHINA"),
    ],
    ["product_id", "product_name"]
)
users.show()
orders.show()
products.show()
```

```
+-------+-------+
|user_id|   city|
+-------+-------+
|     u1| Berlin|
|     u2| Berlin|
|     u3| Munich|
|     u4|Hamburg|
+-------+-------+

+--------+-------+----------+---+-----+
|order_id|user_id|product_id|qty|price|
+--------+-------+----------+---+-----+
|      o1|     u1|        p1|  2| 10.0|
|      o2|     u1|        p2|  1| 30.0|
|      o3|     u2|        p1|  1| 10.0|
|      o4|     u2|        p3|  5|  7.0|
|      o5|     u3|        p2|  3| 30.0|
```

Took 0 sec. Last updated by anonymous at March 01 2026, 1:48:46 AM.

```
%spark.pyspark
revenue = orders.select(sum(orders["qty"]*orders["price"]))
revenue.show()
```

```
+------------------+
|sum((qty * price))|
+------------------+
|             292.0|
+------------------+
```

Took 0 sec. Last updated by anonymous at March 01 2026, 1:48:46 AM.

```
%spark.pyspark
orders=orders.withColumn("revenue", orders["qty"]*orders["price"])
orders.show()
```

```
+--------+-------+----------+---+-----+-------+
|order_id|user_id|product_id|qty|price|revenue|
+--------+-------+----------+---+-----+-------+
|      o1|     u1|        p1|  2| 10.0|   20.0|
|      o2|     u1|        p2|  1| 30.0|   30.0|
|      o3|     u2|        p1|  1| 10.0|   10.0|
|      o4|     u2|        p3|  5|  7.0|   35.0|
|      o5|     u3|        p2|  3| 30.0|   90.0|
|      o6|     u3|        p3|  1|  7.0|    7.0|
|      o7|     u4|        p1| 10| 10.0|  100.0|
+--------+-------+----------+---+-----+-------+
```

Took 1 sec. Last updated by anonymous at March 01 2026, 1:48:47 AM.

```
%spark.pyspark
orders=broadcast(orders).join(users, "user_id")
orders=broadcast(orders).join(products, "product_id")

orders.show()
```

```
+----------+-------+--------+---+-----+-------+-------+------------+
|product_id|user_id|order_id|qty|price|revenue|   city|product_name|
+----------+-------+--------+---+-----+-------+-------+------------+
|        p1|     u4|      o7| 10| 10.0|  100.0|Hamburg|   Ring VOLA|
|        p1|     u2|      o3|  1| 10.0|   10.0| Berlin|   Ring VOLA|
|        p1|     u1|      o1|  2| 10.0|   20.0| Berlin|   Ring VOLA|
|        p2|     u3|      o5|  3| 30.0|   90.0| Munich|  Ring POROG|
|        p2|     u1|      o2|  1| 30.0|   30.0| Berlin|  Ring POROG|
|        p3|     u3|      o6|  1|  7.0|    7.0| Munich|Ring TISHINA|
|        p3|     u2|      o4|  5|  7.0|   35.0| Berlin|Ring TISHINA|
+----------+-------+--------+---+-----+-------+-------+------------+
```

Took 0 sec. Last updated by anonymous at March 01 2026, 1:48:47 AM.

```
%spark.pyspark
import pyspark.sql.functions as F
metric_city = orders.groupBy("city").agg(F.count(orders["order_id"]).alias("orders_cnt"), F.sum(orders["qty"]).alias("qty_sum"), F.sum(orders["revenue"]).alias("revenue_sum"))
metric_city.show()

metric_product_id = orders.groupBy("product_id").agg(F.count(orders["order_id"]).alias("orders_cnt"), F.sum(orders["qty"]).alias("qty_sum"), F.sum(orders["revenue"]).alias("revenue_sum"))
metric_product_id.show()

metric_product_name = orders.groupBy("product_name").agg(F.count(orders["order_id"]).alias("orders_cnt"), F.sum(orders["qty"]).alias("qty_sum"), F.sum(orders["revenue"]).alias("revenue_sum"))
metric_product_name.show()

metric = orders.groupBy("city", "product_id", "product_name").agg(F.count(orders["order_id"]).alias("orders_cnt"), F.sum(orders["qty"]).alias("qty_sum"), F.sum(orders["revenue"]).alias("revenue_sum"))
metric.show()
```

```
+-------+----------+-------+-----------+
|   city|orders_cnt|qty_sum|revenue_sum|
+-------+----------+-------+-----------+
| Berlin|         4|      9|       95.0|
|Hamburg|         1|     10|      100.0|
| Munich|         2|      4|       97.0|
+-------+----------+-------+-----------+

+----------+----------+-------+-----------+
|product_id|orders_cnt|qty_sum|revenue_sum|
+----------+----------+-------+-----------+
|        p1|         3|     13|      130.0|
|        p2|         2|      4|      120.0|
|        p3|         2|      6|       42.0|
+----------+----------+-------+-----------+

+------------+----------+-------+-----------+
|product_name|orders_cnt|qty_sum|revenue_sum|
+------------+----------+-------+-----------+
|   Ring VOLA|         3|     13|      130.0|
|Ring TISHINA|         2|      6|       42.0|
|  Ring POROG|         2|      4|      120.0|
+------------+----------+-------+-----------+

+-------+----------+------------+----------+-------+-----------+
|   city|product_id|product_name|orders_cnt|qty_sum|revenue_sum|
+-------+----------+------------+----------+-------+-----------+
| Berlin|        p1|   Ring VOLA|         2|      3|       30.0|
|Hamburg|        p1|   Ring VOLA|         1|     10|      100.0|
| Berlin|        p2|  Ring POROG|         1|      1|       30.0|
| Munich|        p3|Ring TISHINA|         1|      1|        7.0|
| Munich|        p2|  Ring POROG|         1|      3|       90.0|
| Berlin|        p3|Ring TISHINA|         1|      5|       35.0|
+-------+----------+------------+----------+-------+-----------+
```

Took 2 sec. Last updated by anonymous at March 01 2026, 1:48:49 AM.

```
%spark.pyspark
from pyspark.sql.window import Window
from pyspark.sql.functions import row_number, desc
funk = Window.partitionBy("city").orderBy(desc("revenue_sum"))

top2 = metric.withColumn("rn", row_number().over(funk)).filter("rn<3").drop("rn")
top2.show()
```

```
+-------+----------+------------+----------+-------+-----------+
|   city|product_id|product_name|orders_cnt|qty_sum|revenue_sum|
+-------+----------+------------+----------+-------+-----------+
| Berlin|        p3|Ring TISHINA|         1|      5|       35.0|
| Berlin|        p1|   Ring VOLA|         2|      3|       30.0|
|Hamburg|        p1|   Ring VOLA|         1|     10|      100.0|
| Munich|        p2|  Ring POROG|         1|      3|       90.0|
| Munich|        p3|Ring TISHINA|         1|      1|        7.0|
+-------+----------+------------+----------+-------+-----------+
```

Took 1 sec. Last updated by anonymous at March 01 2026, 1:48:50 AM.

```
%spark.pyspark
path = "hdfs:///tmp/sandbox_zeppelin/mart_city_top_products/"
top2.write.mode("overwrite").parquet(path)
```

Took 0 sec. Last updated by anonymous at March 01 2026, 1:48:50 AM.

```
%spark.pyspark
spark.read.parquet(path).show()
```

```
+-------+----------+------------+----------+-------+-----------+
|   city|product_id|product_name|orders_cnt|qty_sum|revenue_sum|
+-------+----------+------------+----------+-------+-----------+
| Berlin|        p3|Ring TISHINA|         1|      5|       35.0|
| Berlin|        p1|   Ring VOLA|         2|      3|       30.0|
|Hamburg|        p1|   Ring VOLA|         1|     10|      100.0|
| Munich|        p2|  Ring POROG|         1|      3|       90.0|
| Munich|        p3|Ring TISHINA|         1|      1|        7.0|
+-------+----------+------------+----------+-------+-----------+
```

Took 1 sec. Last updated by anonymous at March 01 2026, 1:48:51 AM.