Master PySpark: From Zero to Big Data Hero!!

union and unionAll in PySpark

Overview

- **Purpose**: Both union and unionAll are used to combine two DataFrames into a single DataFrame.
- **DataFrame Compatibility**: The two DataFrames must have the same schema (i.e., the same column names and data types) to perform the union operation.

union()

- **Functionality**:
  - Combines two DataFrames and retains all rows, duplicate rows from the result.
- **Behavior**:
  - The union() method doesnot retains unique rows across both DataFrames, resulting in a DataFrame with duplicates.

unionAll()

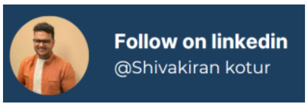
- **Functionality**:
  - Combines two DataFrames and retains all rows, including duplicates.
- **Behavior**:
  - The unionAll() method performs the union operation but does not eliminate duplicate rows, similar to Unionall

Syntax

```
# Using union to retain all rows including duplicates
unioned_df = df1.union(df2)

# Using unionAll to retain all rows including duplicates
unionAll_df = df1.unionAll(df2)
```

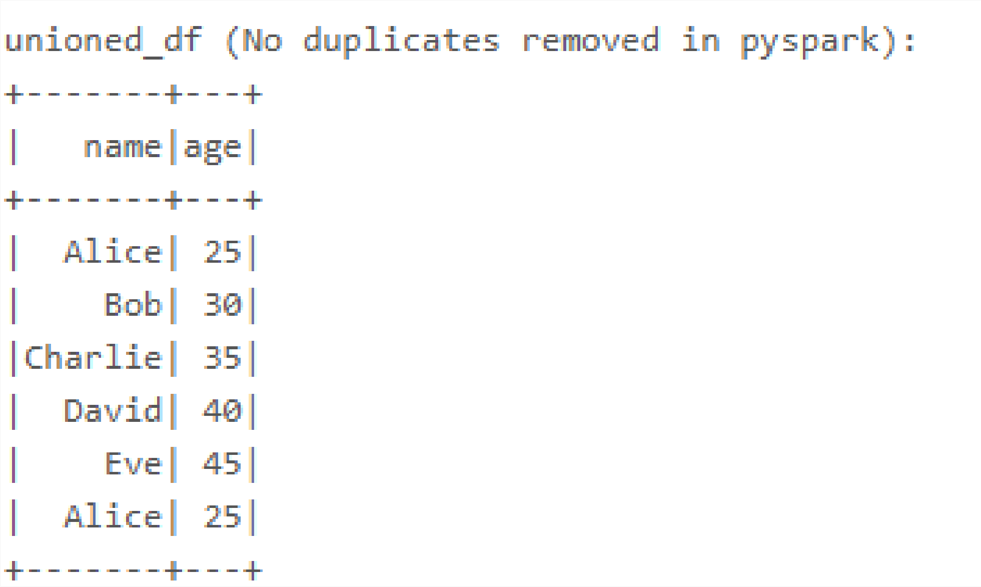
Example Code

Here’s a complete example demonstrating both union and unionAll:



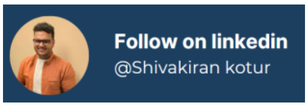
Follow me on LinkedIn – Shivakiran kotur

```python
from pyspark.sql import SparkSession

# Initialize Spark session
spark = SparkSession.builder.appName("UnionExample").getOrCreate()

# Sample DataFrames
data1 = [("Alice", 25), ("Bob", 30), ("Charlie", 35)]
data2 = [("David", 40), ("Eve", 45), ("Alice", 25)]
columns = ["name", "age"]

df1 = spark.createDataFrame(data1, columns)
df2 = spark.createDataFrame(data2, columns)

# Using union to retain all rows including duplicates
unioned_df = df1.union(df2)

# Using unionAll to retain all rows
unionAll_df = df1.unionAll(df2)

# Show the results
print("unioned_df (No duplicates removed):")
unioned_df.show()
```

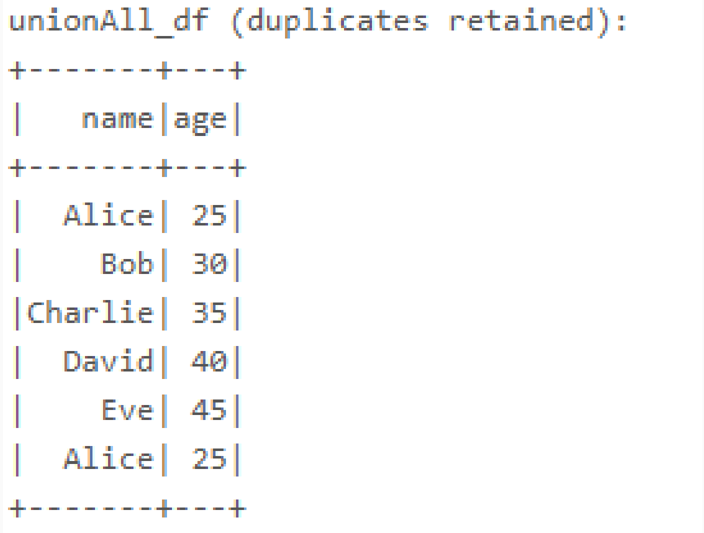


```
unioned_df (No duplicates removed in pyspark):
+-------+---+
|   name|age|
+-------+---+
|  Alice| 25|
|    Bob| 30|
|Charlie| 35|
|  David| 40|
|    Eve| 45|
|  Alice| 25|
+-------+---+
```

```python
print("unionAll_df (duplicates retained):")
unionAll_df.show()
```

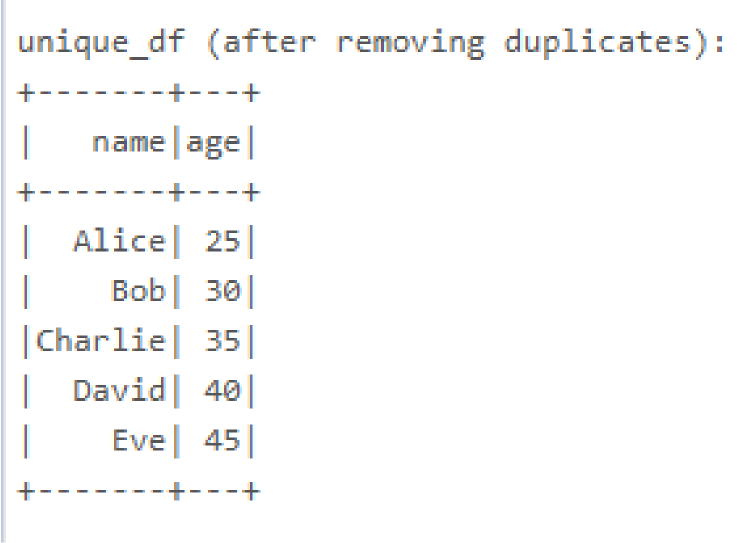


Follow me on LinkedIn – Shivakiran kotur



```
unionAll_df (duplicates retained):
+-------+---+
|   name|age|
+-------+---+
|  Alice| 25|
|    Bob| 30|
|Charlie| 35|
|  David| 40|
|    Eve| 45|
|  Alice| 25|
+-------+---+
```

```
# Remove duplicate rows and create a new DataFrame
unique_df = unioned_df.dropDuplicates()
# or
unique_df = unioned_df.distinct()

print("unique_df (after removing duplicates):")
unique_df.show()
```



```
unique_df (after removing duplicates):
+-------+---+
|   name|age|
+-------+---+
|  Alice| 25|
|    Bob| 30|
|Charlie| 35|
|  David| 40|
|    Eve| 45|
+-------+---+
```