Bytewise Fellowship

Task 10: Reading a CSV using PySpark

Name	Muhammad Ahsan Saleem
Date	5th August, 2024
Submitted to:	Muhammad Bilal

Task 10 - Reading a CSV using PySpark

Task Statement: Learn how to read a CSV file using PySpark and display the resulting Spark DataFrame.

Requirements:

- Use PySpark instead of Pandas to read the CSV file provided earlier.
- Display the first few rows of the Spark DataFrame to verify successful extraction.

Steps and Code:

In this section, I will outline the steps taken to complete the task of reading a CSV using PySpark. This includes initializing the Spark session, reading the CSV file into a Spark DataFrame, and displaying the first few rows to verify successful data extraction. Each step is accompanied by the corresponding Python code.

Step 1: Import SparkSession:

- SparkSession is the entry point to programming with PySpark.

Code:

```
from pyspark.sql import SparkSession
```

Step 2: Initialize Spark session:

- This initializes a Spark session with the application name "SalesDataProcessing". getOrCreate() ensures that a new session is created if it doesn't already exist.

Code:

```
# Initialize the SparkSession
spark = SparkSession.builder.appName("SalesDataProcessing").getOrCreate()
```

Step 3: Read the CSV file:

- file_path is the location of the CSV file.
- spark.read.csv() reads the CSV file into a DataFrame.
- header=True specifies that the first row of the CSV file contains headers.
- inferSchema=True enables Spark to automatically infer the data types of each column.

Code:

```
# Read the CSV file
file_path = r'C:\Users\mahsa\Desktop\Data\Bytewise Fellowship\Daily Tasks\Month 1\Task 10\Resources\dataset.csv'
sales_data = spark.read.csv(file_path, header=True, inferSchema=True)
```

Step 4: Display the first few rows of the DataFrame:

- This displays the first 5 rows of the DataFrame to verify that the data has been read correctly.

Code:

```
# Display the first few rows
sales_data.show(5)
```

Output:

```
+--------------------+--------+----------+-------+---------+-------------------+--------------------+--------------+--------------------+---------------+--------------------+
|             item_id|order_id|product_id| amount|   status|     item_timestamp|            location| customer_name|      customer_phone|        country|         description|
+--------------------+--------+----------+-------+---------+-------------------+--------------------+--------------+--------------------+---------------+--------------------+
|4dc01ae9-c1a8-461...|  160794|       647|2237.23|Cancelled|2024-01-13 21:34:...|       East Cameron|Richard Stevens|(774)709-6342x106|      Guatemala|Room as address h...|
|cafaa69b-f0c5-42c...|  105101|       127|2029.17|     NULL|2024-04-24 03:22:...|   East Richardville|     Keith Lamb|924-443-4084x8236|Saint Barthelemy|Nice beat despite...|
|77944e0c-f500-456...|  510841|       243| 848.88|Cancelled|2024-05-29 17:05:...|South Christinaburgh|  Patrick Allen| 001-734-642-3018|     Mauritania|Accept part crime...|
|1019711d-53c9-401...|  259964|       209| 614.64| Returned|2024-01-03 02:18:...|    South Jeremybury|    Wendy White|  +1-210-390-0363|       Cameroon|Top huge old beha...|
|8e4497f4-78f2-495...|  270130|       637| 2898.0|     NULL|2024-05-14 21:24:...|         Duncanland|   Dustin Hicks|     230.673.9935|       Maldives|Style there TV so...|
+--------------------+--------+----------+-------+---------+-------------------+--------------------+--------------+--------------------+---------------+--------------------+
only showing top 5 rows
```

Conclusion:

In this task, we successfully read a CSV file using PySpark and displayed the first few rows of the resulting DataFrame. This basic operation is crucial for understanding how to load and inspect data using PySpark, setting the stage for more advanced data processing and analysis tasks.