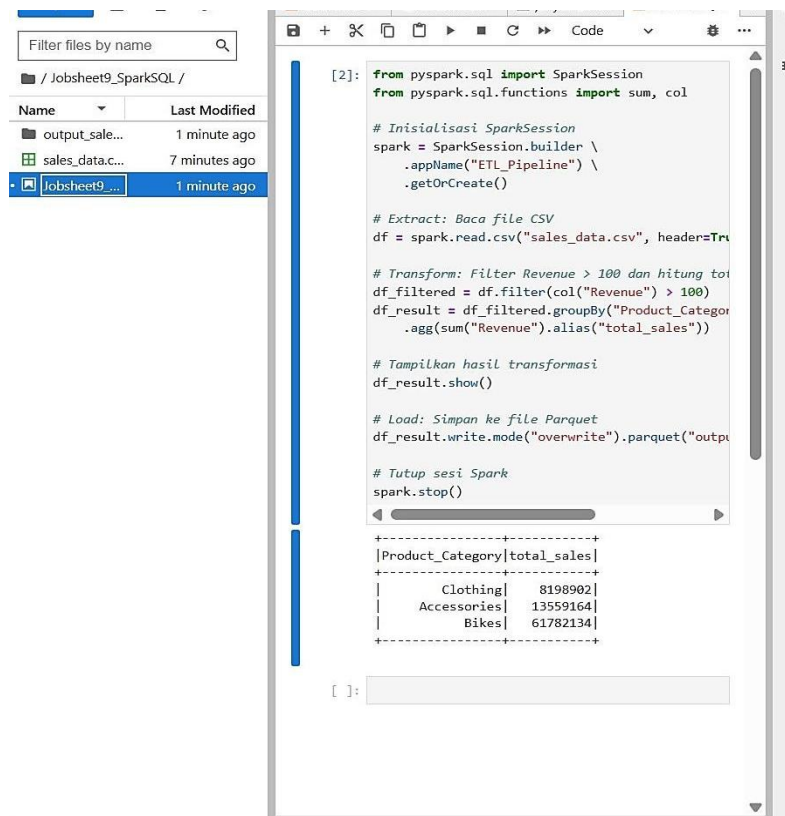


NAMA : HERTIN NURHAYATI
KELAS : SIB 3E
NIM : 2241760025

1. Praktikum: Membangun ETL Pipeline

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.



```
[2]: from pyspark.sql import SparkSession
from pyspark.sql.functions import sum, col

# Inisialisasi SparkSession
spark = SparkSession.builder \
    .appName("ETL_Pipeline") \
    .getOrCreate()

# Extract: Baca file CSV
df = spark.read.csv("sales_data.csv", headers=True)

# Transform: Filter Revenue > 100 dan hitung total penjualan per kategori
df_filtered = df.filter(col("Revenue") > 100)
df_result = df_filtered.groupBy("Product_Category") \
    .agg(sum("Revenue").alias("total_sales"))

# Tampilkan hasil transformasi
df_result.show()

# Load: Simpan ke file Parquet
df_result.write.mode("overwrite").parquet("output_sales.parquet")

# Tutup sesi Spark
spark.stop()
```

Product_Category	total_sales
Clothing	8198902
Accessories	13559164
Bikes	61782134

2. Analisis Data Retail

Dataset

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

Tugas

1. Hitung total pendapatan per bulan.

Filter files by name

/ Jobsheet9_SparkSQL /

Name	Last Modified
output_sale...	12 minutes ago
sales_data.c...	18 minutes ago
Jobsheet9_...	37 seconds ago

```
[7]: from pyspark.sql import SparkSession
from pyspark.sql.functions import month, sum, co

# 1. Start ulang SparkSession
spark = SparkSession.builder \
    .appName("Sales Analysis") \
    .getOrCreate()

# 2. Load file CSV
df = spark.read.csv("sales_data.csv", header=True)

# 3. Hitung pendapatan per bulan
df_revenue = df.withColumn("month", month("Date")) \
    .groupBy("month") \
    .agg(sum(df["Unit_Price"] * df["Order_Quantit

df_revenue.show()

# 4. Produk terlaris
df_top_products = df.groupBy("Product") \
    .agg(count("*").alias("total_orders")) \
    .orderBy("total_orders", ascending=False) \
    .limit(5)

df_top_products.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.

Filter files by name

/ Jobsheet9_SparkSQL /

Name	Last Modified
output_sale...	13 minutes ago
sales_data.c...	19 minutes ago
Jobsheet9_...	1 minute ago

```
df_revenue.show()

# 4. Produk terlaris
df_top_products = df.groupBy("Product") \
    .agg(count("*").alias("total_orders")) \
    .orderBy("total_orders", ascending=False) \
    .limit(5)

df_top_products.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

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 hasil dalam format Parquet.

