Nama : Muhammad Rifky Harto

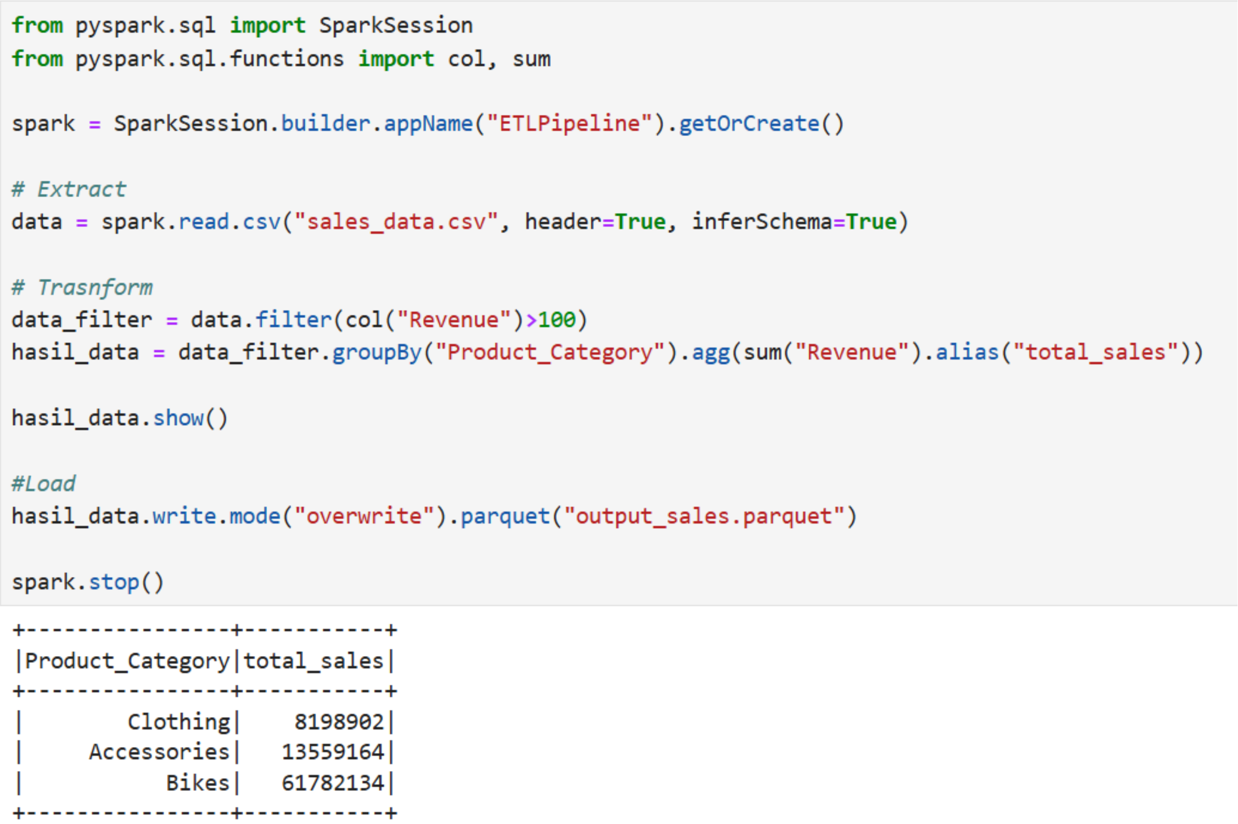
NIM : 2241720176

Kelas : TI-3D

Praktikum: Membangun ETL pipeline

Tugas

1. **Extract**: Baca data dari file CSV (sales_data.csv).
2. **Transform**:
   - Filter transaksi dengan Revenue > $100.
   - Hitung total penjualan per kategori.
3. **Load**: Simpan hasil ke Parquet.



```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, sum

spark = SparkSession.builder.appName("ETLPipeline").getOrCreate()

# Extract
data = spark.read.csv("sales_data.csv", header=True, inferSchema=True)

# Trasnform
data_filter = data.filter(col("Revenue")>100)
hasil_data = data_filter.groupBy("Product_Category").agg(sum("Revenue").alias("total_sales"))

hasil_data.show()

#Load
hasil_data.write.mode("overwrite").parquet("output_sales.parquet")

spark.stop()
```

```
+----------------+-----------+
|Product_Category|total_sales|
+----------------+-----------+
|        Clothing|    8198902|
|     Accessories|   13559164|
|           Bikes|   61782134|
+----------------+-----------+
```

Analisis Data Retail

Dataset

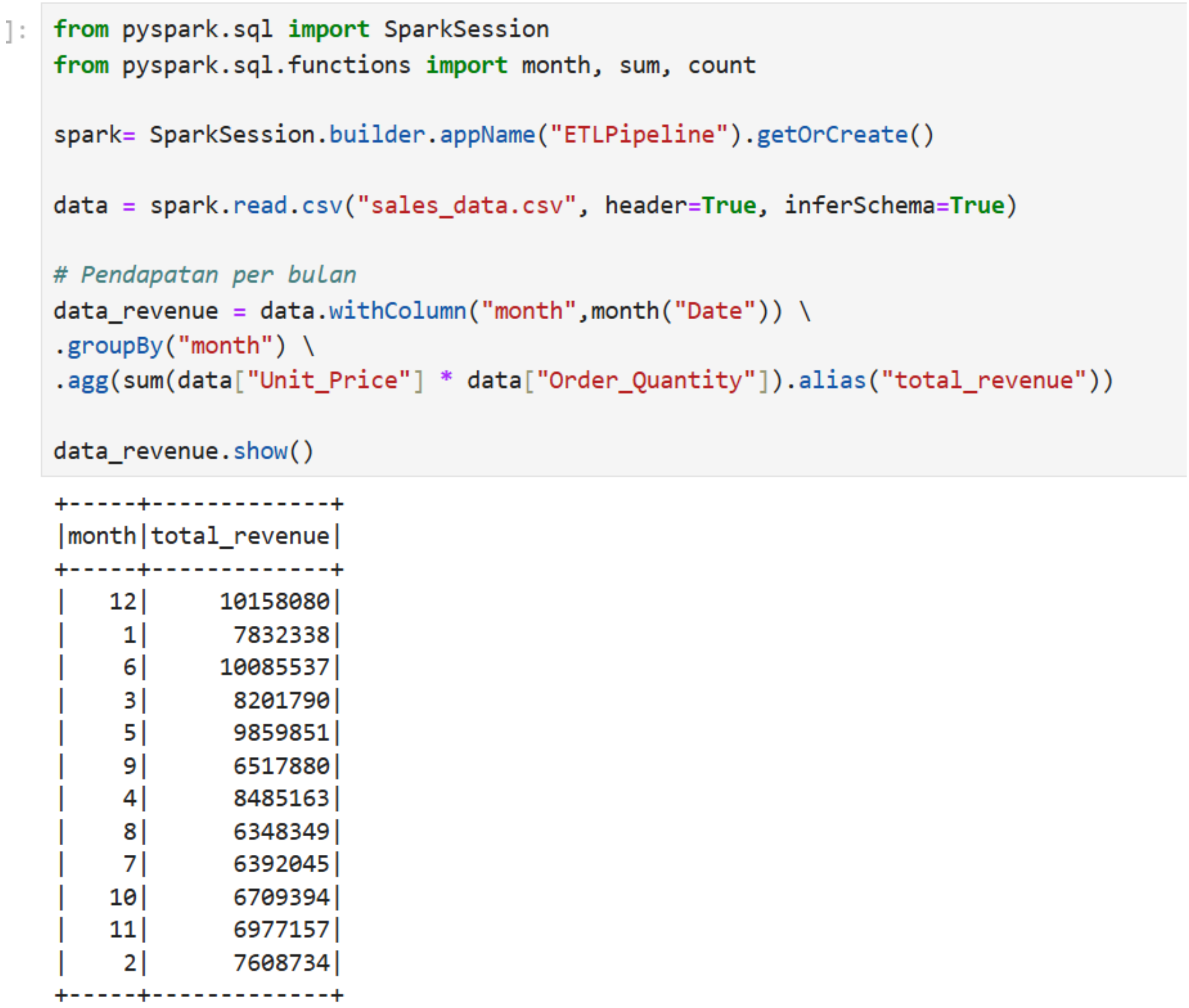
- **Format**: CSV (sales_data.csv)

Tugas

1. Hitung total pendapatan per bulan.
2. Identifikasi 5 produk terlaris.
3. Simpan hasil dalam format Parquet.

Solusi

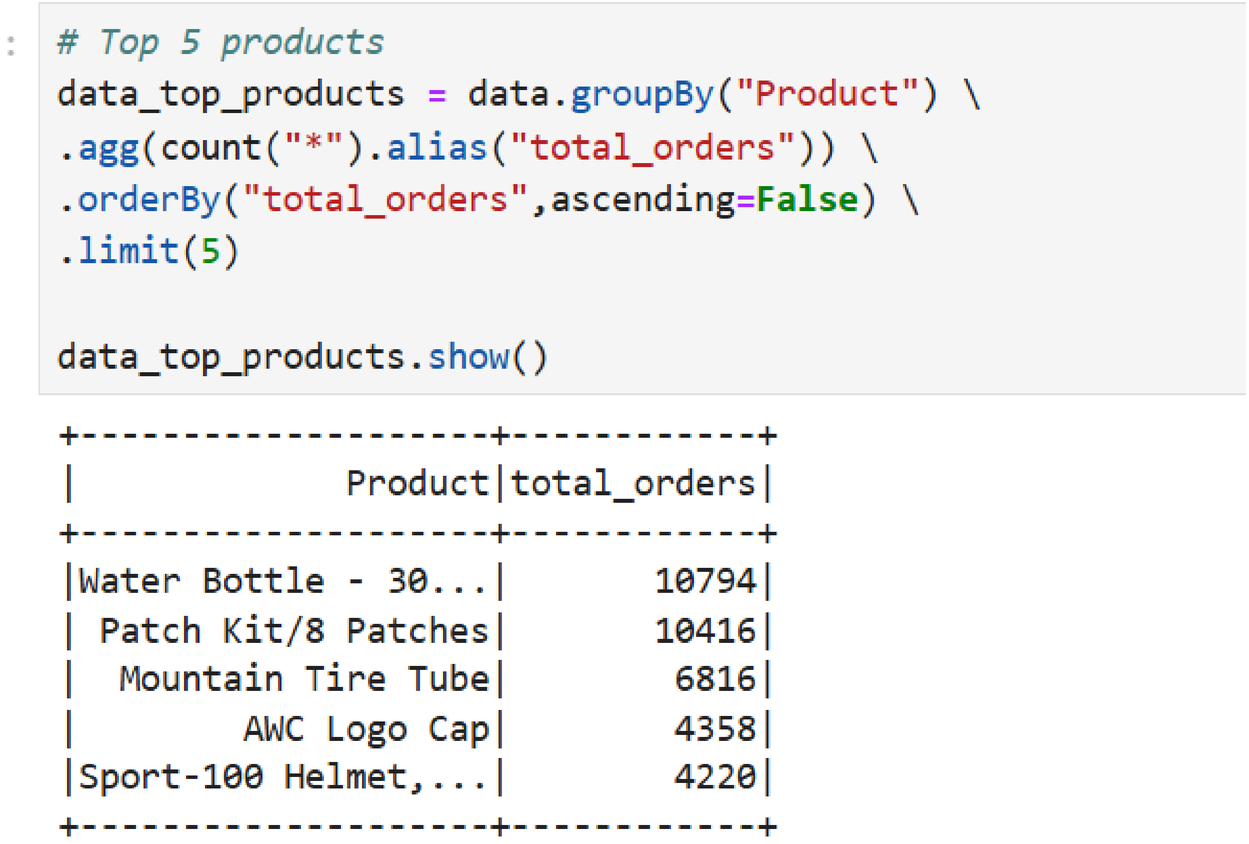
1. Pendapatan perbulan



```
from pyspark.sql import SparkSession
from pyspark.sql.functions import month, sum, count

spark= SparkSession.builder.appName("ETLPipeline").getOrCreate()

data = spark.read.csv("sales_data.csv", header=True, inferSchema=True)

# Pendapatan per bulan
data_revenue = data.withColumn("month",month("Date")) \
.groupBy("month") \
.agg(sum(data["Unit_Price"] * data["Order_Quantity"]).alias("total_revenue"))

data_revenue.show()
```

```
+-----+-------------+
|month|total_revenue|
+-----+-------------+
|   12|     10158080|
|    1|      7832338|
|    6|     10085537|
|    3|      8201790|
|    5|      9859851|
|    9|      6517880|
|    4|      8485163|
|    8|      6348349|
|    7|      6392045|
|   10|      6709394|
|   11|      6977157|
|    2|      7608734|
+-----+-------------+
```

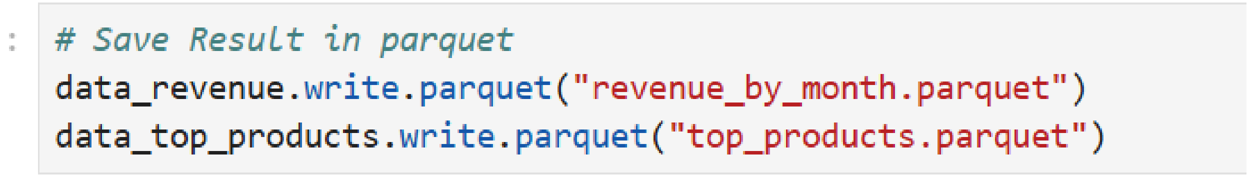
2. Identifikasi 5 Produk terlaris



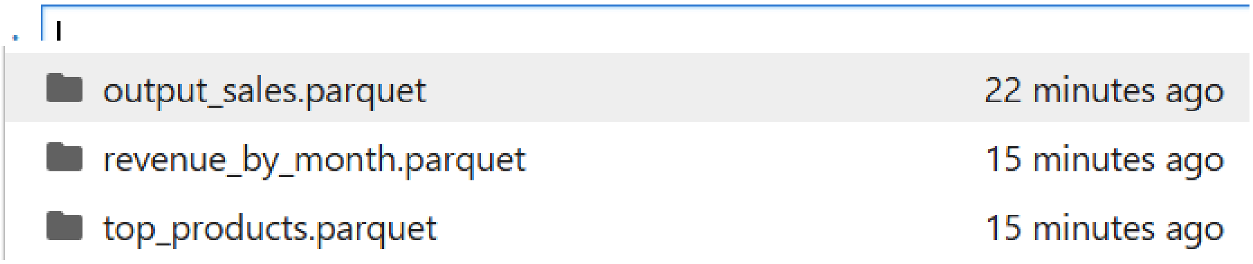
```python
# Top 5 products
data_top_products = data.groupBy("Product") \
.agg(count("*").alias("total_orders")) \
.orderBy("total_orders",ascending=False) \
.limit(5)

data_top_products.show()
```

```
+--------------------+------------+
|             Product|total_orders|
+--------------------+------------+
|Water Bottle - 30...|       10794|
| Patch Kit/8 Patches|       10416|
|  Mountain Tire Tube|        6816|
|        AWC Logo Cap|        4358|
|Sport-100 Helmet,...|        4220|
+--------------------+------------+
```

3. simpan dalam format parquet



```python
# Save Result in parquet
data_revenue.write.parquet("revenue_by_month.parquet")
data_top_products.write.parquet("top_products.parquet")
```



output_sales.parquet	22 minutes ago
revenue_by_month.parquet	15 minutes ago
top_products.parquet	15 minutes ago