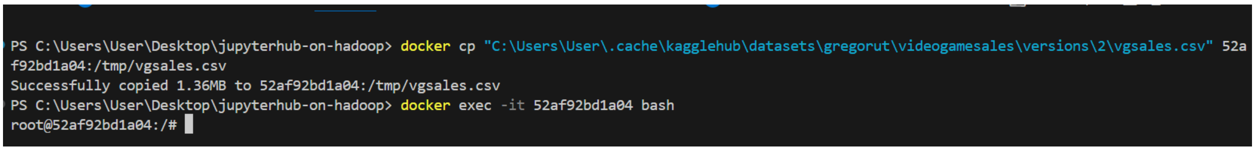
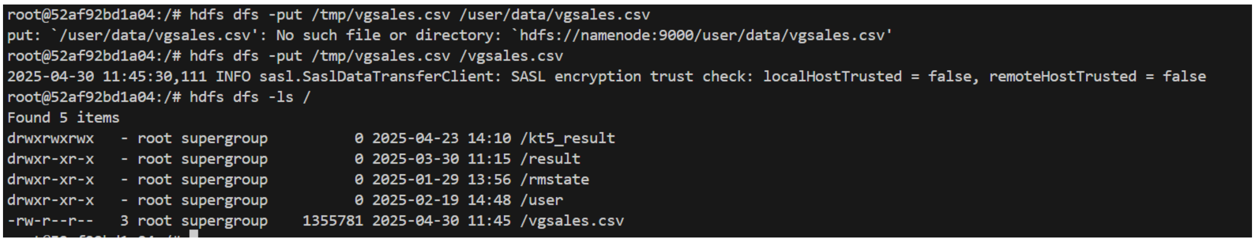
*Демонстрация ключевых команд hdfs*

1) Копирование файла в контейнер



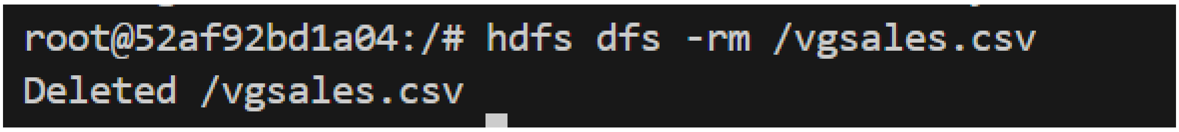
```
PS C:\Users\User\Desktop\jupyterhub-on-hadoop> docker cp "C:\Users\User\.cache\kagglehub\datasets\gregorut\videogamesales\versions\2\vgsales.csv" 52a
f92bd1a04:/tmp/vgsales.csv
Successfully copied 1.36MB to 52af92bd1a04:/tmp/vgsales.csv
PS C:\Users\User\Desktop\jupyterhub-on-hadoop> docker exec -it 52af92bd1a04 bash
root@52af92bd1a04:/#
```

2) Отправка датасета в корневую директорию hdfs



```
root@52af92bd1a04:/# hdfs dfs -put /tmp/vgsales.csv /user/data/vgsales.csv
put: `/user/data/vgsales.csv': No such file or directory: `hdfs://namenode:9000/user/data/vgsales.csv'
root@52af92bd1a04:/# hdfs dfs -put /tmp/vgsales.csv /vgsales.csv
2025-04-30 11:45:30,111 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@52af92bd1a04:/# hdfs dfs -ls /
Found 5 items
drwxrwxrwx   - root supergroup          0 2025-04-23 14:10 /kt5_result
drwxr-xr-x   - root supergroup          0 2025-03-30 11:15 /result
drwxr-xr-x   - root supergroup          0 2025-01-29 13:56 /rmstate
drwxr-xr-x   - root supergroup          0 2025-02-19 14:48 /user
-rw-r--r--   3 root supergroup    1355781 2025-04-30 11:45 /vgsales.csv
```

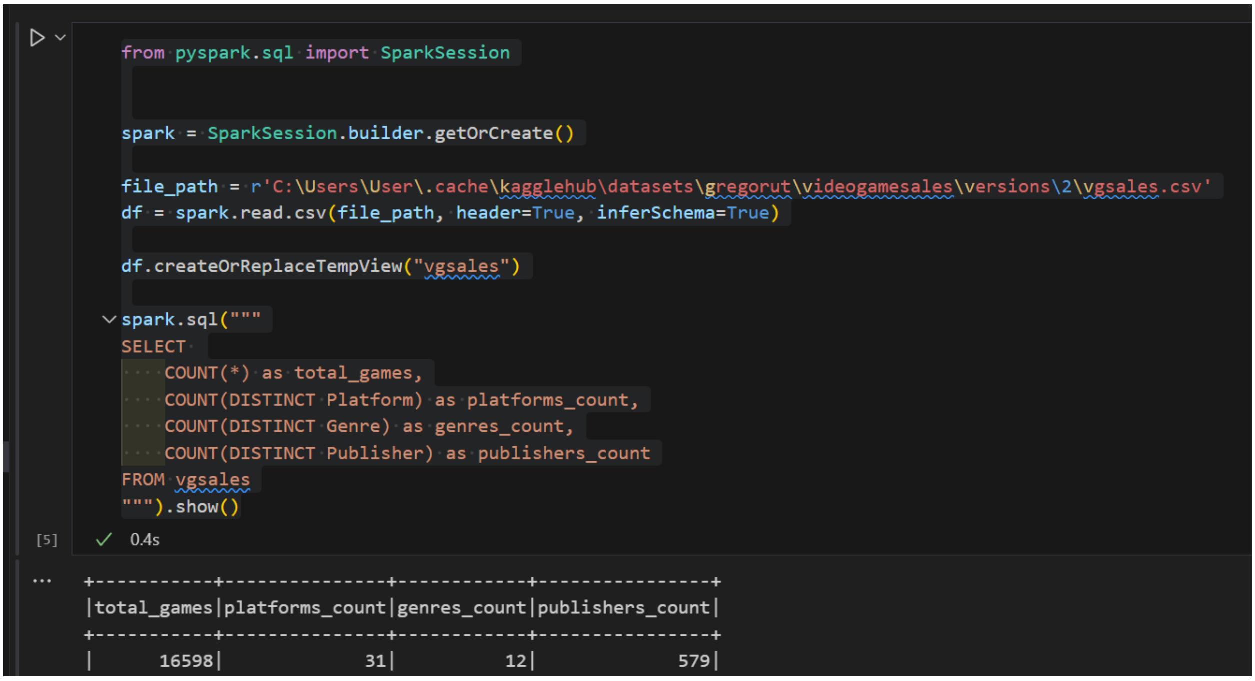
3) Удаление файла из hdfs



```
root@52af92bd1a04:/# hdfs dfs -rm /vgsales.csv
Deleted /vgsales.csv
```

*Отработка SQL-запросов через Spark SQL*

*1) Создание представления и вывод базовый сведений*

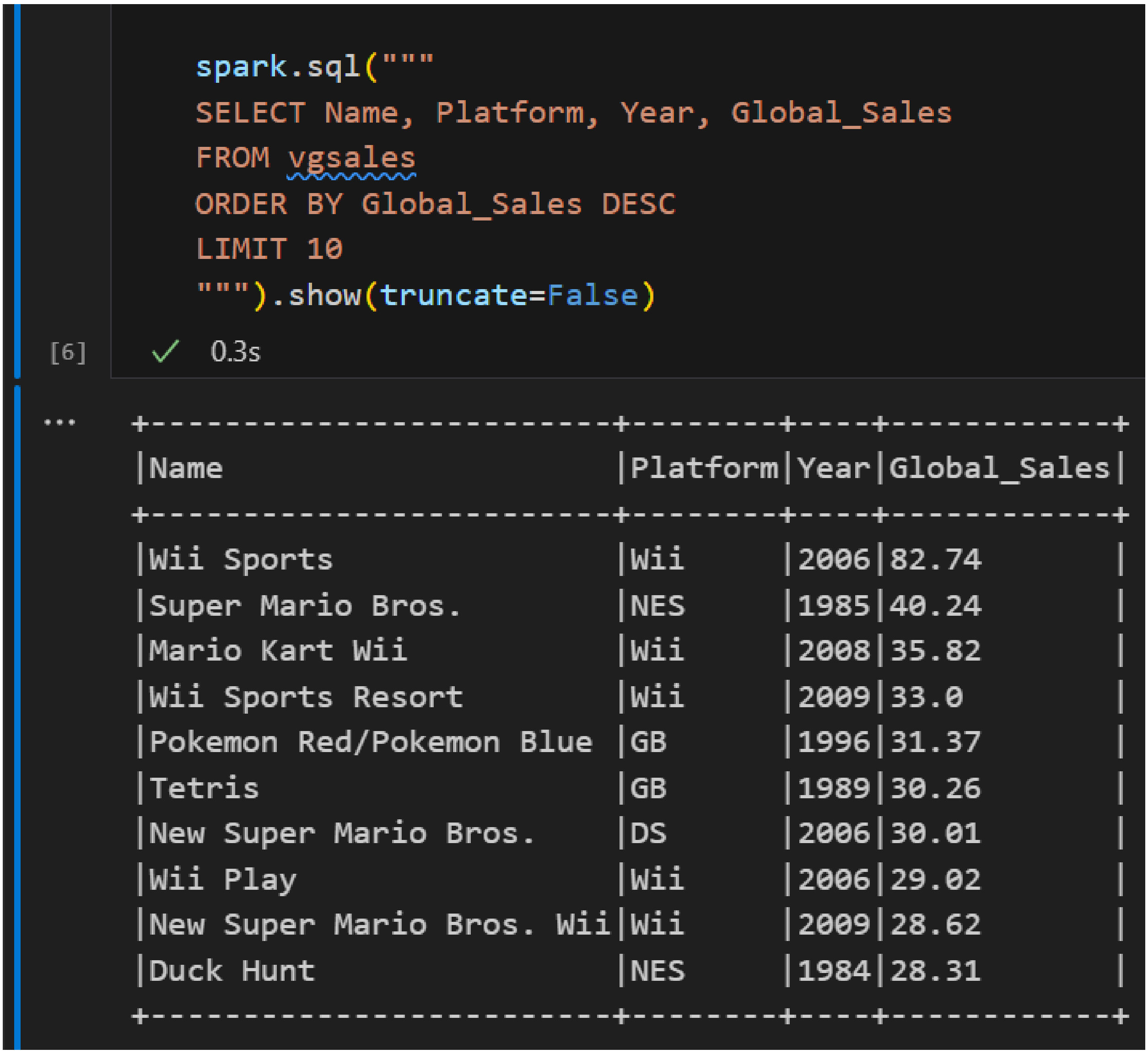


```
from pyspark.sql import SparkSession

spark = SparkSession.builder.getOrCreate()

file_path = r'C:\Users\User\.cache\kagglehub\datasets\gregorut\videogamesales\versions\2\vgsales.csv'
df = spark.read.csv(file_path, header=True, inferSchema=True)

df.createOrReplaceTempView("vgsales")

spark.sql("""
SELECT
    COUNT(*) as total_games,
    COUNT(DISTINCT Platform) as platforms_count,
    COUNT(DISTINCT Genre) as genres_count,
    COUNT(DISTINCT Publisher) as publishers_count
FROM vgsales
""").show()
```

[5] ✓ 0.4s

```
+-----------+---------------+------------+----------------+
|total_games|platforms_count|genres_count|publishers_count|
+-----------+---------------+------------+----------------+
|      16598|             31|          12|             579|
```

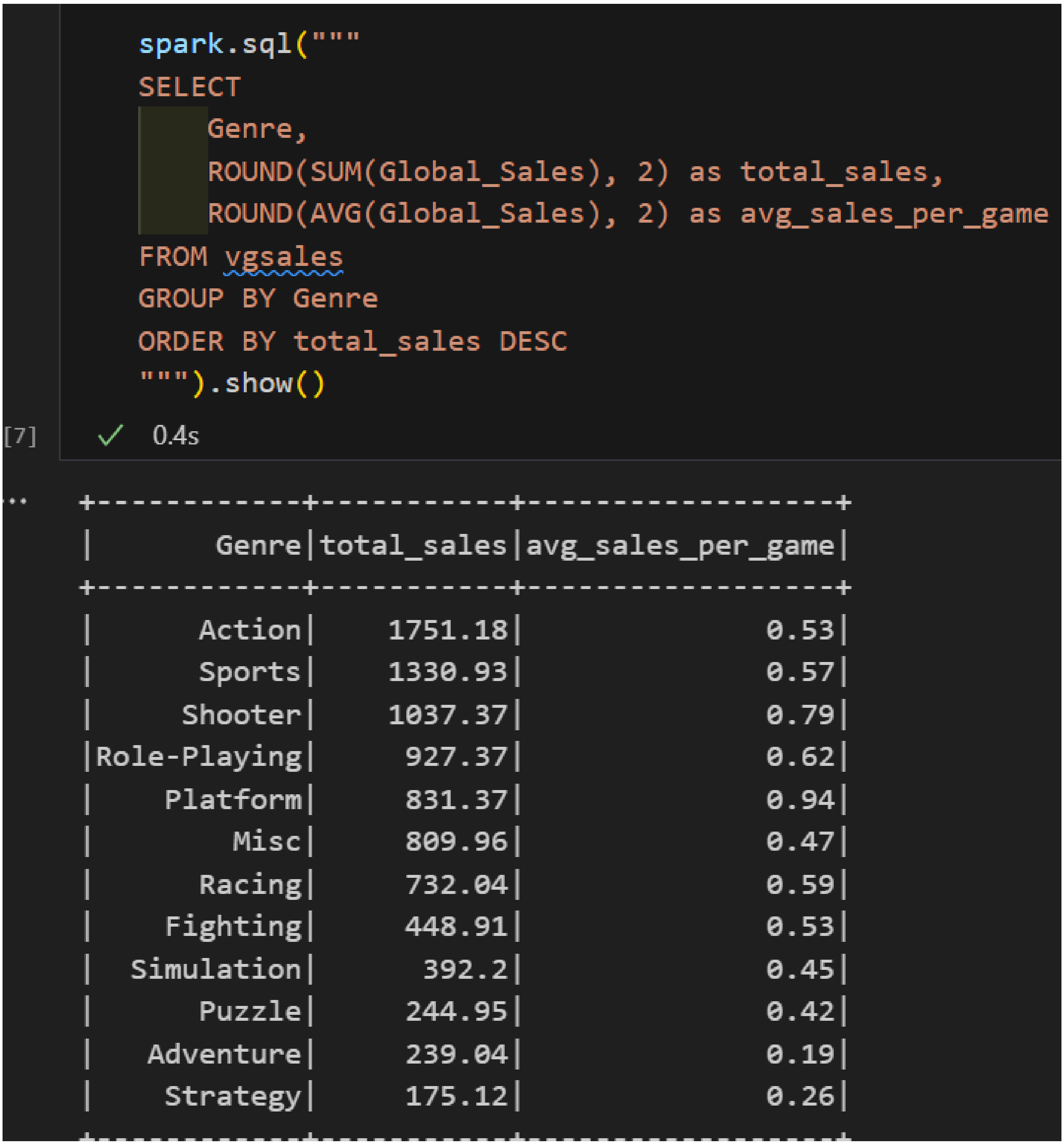
*2) Вывод десяти самых продаваемых игр*



```
spark.sql("""
SELECT Name, Platform, Year, Global_Sales
FROM vgsales
ORDER BY Global_Sales DESC
LIMIT 10
""").show(truncate=False)
```

```
+--------------------------+--------+----+------------+
|Name                      |Platform|Year|Global_Sales|
+--------------------------+--------+----+------------+
|Wii Sports                |Wii     |2006|82.74       |
|Super Mario Bros.         |NES     |1985|40.24       |
|Mario Kart Wii            |Wii     |2008|35.82       |
|Wii Sports Resort         |Wii     |2009|33.0        |
|Pokemon Red/Pokemon Blue  |GB      |1996|31.37       |
|Tetris                    |GB      |1989|30.26       |
|New Super Mario Bros.     |DS      |2006|30.01       |
|Wii Play                  |Wii     |2006|29.02       |
|New Super Mario Bros. Wii|Wii     |2009|28.62       |
|Duck Hunt                 |NES     |1984|28.31       |
+--------------------------+--------+----+------------+
```

*3) Вывод продажи по жанрам*



```
spark.sql("""
SELECT
    Genre,
    ROUND(SUM(Global_Sales), 2) as total_sales,
    ROUND(AVG(Global_Sales), 2) as avg_sales_per_game
FROM vgsales
GROUP BY Genre
ORDER BY total_sales DESC
""").show()
```

[7] ✓ 0.4s

```
+------------+-----------+------------------+
|       Genre|total_sales|avg_sales_per_game|
+------------+-----------+------------------+
|      Action|    1751.18|              0.53|
|      Sports|    1330.93|              0.57|
|     Shooter|    1037.37|              0.79|
|Role-Playing|     927.37|              0.62|
|    Platform|     831.37|              0.94|
|        Misc|     809.96|              0.47|
|      Racing|     732.04|              0.59|
|    Fighting|     448.91|              0.53|
|  Simulation|      392.2|              0.45|
|      Puzzle|     244.95|              0.42|
|   Adventure|     239.04|              0.19|
|    Strategy|     175.12|              0.26|
+------------+-----------+------------------+
```

*Итог*

Представленные hdfs команды и код на PySpark позволяют манипулировать данными внутри hdfs и предоставлять аналитику по предоставленному датасету