PySpark

In [13]:
```
from pyspark.sql import SparkSession
```

In [23]:
```
# Initialize a Spark session
spark = SparkSession.builder.appName("RemoveDuplicates").getOrCreate()
```

In [24]:
```
# Create a DataFrame with duplicate rows
data = [("Alice", 25), ("Bob", 30), ("Alice", 25), ("Charlie", 35), ("
columns = ["Name", "Age"]
```

In [25]:
```
df = spark.createDataFrame(data, columns)
```

In [26]:
```
df.show()
```

```
+-------+---+
|   Name|Age|
+-------+---+
|  Alice| 25|
|    Bob| 30|
|  Alice| 25|
|Charlie| 35|
|  David| 40|
|    Bob| 30|
|    Bob|  3|
+-------+---+
```

In [27]:
```
# Remove duplicates and keep the first occurrence
df_no_duplicates = df.dropDuplicates()
```

In [28]:
```
# Display the DataFrame without duplicates
df_no_duplicates.show()
```

```
+-------+---+
|   Name|Age|
+-------+---+
|  Alice| 25|
|    Bob| 30|
|Charlie| 35|
|  David| 40|
|    Bob|  3|
+-------+---+
```

In [29]:
```
# Remove duplicates and based on one column
df_no_duplicates = df.dropDuplicates(subset=["Name"])

df_no_duplicates.show()
```

```
[Stage 15:=====================>                                    (
3 + 5) / 8]

+-------+---+
|   Name|Age|
+-------+---+
|  Alice| 25|
|    Bob| 30|
|Charlie| 35|
|  David| 40|
+-------+---+
```