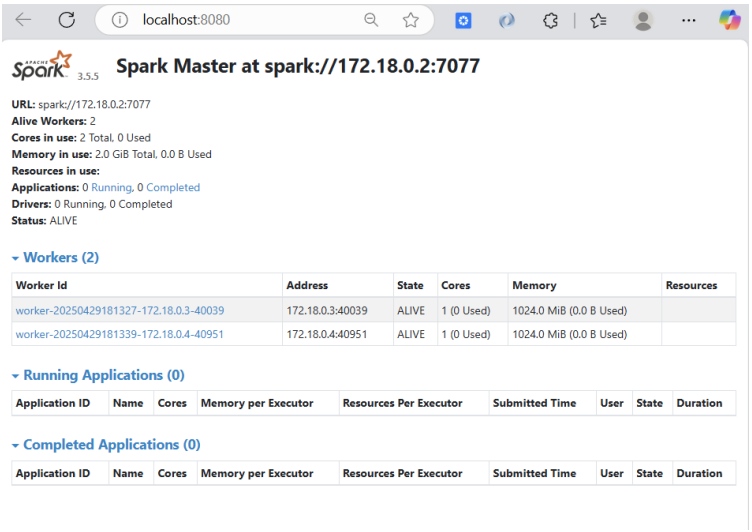
NAMA : SUKMA BAGUS WAHASDWIKA
NIM : 2241720223
KELAS : TI - 3D

BIG DATA – 09 Spark SQL

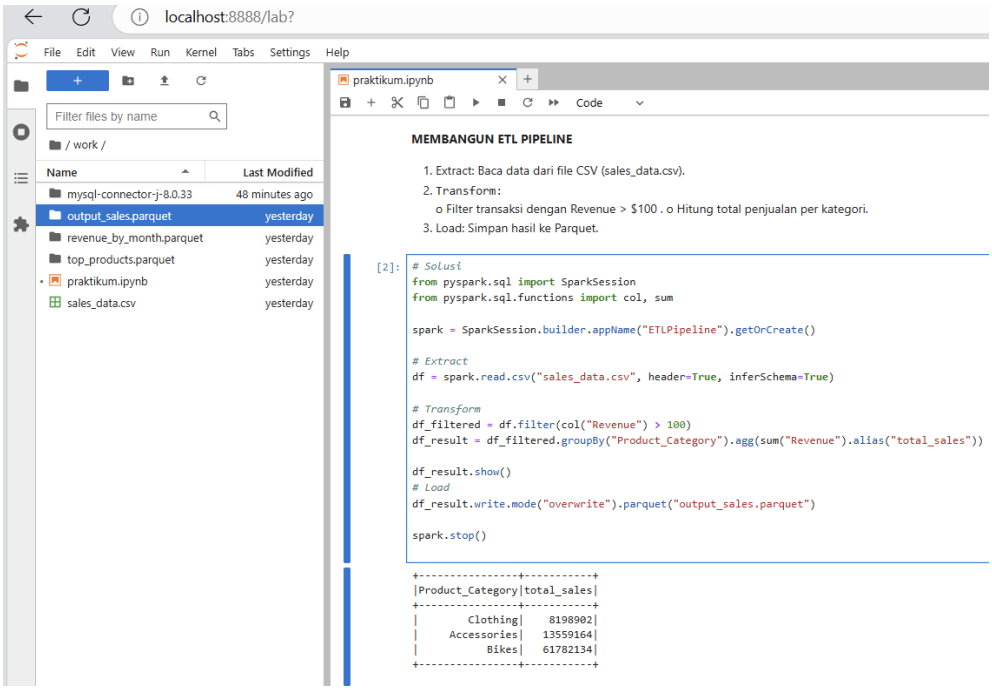
Siapkan lingkungan Spark Cluster



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. (Hasil Praktikum di bawah ini)



NAMA : SUKMA BAGUS WAHASDWIKA
NIM : 2241720223
KELAS : TI - 3D

BIG DATA – 09 Spark SQL

Analisis Data Retail

Dataset

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

Tugas

1. Hitung total pendapatan per bulan.

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.

[3]:
```
# Solusi 1 : Pendapatan perbulan
from pyspark.sql import SparkSession
from pyspark.sql.functions import month, sum, count

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

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

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

df_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.

[4]:
```
# Solusi 2 : 5 produk terlaris
df_top_products = df.groupBy("Product") \
                    .agg(count("*").alias("total_orders")) \
                    .orderBy("total_orders", ascending=False) \
                    .limit(5)

df_top_products.show()
```

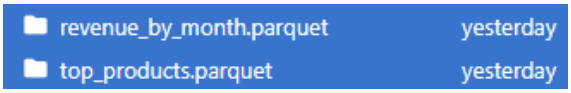
```
+--------------------+------------+
|             Product|total_orders|
+--------------------+------------+
|Water Bottle - 30...|       10794|
| Patch Kit/8 Patches|       10416|
|  Mountain Tire Tube|        6816|
|        AWC Logo Cap|        4358|
|Sport-100 Helmet,...|        4220|
+--------------------+------------+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

NAMA : SUKMA BAGUS WAHASDWIKA
NIM : 2241720223
KELAS : TI - 3D

BIG DATA – 09 Spark SQL

3. Simpan hasil dalam format Parquet.

```
[5]: # Solusi 3 : Simpan hasil
df_revenue.write.parquet("revenue_by_month.parquet")
df_top_products.write.parquet("top_products.parquet")
```



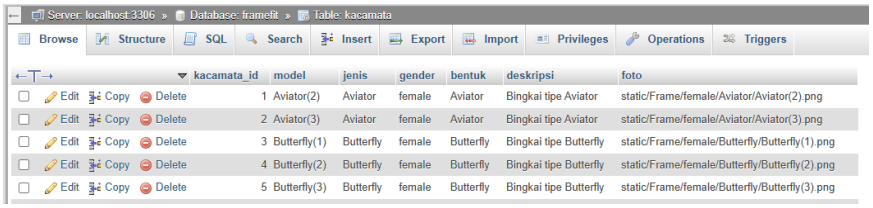
Evaluasi

Soal Latihan

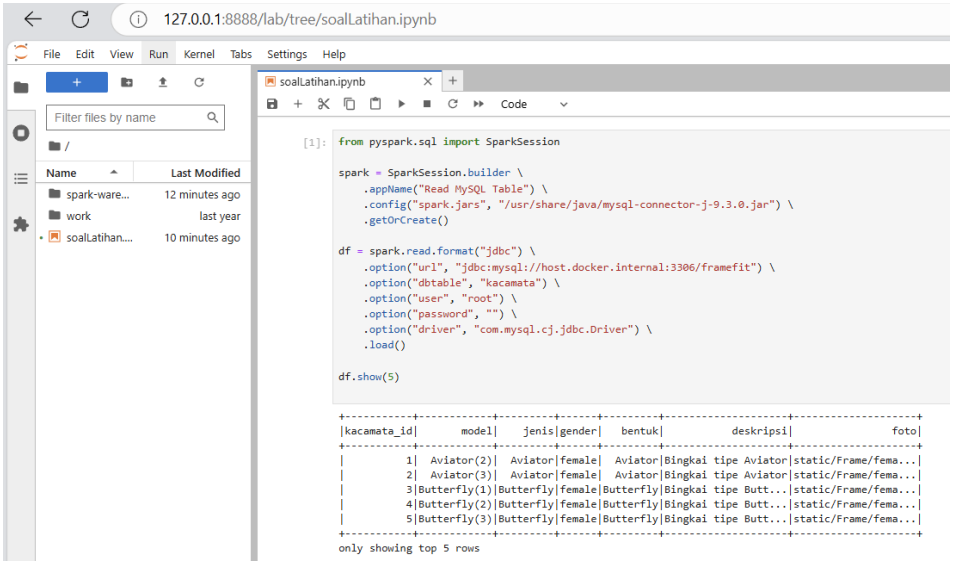
1. Baca data dari table di database MySQL anda menggunakan Spark,

   DB : framefit

   Tabel : kacamata



- Konfigurasi JDBC, download file mysql-connector-j-9.3.0.jar. Kemudian file mysql-connector-j-9.3.0.jar di copy ke spark-notebook dengan perintah docker cp mysql-connector-j-9.3.0.jar beautiful_bassi:/usr/share/java/.
- Dilanjutkan dengan menjalankan kode progam seperti dibawah



NAMA : SUKMA BAGUS WAHASDWIKA
NIM : 2241720223
KELAS : TI - 3D

BIG DATA – 09 Spark SQL

2. Buat query Spark SQL untuk menghitung Jumlah row dalam table tersebut

```
df.createOrReplaceTempView("kacamata")

jumlah_row = spark.sql("SELECT COUNT(*) AS jumlah_row FROM kacamata").show()
```

```
+----------+
|jumlah_row|
+----------+
|        62|
+----------+
```

Kesimpulan

- Spark SQL menyediakan antarmuka terstruktur untuk pemrosesan data besar.
- DataFrame & Dataset APIs memungkinkan manipulasi data dengan sintaks mirip SQL.
- DataSources API mendukung integrasi dengan berbagai format penyimpanan.