Spark – Append or Concatenate two Datasets – Example

Append or Concatenate Datasets

Spark provides union() method in Dataset class to concatenate or append a Dataset to another. Dataset Union can only be performed on Datasets with the same number of columns.

Syntax of Dataset.union() method

```
public Dataset<Row>join(Dataset<?> right)
```

Returns Dataset with specified Dataset concatenated/appended to this Dataset.

Steps to Concatenate two Datasets

To append or concatenate two Datasets

1. Use Dataset.union() method on the first dataset and provide second Dataset as argument.

Example – Concatenate two Datasets

In the following example, we have two Datasets with employee information read from different data files. We shall concatenate these two Datasets.

ConcatenateDatasets.java

```
import
org.apache.spark.sql.D
```

```
import org.apache.spark.sql.Dataset;
import org.apache.spark.sql.Row;
import org.apache.spark.sql.SparkSession;

public class ConcatenateDatasets {

    public static void main(String[] args) {
        // configure spark
        SparkSession spark = SparkSession
                .builder()
                .appName("Spark Example - Append/Concatenate two Datasets")
                .master("local[2]")
                .getOrCreate();

        Dataset<Row> ds1 = spark.read().json("data/employees.json");
        Dataset<Row> ds2 = spark.read().json("data/employees2.json");

        // print dataset
        System.out.println("Dataset 1\n===============");
        ds1.show();
        System.out.println("Dataset 2\n===============");
        ds1.show();

        // concatenate datasets
        Dataset<Row> ds3 = ds1.union(ds2);

        System.out.println("Dataset 3 = Dataset 1 + Dataset 2\n==============================");
        ds3.show();

        spark.stop();
    }
}
```

Output

```
Dataset 1
===============
```

```
Dataset 1
==============
+-------+------+
| name|salary|
+-------+------+
|Michael| 3000|
| Andy| 4500|
| Justin| 3500|
| Berta| 4000|
| Raju| 3000|
+-------+------+

Dataset 2
==============
+-------+------+
| name|salary|
+-------+------+
|Michael| 3000|
| Andy| 4500|
| Justin| 3500|
| Berta| 4000|
| Raju| 3000|
+-------+------+

Dataset 3 = Dataset 1 + Dataset 2
==============================
+-------+------+
| name|salary|
+-------+------+
|Michael| 3000|
| Andy| 4500|
| Justin| 3500|
| Berta| 4000|
| Raju| 3000|
| Chandy| 4500|
| Joey| 3500|
| Mon| 4000|
| Rachel| 4000|
+-------+------+
```

General Pitfalls while concatenating Datasets

If number of columns in the two Datasets do not match, you would get an exception as shown below :

```
Exception in thread
"main"
```

```
Exception in thread "main" org.apache.spark.sql.AnalysisException: Union can only be performed on tables with the same number of columns, but the first table has 2 columns and the second table has 3 columns;;
'Union
:- Relation[name#8,salary#9L] json
+- Relation[name#21,nn#22L,salary#23L] json
```

In the above case, there are two columns in the first Dataset, while the second Dataset has three columns.

Conclusion :

In this Spark Tutorial – Concatenate two Datasets, we have learnt to use Dataset.union() method to **append a Dataset to another** with same number of columns.

- Spark RDD - Containing Custom Class Objects
- Spark RDD - Map
- Spark RDD - FlatMap
- Spark RDD - Filter
- Spark RDD - Distinct
- Spark RDD - Reduce

Spark Dataseet

- Spark - Read JSON file to Dataset
- Spark - Write Dataset to JSON file
- Spark - Add new Column to Dataset
- Spark - Concatenate Datasets

Spark MLlib (Machine Learning Library)

- Spark MLlib Tutorial
- KMeans Clustering & Classification
- Decision Tree Classification
- Random Forest Classification
- Naive Bayes Classification
- Logistic Regression Classification
- Topic Modelling

Spark SQL

- Spark SQL Tutorial
- Spark SQL - Load JSON file and execute SQL Query

Spark Others

- Spark Interview Questions