TutorialKart

Spark – Add new column to Dataset – Example

Spark – Add new column to Dataset

A new column could be added to an existing Dataset using Dataset.withColumn() method. withColumn accepts two arguments: the column name to be added, and the Column and returns a new Dataset<Row>. The syntax of withColumn() is provided below.

Syntax – withColumn()

The syntax of withColumn() method is

```
public Dataset<Row> withColumn(String colName, Column col)
```

Step by step process to add New Column to Dataset

To add a new column to Dataset in Apache Spark

1. Use withColumn() method of the Dataset.
2. Provide a string as first argument to withColumn() which represents the column name.
3. Useorg.apache.spark.sql.functions class for generating a new Column, to be provided as second argument. Spark functions[https://spark.apache.org/docs/latest/api/java/org/apache/spark/sql/functions.html] class provides methods for many of the mathematical functions like statistical, trigonometrical, etc.

Example – Spark – Add new column to Spark Dataset

In the following example, we shall add a new column with name “new_col” with a constant value. We shall use functions.lit(Object literal) to create a new Column.

DatasetAddColumn.java

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

```
public class DatasetAddColumn {

    public static void main(String[] args) {
        // configure spark
        SparkSession spark = SparkSession
                .builder()
                .appName("Spark Example - Add a new Column to Dataset")
                .master("local[2]")
                .getOrCreate();

        String jsonPath = "data/employees.json";
        Dataset<Row> ds = spark.read().json(jsonPath);

        // dataset before adding enw column
        ds.show();

        // add column to ds
        Dataset<Row> newDs = ds.withColumn("new_col",functions.lit(1));

        // print dataset after adding new column
        newDs.show();

        spark.stop();
    }
}
```

Output

```
+-------+------+
|   name|salary|
+-------+------+
|Michael|  3000|
|   Andy|  4500|
| Justin|  3500|
|  Berta|  4000|
|   Raju|  3000|
| Chandy|  4500|
|   Joey|  3500|
|    Mon|  4000|
| Rachel|  4000|
+-------+------+



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

Conclusion

In this Spark Tutorial – **Add new Column to existing DataSet**, we have learnt to use Dataset.withColumn() method and functions class to add a new column to a Dataset.

Learn Apache Spark

- Apache Spark Tutorial
- Install Spark on Ubuntu
- Install Spark on Mac OS
- Scala Spark Shell - Example
- Python Spark Shell - PySpark
- Setup Java Project with Spark
- Spark Scala Application - WordCount Example
- Spark Python Application
- Spark DAG & Physical Execution Plan
- Setup Spark Cluster
- Configure Spark Ecosystem
- Configure Spark Application
- Spark Cluster Managers

Spark RDD

- Spark RDD
- Spark RDD - Print Contents of RDD
- Spark RDD - foreach
- Spark RDD - Create RDD
- Spark Parallelize
- Spark RDD - Read Text File to RDD
- Spark RDD - Read Multiple Text Files to Single RDD
- Spark RDD - Read JSON File to RDD
- 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