Assignment – Day 13

-Sarthak Niranjan Kulkarni (Maverick)

- sarthakkul2311@gmail.com - (+91) 93256 02791

20/11/2024 (Wednesday)

Summary of Modifying DataFrames in PySpark:-

1. **Spark Session Creation:**

- A SparkSession is created using .builder.appName('pyspark - example join').getOrCreate(), allowing the execution of PySpark commands.

2. **DataFrame Creation:**

- A list of tuples (data) is defined with sample information (names, dates of birth, gender, and salary), and this data is loaded into a DataFrame (df) with specified column names ("Name", "DOB", "Gender", "salary").

3. **Column Renaming:**

- The column "DOB" is renamed to "date of birth".
- The column "Name" is renamed to "personname", and the updated DataFrame is shown.

4. **Selecting and Renaming Columns Using selectExpr:**

- "Gender" is renamed to "category".
- "Name" is renamed to "name".
- The resulting DataFrame data is displayed.

5. **Using col() for Column Selection and Aliasing:**

- The select() function with col() is used to select columns explicitly and rename the "salary" column to "Amount" using the alias() function.
- The DataFrame with the renamed column (Amount) is displayed.

Modifying DataFrames in PySpark Practice: -

1. **Renaming columns in a PySpark DataFrame using withColumnRenamed.**

```
# Importing necessary libraries
from pyspark.sql import SparkSession

# Create a spark session
spark = SparkSession.builder.appName('pyspark - example
join').getOrCreate()

# Create data in dataframe
data = [(('SriRam'), '1991-04-01', 'M', 30000),
        (('Sarthak'), '2002-01-23', 'M', 4000),
        (('Rohini'), '1978-09-05', 'M', 4000),
        (('Lakshita'), '2002-08-08', 'F', 4000),
        (('Jenis'), '1980-02-17', 'F', 1200)]

# Column names in dataframe
columns = ["Name", "DOB", "Gender", "salary"]

# Create the spark dataframe
df = spark.createDataFrame(data=data,
                           schema=columns)
df.withColumnRenamed("DOB","date of birth").show()
df.withColumnRenamed("DOB","date of
birth").withColumnRenamed("Name","personname").show()
```

```
▸ (6) Spark Jobs
▸ df: pyspark.sql.dataframe.DataFrame = [Name: string, DOB: string ... 2 more fields]
+--------+-------------+------+------+
|    Name|date of birth|Gender|salary|
+--------+-------------+------+------+
|  SriRam|   1991-04-01|     M| 30000|
| Sarthak|   2002-01-23|     M|  4000|
|  Rohini|   1978-09-05|     M|  4000|
|Lakshita|   2002-08-08|     F|  4000|
|   Jenis|   1980-02-17|     F|  1200|
+--------+-------------+------+------+

+----------+-------------+------+------+
|personname|date of birth|Gender|salary|
+----------+-------------+------+------+
|    SriRam|   1991-04-01|     M| 30000|
|   Sarthak|   2002-01-23|     M|  4000|
|    Rohini|   1978-09-05|     M|  4000|
|  Lakshita|   2002-08-08|     F|  4000|
|     Jenis|   1980-02-17|     F|  1200|
+----------+-------------+------+------+
```

2. Selecting and renaming columns in a PySpark DataFrame using selectExpr.

```
# Importing necessary libraries using select exp
from pyspark.sql import SparkSession

# Create a spark session
spark = SparkSession.builder.appName('pyspark - example
join').getOrCreate()

# Create data in dataframe
data = [(('SriRam'), '1991-04-01', 'M', 30000),
        (('Sarthak'), '2002-01-23', 'M', 4000),
        (('Rohini'), '1978-09-05', 'M', 4000),
        (('Lakshita'), '2002-08-08', 'F', 4000),
        (('Jenis'), '1980-02-17', 'F', 1200)]

# Column names in dataframe
columns = ["Name", "DOB", "Gender", "salary"]

# Create the spark dataframe
df = spark.createDataFrame(data=data,
                           schema=columns)
data = df.selectExpr("Gender as category","DOB","Name as
name","salary")
data.show()
```

```
▸ (3) Spark Jobs
▸ df: pyspark.sql.dataframe.DataFrame = [Name: string, DOB: string ... 2 more fields]
▸ data: pyspark.sql.dataframe.DataFrame = [category: string, DOB: string ... 2 more fields]
+--------+----------+--------+------+
|category|       DOB|    name|salary|
+--------+----------+--------+------+
|       M|1991-04-01|  SriRam| 30000|
|       M|2002-01-23| Sarthak|  4000|
|       M|1978-09-05|  Rohini|  4000|
|       F|2002-08-08|Lakshita|  4000|
|       F|1980-02-17|   Jenis|  1200|
+--------+----------+--------+------+
```

3. Selecting and aliasing columns in a PySpark DataFrame using col() and alias().

```
from pyspark.sql.functions import col

# Select the 'salary' as 'Amount' using aliasing
# Select remaining with their original name
data = df.select(col("Name"),col("DOB"),
                 col("Gender"),
                 col("salary").alias('Amount'))
data.show()
```

```
▸ (3) Spark Jobs
▸ data: pyspark.sql.dataframe.DataFrame = [Name: string, DOB: string ... 2 more fields]
+--------+----------+------+------+
|    Name|       DOB|Gender|Amount|
+--------+----------+------+------+
|  SriRam|1991-04-01|     M| 30000|
| Sarthak|2002-01-23|     M|  4000|
|  Rohini|1978-09-05|     M|  4000|
|Lakshita|2002-08-08|     F|  4000|
|   Jenis|1980-02-17|     F|  1200|
+--------+----------+------+------+
```