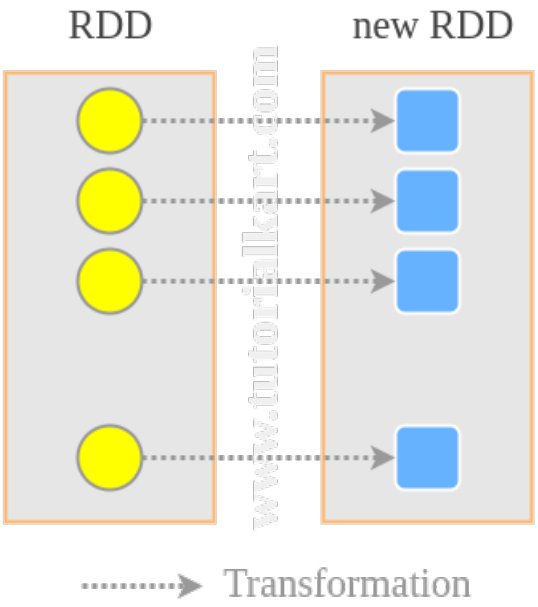
Spark RDD map() – Java & Python Examples

Spark RDD map()

In this Spark Tutorial, we shall learn to map one RDD to another. Mapping is transforming each RDD element using a function and returning a new RDD. Simple example would be calculating logarithmic value of each RDD element (RDD<Integer>) and creating a new RDD with the returned elements.



- Syntax
- Java Examples
- Python Examples

Syntax

```
RDD.map(<function>)
```

where<function> is the transformation function for each of the element of source RDD.

Examples

Java Example 1 – Spark RDD Map Example

In this example, we will an RDD with some integers. We shall then call map() function on this RDD to map

integer items to their logarithmic values The item in RDD is of type Integer, and the output for each item would be Double. So we are mapping an RDD<Integer> to RDD<Double>.

RDDmapExample2.java

```
import java.util.Arrays;

import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;

public class RDDmapExample2 {

    public static void main(String[] args) {
        // configure spark
        SparkConf sparkConf = new SparkConf().setAppName("Read Text to RDD")
                                        .setMaster("local[2]").set("spark.executor.memor
        // start a spark context
        JavaSparkContext sc = new JavaSparkContext(sparkConf);

        // initialize an integer RDD
        JavaRDD<Integer> numbers = sc.parallelize(Arrays.asList(14,21,88,99,455));

        // map each line to number of words in the line
        JavaRDD<Double> log_values = numbers.map(x -> Math.log(x));

        // collect RDD for printing
        for(double value:log_values.collect()){
            System.out.println(value);
        }
    }
}
```

Run this Spark Application and you would get the following output in the console.

```
17/11/28 16:31:11 INFO DAGScheduler: ResultStage 0 (collect at RDDmapExample2.java:23) f
17/11/28 16:31:11 INFO DAGScheduler: Job 0 finished: collect at RDDmapExample2.java:23,
2.6390573296152584
3.044522437723423
4.477336814478207
4.59511985013459
6.12029741895095
17/11/28 16:31:11 INFO SparkContext: Invoking stop() from shutdown hook
```

Java Example 2 – Spark RDD.map()

In this example, we will map an RDD of Strings to an RDD of Integers with each element in the mapped RDD representing the number of words in the input RDD. The final mapping would be RDD<String> -> RDD<Integer>.

RDDmapExample.java

```
import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;

public class RDDmapExample {

    public static void main(String[] args) {
        // configure spark
        SparkConf sparkConf = new SparkConf().setAppName("Read Text to RDD")
                                        .setMaster("local[2]").set("spark.executor.memor
        // start a spark context
        JavaSparkContext sc = new JavaSparkContext(sparkConf);

        // provide path to input text file
        String path = "data/rdd/input/sample.txt";

        // read text file to RDD
        JavaRDD<String> lines = sc.textFile(path);

        // map each line to number of words in the line
        JavaRDD<Integer> n_words = lines.map(x -> x.split(" ").length);

        // collect RDD for printing
        for(int n:n_words.collect()){
            System.out.println(n);
        }
    }
}
```

Following is the input text file we used.

data/rdd/input/sample.txt

```
Welcome to TutorialKart
Learn Apache Spark
Learn to work with RDD
```

Run the above Java Example, and you would get the following output in console.

```
17/11/28 16:25:22 INFO DAGScheduler: ResultStage 0 (collect at RDDmapExample.java:24) fi
17/11/28 16:25:22 INFO DAGScheduler: Job 0 finished: collect at RDDmapExample.java:24, t
3
3
5
17/11/28 16:25:22 INFO SparkContext: Invoking stop() from shutdown hook
```

We have successfully created a new RDD with strings transformed to number of words in it.

Python Example 1 – Spark RDD.map()

In this example, we will map integers in RDD to their logarithmic values using Python.

spark-rdd-map-example-2.py

```
import sys, math

from pyspark import SparkContext, SparkConf

if __name__ == "__main__":

  # create Spark context with Spark configuration
  conf = SparkConf().setAppName("Map Numbers to their Log Values - Python")
  sc = SparkContext(conf=conf)

  # read input text file to RDD
  numbers = sc.parallelize([14,21,88,99,455])

  # map lines to n_words
  log_values = numbers.map(lambda n : math.log10(n))

  # collect the RDD to a list
  llist = log_values.collect()

  # print the list
  for line in llist:
    print line
```

Run the following command to submit this Python program to run as Spark Application.

```
$ spark-submit spark-rdd-map-example-2.py
```

Following is the output of this Python Application in console.

```
17/11/28 19:40:42 INFO DAGScheduler: ResultStage 0 (collect at /home/arjun/workspace/spa
17/11/28 19:40:42 INFO DAGScheduler: Job 0 finished: collect at /home/arjun/workspace/sp
1.14612803568
1.3.6.0929473
1.94448267215
1.9956351946
2.65801139666
17/11/28 19:40:42 INFO SparkContext: Invoking stop() from shutdown hook
```

Python Example 2 – Spark RDD.map()

In this example, we will map sentences to number of words in the sentence.

spark-rdd-map-example.py

```
import sys
```

```
from pyspark import SparkContext, SparkConf

if __name__ == "__main__":

  # create Spark context with Spark configuration
  conf = SparkConf().setAppName("Read Text to RDD - Python")
  sc = SparkContext(conf=conf)

  # read input text file to RDD
  lines = sc.textFile("/home/arjun/workspace/spark/sample.txt")

  # map lines to n_words
  n_words = lines.map(lambda line : len(line.split()))

  # collect the RDD to a list
  llist = n_words.collect()

  # print the list
  for line in llist:
    print line
```

Run the above python program using following spark-submit command.

```
$ spark-submit spark-rdd-map-example.py
```

```
17/11/28 19:31:44 INFO DAGScheduler: ResultStage 0 (collect at /home/arjun/workspace/spa
17/11/28 19:31:44 INFO DAGScheduler: Job 0 finished: collect at /home/arjun/workspace/sp
3
3
5
17/11/28 19:31:44 INFO SparkContext: Invoking stop() from shutdown hook
```

Conclusion

In this Spark Tutorial, we learned the syntax and examples for RDD.map() method.

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