Raport

Lab. 4

Autor Maciej Tonderski

1. **Cel ćwieczenia**

Celem tego ćwiczenia jest przeprowadzenie analizy danych transakcyjnych detalicznych za pomocą Apache Spark SQL. Studenci będą pracować na zbiorze danych "Online Retail", który zawiera informacje o transakcjach detalicznych, takie jak identyfikator transakcji, kod produktu, ilość, cena jednostkowa, kraj itp. Zadanie będzie polegać na wczytaniu danych, wykonaniu różnych analiz, takich jak analiza sprzedaży, trendów produktowych, analiza krajów itp., oraz prezentacji wyników za pomocą odpowiednich zapytań SQL i wizualizacji.

Do ćwiczenia stworzono skrypt:

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._
import org.apache.spark.sql.expressions.Window

val spark = SparkSession.builder.appName("Online Retail
Analysis").getOrCreate()

println("Step 1: Reading the Data")
// Step 1: Reading the Data
val df = spark.read.option("header",
"true").option("inferSchema", "true").csv("/opt/spark-
data/OnlineRetail.csv")
println("Displaying the first 20 rows of the dataset:")
df.show()

println("Step 2: Data Cleaning")
// Step 2: Data Cleaning
println("Checking for missing values and dropping rows with
missing values")
val cleanedDF = df.na.drop()
println("Checking for duplicates and dropping duplicate rows")
val deduplicatedDF = cleanedDF.dropDuplicates()
println("Displaying the cleaned data:")
```

```
deduplicatedDF.show()

println("Step 3: Sales Analysis")
// Step 3: Sales Analysis
println("Calculating the total sales value for each
transaction")
val salesDF = deduplicatedDF.withColumn("TotalSale",
col("Quantity") * col("UnitPrice"))
val totalSalesDF =
salesDF.groupBy("InvoiceNo").agg(sum("TotalSale").alias("Total
Sale"))
println("Displaying the total sales per transaction, sorted by
total sales:")
totalSalesDF.orderBy(desc("TotalSale")).show()

println("Step 4: Product Trends Analysis")
// Step 4: Product Trends Analysis
println("Analyzing product sales trends over time")
val windowSpec =
Window.partitionBy("StockCode").orderBy("InvoiceDate")
val salesWithLagDF = salesDF.withColumn("PreviousTotalSale",
lag("TotalSale", 1).over(windowSpec))
val salesTrendDF =
salesWithLagDF.withColumn("PercentageChange",
  (col("TotalSale") - col("PreviousTotalSale")) /
col("PreviousTotalSale") * 100)
println("Displaying the product sales trends with percentage
changes:")
salesTrendDF.show()

println("Step 5: Geographical Analysis")
// Step 5: Geographical Analysis
println("Calculating total sales by country")
val countrySalesDF =
salesDF.groupBy("Country").agg(sum("TotalSale").alias("TotalSa
le"))
println("Displaying the total sales per country, sorted by
total sales:")
```

```
countrySalesDF.orderBy(desc("TotalSale")).show()

// Save results to CSV
println("Saving results to CSV files")
totalSalesDF.write.option("header", "true").csv("/opt/spark-
data/output/total_sales")
salesTrendDF.write.option("header", "true").csv("/opt/spark-
data/output/sales_trend")
countrySalesDF.write.option("header", "true").csv("/opt/spark-
data/output/country_sales")

println("Analysis complete. Results saved to /opt/spark-
data/output/")
spark.stop()
```

Następnie w celu uruchomienia kodu stworzono dockerfile budujący obraz wraz z potrzebnymi bibliotekami:

```
# Use an official Spark base image
FROM bitnami/spark:latest

# Copy the Spark application code into the Docker image
COPY lab3.scala /opt/bitnami/spark/work-dir/

# Set the working directory
WORKDIR /opt/bitnami/spark/work-dir/

# Run the Spark Shell with the application code
CMD ["spark-shell", "-i", "lab3.scala"]
```

A na koniec do uruchomienia kodu stworzono docker-compose-yml file which creates master-slave spark system to execute code with attached volumes:

```
version: "3.7"
services:
  spark-master:
    image: bitnami/spark:latest
    container_name: spark-master
    environment:
```

```
      - SPARK_MODE=master
    ports:
      - "8080:8080"
      - "7077:7077"

  spark-worker:
    image: bitnami/spark:latest
    container_name: spark-worker
    environment:
      - SPARK_MODE=worker
      - SPARK_MASTER_URL=spark://spark-master:7077
    depends_on:
      - spark-master
    ports:
      - "8081:8081"

  spark-driver:
    build: .
    container_name: spark-driver
    environment:
      - SPARK_MASTER_URL=spark://spark-master:7077
    depends_on:
      - spark-master
      - spark-worker
    volumes:
      - ./data:/opt/spark-data
```

To run it type in a console:

```
docker compose up --build
```

This action will build driver container and create rest of the needed containers alongside with respected networks.

2. **Wykonanie**

Zrealizowany skrypt wykonał kod którego poszególne kroki zostały opisane w załączeniu wklejony wynik z konsoli po uruchomieniu programu:

```
Step 1: Reading the Data
Displaying the first 20 rows of the dataset:
+---------+---------+--------------------+--------+--------------+---------+----------+--------------+
|InvoiceNo|StockCode|         Description|Quantity|   InvoiceDate|UnitPrice|CustomerID|       Country|
+---------+---------+--------------------+--------+--------------+---------+----------+--------------+
|   536365|   85123A|WHITE HANGING HEA...|       6|12/1/2010 8:26|     2.55|     17850|United Kingdom|
|   536365|    71053| WHITE METAL LANTERN|       6|12/1/2010 8:26|     3.39|     17850|United Kingdom|
|   536365|   84406B|CREAM CUPID HEART...|       8|12/1/2010 8:26|     2.75|     17850|United Kingdom|
|   536365|   84029G|KNITTED UNION FLA...|       6|12/1/2010 8:26|     3.39|     17850|United Kingdom|
|   536365|   84029E|RED WOOLLY HOTTIE...|       6|12/1/2010 8:26|     3.39|     17850|United Kingdom|
|   536365|    22752|SET 7 BABUSHKA NE...|       2|12/1/2010 8:26|     7.65|     17850|United Kingdom|
|   536365|    21730|GLASS STAR FROSTE...|       6|12/1/2010 8:26|     4.25|     17850|United Kingdom|
|   536366|    22633|HAND WARMER UNION...|       6|12/1/2010 8:28|     1.85|     17850|United Kingdom|
|   536366|    22632|HAND WARMER RED P...|       6|12/1/2010 8:28|     1.85|     17850|United Kingdom|
|   536367|    84879|ASSORTED COLOUR B...|      32|12/1/2010 8:34|     1.69|     13047|United Kingdom|
|   536367|    22745|POPPY'S PLAYHOUSE...|       6|12/1/2010 8:34|      2.1|     13047|United Kingdom|
|   536367|    22748|POPPY'S PLAYHOUSE...|       6|12/1/2010 8:34|      2.1|     13047|United Kingdom|
|   536367|    22749|FELTCRAFT PRINCES...|       8|12/1/2010 8:34|     3.75|     13047|United Kingdom|
|   536367|    22310|IVORY KNITTED MUG...|       6|12/1/2010 8:34|     1.65|     13047|United Kingdom|
|   536367|    84969|BOX OF 6 ASSORTED...|       6|12/1/2010 8:34|     4.25|     13047|United Kingdom|
|   536367|    22623|BOX OF VINTAGE JI...|       3|12/1/2010 8:34|     4.95|     13047|United Kingdom|
|   536367|    22622|BOX OF VINTAGE AL...|       2|12/1/2010 8:34|     9.95|     13047|United Kingdom|
|   536367|    21754|HOME BUILDING BLO...|       3|12/1/2010 8:34|     5.95|     13047|United Kingdom|
|   536367|    21755|LOVE BUILDING BLO...|       3|12/1/2010 8:34|     5.95|     13047|United Kingdom|
|   536367|    21777|RECIPE BOX WITH M...|       4|12/1/2010 8:34|     7.95|     13047|United Kingdom|
+---------+---------+--------------------+--------+--------------+---------+----------+--------------+
only showing top 20 rows
```

```
Step 2: Data Cleaning
Checking for missing values and dropping rows with missing values
Checking for duplicates and dropping duplicate rows
Displaying the cleaned data:
+---------+---------+--------------------+--------+--------------+---------+----------+--------------+
|InvoiceNo|StockCode|         Description|Quantity|    InvoiceDate|UnitPrice|CustomerID|       Country|
+---------+---------+--------------------+--------+--------------+---------+----------+--------------+
|   536367|    22745|POPPY'S PLAYHOUSE...|       6| 12/1/2010 8:34|      2.1|     13047|United Kingdom|
|   536368|    22960|JAM MAKING SET WI...|       6| 12/1/2010 8:34|     4.25|     13047|United Kingdom|
|   536388|    22915|ASSORTED BOTTLE T...|      12| 12/1/2010 9:59|     0.42|     16250|United Kingdom|
|   536401|    21464|DISCO BALL ROTATO...|       1|12/1/2010 11:21|     4.25|     15862|United Kingdom|
|   536412|    22569|FELTCRAFT CUSHION...|       2|12/1/2010 11:49|     3.75|     17920|United Kingdom|
|   536425|    22645|CERAMIC HEART FAI...|      12|12/1/2010 12:08|     1.45|     13758|United Kingdom|
|   536488|    22376|AIRLINE BAG VINTA...|       1|12/1/2010 12:31|     4.25|     17897|United Kingdom|
|   536520|    21930|JUMBO STORAGE BAG...|       1|12/1/2010 12:43|     1.95|     14729|United Kingdom|
|   536534|    22866|HAND WARMER SCOTT...|      12|12/1/2010 13:33|      2.1|     15350|United Kingdom|
|   536540|   85136A|YELLOW SHARK HELI...|       2|12/1/2010 14:05|     7.95|     14911|          EIRE|
|   536562|   79302M|ART LIGHTS,FUNK M...|       6|12/1/2010 15:08|     2.95|     13468|United Kingdom|
|   536569|    22941|CHRISTMAS LIGHTS ...|       1|12/1/2010 15:35|      8.5|     16274|United Kingdom|
|   536571|    21352|EUCALYPTUS & PINE...|       2|12/1/2010 15:37|     6.75|     14696|United Kingdom|
|   536591|    21985|PACK OF 12 HEARTS...|       4|12/1/2010 16:58|     0.29|     14606|United Kingdom|
|   536624|    22672|FRENCH BATHROOM S...|      12|12/2/2010 10:45|     1.65|     13418|United Kingdom|
|   536624|    21843|RED RETROSPOT CAK...|       4|12/2/2010 10:45|    10.95|     13418|United Kingdom|
|   536630|    22752|SET 7 BABUSHKA NE...|       2|12/2/2010 10:56|     7.65|     17850|United Kingdom|
|   536635|    22441|GROW YOUR OWN BAS...|       8|12/2/2010 11:22|      2.1|     15955|United Kingdom|
|   536667|    22594|CHRISTMAS GINGHAM...|      24|12/2/2010 12:09|     0.85|     15260|United Kingdom|
|   536671|    22740|        POLKADOT PEN|      48|12/2/2010 12:10|     0.85|     13305|United Kingdom|
+---------+---------+--------------------+--------+--------------+---------+----------+--------------+
only showing top 20 rows
```

```
Step 3: Sales Analysis
Calculating the total sales value for each transaction
Displaying the total sales per transaction, sorted by total sales:
+---------+------------------+
|InvoiceNo|         TotalSale|
+---------+------------------+
|   581483|          168469.6|
|   541431|           77183.6|
|   556444|           38970.0|
|   567423|          31698.16|
|   556917|          22775.93|
|   572209|           22206.0|
|   567381|22104.800000000003|
|   563614|          21880.44|
|   550461|           21535.9|
|   572035|          20277.92|
|   563076|19150.660000000003|
|   562439|18841.480000000003|
|   541220|16774.719999999998|
|   545475|          16726.84|
|   556255|           16488.0|
|   537659|          15885.49|
|   548011|          15719.56|
|   569650|          15643.77|
|   540815|15160.900000000001|
|   552883|14415.740000000002|
+---------+------------------+
only showing top 20 rows
```

```
Step 4: Product Trends Analysis
Analyzing product sales trends over time
Displaying the product sales trends with percentage changes:
+---------+---------+--------------------+--------+---------------+---------+----------+--------------+---------+-----------------+--------------------+
|InvoiceNo|StockCode|         Description|Quantity|    InvoiceDate|UnitPrice|CustomerID|       Country|TotalSale|PreviousTotalSale|  PercentageChange|
+---------+---------+--------------------+--------+---------------+---------+----------+--------------+---------+-----------------+--------------------+
|   540595|    10133|COLOURING PENCILS...|      20| 1/10/2011 11:35|     0.42|     14321|United Kingdom|      8.4|             NULL|                NULL|
|   541594|    10133|COLOURING PENCILS...|      20| 1/19/2011 15:30|     0.42|     14916|United Kingdom|      8.4|              8.4|                 0.0|
|   541699|    10133|COLOURING PENCILS...|      40|  1/21/2011 9:03|     0.42|     17744|United Kingdom|     16.8|              8.4|               100.0|
|   542103|    10133|COLOURING PENCILS...|      10| 1/25/2011 13:26|     0.85|     13198|United Kingdom|      8.5|             16.8|-49.404761904761905|
|   536446|    10133|COLOURING PENCILS...|       5| 12/1/2010 12:15|     0.85|     15983|United Kingdom|     4.25|              8.5|               -50.0|
|   539222|    10133|COLOURING PENCILS...|      10|12/16/2010 13:00|     0.85|     14621|United Kingdom|      8.5|             4.25|               100.0|
|   539240|    10133|COLOURING PENCILS...|      20|12/16/2010 13:15|     0.42|     15194|United Kingdom|      8.4|              8.5|  -1.17647058823529|
|   539252|    10133|COLOURING PENCILS...|      20|12/16/2010 14:17|     0.42|     17744|United Kingdom|      8.4|              8.4|                 0.0|
|   539353|    10133|COLOURING PENCILS...|      20|12/17/2010 11:30|     0.42|     12782|      Portugal|      8.4|              8.4|                 0.0|
|   539477|    10133|COLOURING PENCILS...|      10|12/19/2010 14:58|     0.85|     18245|United Kingdom|      8.5|              8.4| 1.1904761904761862|
|   539933|    10133|COLOURING PENCILS...|      10|12/23/2010 11:24|     0.85|     15235|United Kingdom|      8.5|              8.5|                 0.0|
|   537126|    10133|COLOURING PENCILS...|       1| 12/5/2010 12:13|     0.85|     18118|United Kingdom|     0.85|              8.5|               -90.0|
|   537155|    10133|COLOURING PENCILS...|       3| 12/5/2010 13:05|     0.85|     12748|United Kingdom|     2.55|             0.85| 199.99999999999997|
|   537225|    10133|COLOURING PENCILS...|      10| 12/5/2010 16:41|     0.85|     12748|United Kingdom|      8.5|             2.55| 233.33333333333334|
|   537374|    10133|COLOURING PENCILS...|       2| 12/6/2010 12:55|     0.85|     17259|United Kingdom|      1.7|              8.5|               -80.0|
|   538064|    10133|COLOURING PENCILS...|       2| 12/9/2010 13:47|     0.85|     15271|United Kingdom|      1.7|              1.7|                 0.0|
|   538070|    10133|COLOURING PENCILS...|      10| 12/9/2010 14:08|     0.85|     16519|United Kingdom|      8.5|              1.7|               400.0|
|   543982|    10133|COLOURING PENCILS...|      10|  2/15/2011 8:21|     0.85|     15358|United Kingdom|      8.5|              8.5|                 0.0|
|   544586|    10133|COLOURING PENCILS...|       4| 2/21/2011 15:04|     0.85|     17338|United Kingdom|      3.4|              8.5|               -60.0|
|   544931|    10133|COLOURING PENCILS...|      10| 2/24/2011 18:59|     0.85|     13501|   Switzerland|      8.5|              3.4|               150.0|
+---------+---------+--------------------+--------+---------------+---------+----------+--------------+---------+-----------------+--------------------+
only showing top 20 rows
```

```
Step 5: Geographical Analysis
Calculating total sales by country
Displaying the total sales per country, sorted by total sales:
+--------------+------------------+
|       Country|         TotalSale|
+--------------+------------------+
| United Kingdom| 6747156.154000061|
|    Netherlands|284661.54000000004|
|           EIRE|250001.78000000014|
|        Germany| 221509.4699999998|
|         France|196626.04999999987|
|      Australia|         137009.77|
|    Switzerland|55739.399999999994|
|          Spain| 54756.02999999998|
|        Belgium| 40910.96000000001|
|         Sweden| 36585.41000000001|
|          Japan|          35340.62|
|         Norway|          35163.46|
|       Portugal|28995.760000000013|
|         Finland|22326.739999999998|
|Channel Islands|          20076.39|
|        Denmark|18768.140000000003|
|          Italy|16890.510000000002|
|         Cyprus|12858.759999999998|
|        Austria|          10154.32|
|      Singapore|           9120.39|
+--------------+------------------+
only showing top 20 rows
```

3. **Wnioski:**

Przeprowadzone ćwiczenie pokazało, jak efektywnie można wykorzystać Apache Spark SQL do analizy dużych zbiorów danych transakcyjnych. Dzięki zastosowaniu technik czyszczenia danych, grupowania i agregacji, udało się uzyskać wartościowe informacje, które mogą być użyteczne w podejmowaniu decyzji biznesowych. Analiza sprzedaży, trendów produktowych oraz analiza geograficzna dostarczyły istotnych wniosków, które mogą pomóc w optymalizacji strategii sprzedażowej i marketingowej. Dodatkowo, zastosowanie środowiska Docker pozwoliło na łatwe uruchomienie i skalowanie aplikacji w różnych środowiskach, co zwiększa elastyczność i efektywność przeprowadzanych analiz.