

# Spark by {Examples} (https://sparkbyexamples.com/)

## Spark Tutorial

[Spark – Installation on Windows \(https://sparkbyexamples.com/spark/apache-spark-installation-on-windows/\)](https://sparkbyexamples.com/spark/apache-spark-installation-on-windows/)

[Spark – Installation on Linux | Ubuntu \(https://sparkbyexamples.com/spark/spark-installation-on-linux-ubuntu/\)](https://sparkbyexamples.com/spark/spark-installation-on-linux-ubuntu/)

[Spark – Cluster Setup with Hadoop Yarn \(https://sparkbyexamples.com/spark/spark-setup-on-hadoop-yarn/\)](https://sparkbyexamples.com/spark/spark-setup-on-hadoop-yarn/)

[Spark – Web/Application UI \(https://sparkbyexamples.com/spark/spark-web-ui-understanding/\)](https://sparkbyexamples.com/spark/spark-web-ui-understanding/)

[Spark – Setup with Scala and IntelliJ \(https://sparkbyexamples.com/spark/spark-setup-run-with-scala-intellij/\)](https://sparkbyexamples.com/spark/spark-setup-run-with-scala-intellij/)

[Spark – How to Run Examples From this Site on IntelliJ IDEA \(https://sparkbyexamples.com/spark/how-to-run-spark-examples-from-intellij/\)](https://sparkbyexamples.com/spark/how-to-run-spark-examples-from-intellij/)

[Spark – SparkSession \(https://sparkbyexamples.com/spark/sparksession-explained-with-examples/\)](https://sparkbyexamples.com/spark/sparksession-explained-with-examples/)

[Spark – SparkContext \(https://sparkbyexamples.com/spark/spark-sparkcontext/\)](https://sparkbyexamples.com/spark/spark-sparkcontext/)

## Spark RDD Tutorial

[Spark RDD – Parallelize \(https://sparkbyexamples.com/apache-spark-rdd/how-to-create-an-rdd-using-parallelize/\)](https://sparkbyexamples.com/apache-spark-rdd/how-to-create-an-rdd-using-parallelize/)

[Spark RDD – Read text file \(https://sparkbyexamples.com/apache-spark-rdd/spark-read-multiple-text-files-into-a-single-rdd/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-read-multiple-text-files-into-a-single-rdd/)

[PySpark \(https://sparkbyexamples.com/pyspark-tutorial/\)](https://sparkbyexamples.com/pyspark-tutorial/)

[Hive \(https://sparkbyexamples.com/apache-hive-tutorial/\)](https://sparkbyexamples.com/apache-hive-tutorial/)

[HBase \(https://sparkbyexamples.com/apache-hbase-tutorial/\)](https://sparkbyexamples.com/apache-hbase-tutorial/)

[Kafka \(https://sparkbyexamples.com/apache-kafka-tutorials-with-examples/\)](https://sparkbyexamples.com/apache-kafka-tutorials-with-examples/)

[FAQ's \(https://sparkbyexamples.com/spark-faq/\)](https://sparkbyexamples.com/spark-faq/)

[More Questions \(https://sparkbyexamples.com/more-questions/\)](https://sparkbyexamples.com/more-questions/)

Chocolate Bouquet @ Rs549  
Ferns N Petals

## Spark map() vs flatMap() with Examples

What is the difference between Spark map() vs flatMap() is a most asked interview question, if you are taking an interview on Spark (Java/Scala/PySpark), so let's understand the differences with examples. Regardless of an interview, you have to know the differences as this is also one of the most used Spark transformations.

- [map\(\)](#) – Spark map() transformation applies a function to each row in a DataFrame/Dataset and returns the new transformed Dataset.



[Spark RDD – Read CSV  
\(https://sparkbyexamples.com/apache-spark-rdd/spark-load-csv-file-into-rdd/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-load-csv-file-into-rdd/)

---

[Spark RDD – Create RDD  
\(https://sparkbyexamples.com/apache-spark-rdd/different-ways-to-create-spark-rdd/\)](https://sparkbyexamples.com/apache-spark-rdd/different-ways-to-create-spark-rdd/)

---

[Spark RDD – Create Empty RDD  
\(https://sparkbyexamples.com/apache-spark-rdd/spark-how-to-create-an-empty-rdd/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-how-to-create-an-empty-rdd/)

---

[Spark RDD – Transformations  
\(https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-transformations/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-transformations/)

---

[Spark RDD – Actions  
\(https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-actions/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-actions/)

---

[Spark RDD – Pair Functions  
\(https://sparkbyexamples.com/apache-spark-rdd/spark-pair-rdd-functions/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-pair-rdd-functions/)

---

[Spark RDD – Repartition and Coalesce  
\(https://sparkbyexamples.com/spark/spark-repartition-vs-coalesce/\)](https://sparkbyexamples.com/spark/spark-repartition-vs-coalesce/)

---

[Spark RDD – Shuffle Partitions  
\(https://sparkbyexamples.com/spark/spark-shuffle-partitions/\)](https://sparkbyexamples.com/spark/spark-shuffle-partitions/)

---

[Spark RDD – Cache vs Persist  
\(https://sparkbyexamples.com/spark/spark-difference-between-cache-and-persist/\)](https://sparkbyexamples.com/spark/spark-difference-between-cache-and-persist/)

---

[Spark RDD – Persistence Storage Levels  
\(https://sparkbyexamples.com/spark/spark-persistence-storage-levels/\)](https://sparkbyexamples.com/spark/spark-persistence-storage-levels/)

---

[Spark RDD – Broadcast Variables  
\(https://sparkbyexamples.com/spark/spark-broadcast-variables/\)](https://sparkbyexamples.com/spark/spark-broadcast-variables/)

---

[Spark RDD – Accumulator Variables  
\(https://sparkbyexamples.com/spark/spark-accumulators/\)](https://sparkbyexamples.com/spark/spark-accumulators/)

---

[Spark RDD – Convert RDD to DataFrame  
\(https://sparkbyexamples.com/spark/spark-convert-rdd-to-dataframe/\)](https://sparkbyexamples.com/spark/spark-convert-rdd-to-dataframe/)

- [flatMap\(\)](#) – Spark flatMap() transformation flattens the DataFrame/Dataset after applying the function on every element and returns a new transformed Dataset. The returned Dataset will return more rows than the current DataFrame. It is also referred to as a one-to-many transformation function. This is one of the major differences between flatMap() and map()

### Key points

- Both map() & flatMap() returns Dataset (DataFrame=Dataset[Row]).
- Both these transformations are narrow meaning they do not result in Spark Data Shuffle (<https://sparkbyexamples.com/spark/spark-shuffle-partitions/>).
- flatMap() results in redundant data on some columns.
- One of the use cases of flatMap() is to flatten column which contains arrays, list, or any nested collection(one cell with one value).
- map() always return the same size/records as in input DataFrame whereas flatMap() returns many records for each record (one-many).

## Spark map vs flatMap with Examples

Let's see the difference with an example. First, let's create a DataFrame that I will use for map() and flatMap() transformation.

[apache-spark-rdd/convert-spark-rdd-to-dataframe-dataset/](#)

## Spark SQL Tutorial

[Spark SQL – Create DataFrame](#)

(<https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/>).

[Spark SQL – Select Columns](#)

(<https://sparkbyexamples.com/spark/spark-select-columns-from-dataframe/>).

[Spark SQL – Add and Update Column \(withColumn\)](#)

(<https://sparkbyexamples.com/spark/spark-dataframe-withcolumn/>).

[Spark SQL – Rename Nested Column](#)

(<https://sparkbyexamples.com/spark/rename-a-column-on-spark-dataframes/>).

[Spark SQL – Drop column](#)

(<https://sparkbyexamples.com/spark/spark-drop-column-from-dataframe-dataset/>).

[Spark SQL – Where | Filter](#)

(<https://sparkbyexamples.com/spark/spark-dataframe-where-filter/>).

[Spark SQL – When Otherwise](#)

(<https://sparkbyexamples.com/spark/spark-case-when-otherwise-example/>).

[Spark SQL – Collect data to Driver](#)

(<https://sparkbyexamples.com/spark/spark-dataframe-collect/>).

[Spark SQL – Distinct](#)

(<https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/>).

[Spark SQL- Pivot Table DataFrame](#)

(<https://sparkbyexamples.com/spark/how-to-pivot-table-and-unpivot-a-spark-dataframe/>).

[Spark SQL – Data Types](#)

(<https://sparkbyexamples.com/spark/spark-sql-dataframe-data-types/>).

```
val data = Seq("Project Gutenberg's",
               "Alice's Adventures in Wonderland",
               "Project Gutenberg's",
               "Adventures in Wonderland",
               "Project Gutenberg's")

import spark.sqlContext.implicits._
val df = data.toDF("data")
df.show()
```

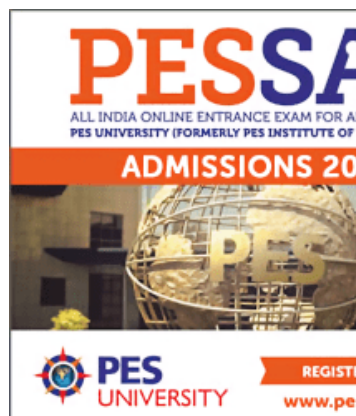
//Output

```
+-----+
|          data |
+-----+
| Project Gutenberg's |
| Alice's Adventure... |
| Project Gutenberg's |
| Adventures in Won... |
| Project Gutenberg's |
+-----+
```

## Spark Map Transformation

Spark `map()` transformation applies a function to each row in a `DataFrame/Dataset` and returns the new transformed `Dataset`.

If you notice the below signatures, both these functions return `Dataset[U]` but not `DataFrame` (`DataFrame=Dataset[Row]`). If you want a `DataFrame` as output then you need to convert the `Dataset` to `DataFrame` using `toDF()` function.



**Syntax:**

[Spark SQL – StructType | StructField](#)  
(<https://sparkbyexamples.com/spark/spark-sql-structtype-on-dataframe/>).

[Spark SQL – Schema](#)  
(<https://sparkbyexamples.com/spark/spark-schema-explained-with-examples/>).

[Spark SQL – Groupby](#)  
(<https://sparkbyexamples.com/spark/using-groupby-on-dataframe/>).

[Spark SQL – Sort DataFrame](#)  
(<https://sparkbyexamples.com/spark/spark-how-to-sort-dataframe-column-explained/>).

[Spark SQL – Join Types](#)  
(<https://sparkbyexamples.com/spark/spark-sql-dataframe-join/>).

[Spark SQL – Union and UnionAll](#)  
(<https://sparkbyexamples.com/spark/spark-dataframe-union-and-union-all/>).

[Spark SQL – map\(\) vs mapPartitions\(\)](#)  
(<https://sparkbyexamples.com/spark/spark-map-vs-mappartitions-transformation/>).

[Spark SQL – foreach\(\) vs foreachPartition\(\)](#)  
(<https://sparkbyexamples.com/spark/spark-foreachpartition-vs-foreach-explained/>).

[Spark SQL – map\(\) vs flatMap\(\)](#)  
(<https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/>).

[Spark SQL – Persist and Cache](#)  
(<https://sparkbyexamples.com/spark/spark-dataframe-cache-and-persist-explained/>).

[Spark SQL – UDF \(User Defined Functions\)](#)  
(<https://sparkbyexamples.com/spark/spark-sql-udf/>).

[Spark SQL – Array \(ArrayType\) Column](#)  
(<https://sparkbyexamples.com/spark/spark-array-arraytype-dataframe-column/>).

```
1) map[U](func : scala.Function1,
           : org.apache.spark.sql.
2) map[U](func : org.apache.spa
           : org.apache.spark.sql.
```

**Example:** Here, I have split the value of the column by space using map() transformation, the split() function returns an Array hence the column on DataFrame converted from String to Array Type. you can check this by calling mapDF.printSchema().

```
//Map Transformation
val mapDF=df.map(fun=> {
    fun.getString(0).split(" ")
})
mapDF.show()

//Output
+-----+
|value|
+-----+
|[Project, Gutenberg's]|
|[Alice's, Adventures, in, Wond]|
|[Project, Gutenberg's]|
|[Adventures, in, Wonderland]|
|[Project, Gutenberg's]|
+-----+
```

## Spark flatMap()

### Transformation

Spark flatMap() transformation flattens the DataFrame column after applying the function on every element and returns a new DataFrame respectively.

The returned DataFrame can have the same count or more elements than the current DataFrame. This is one of the major differences between flatMap() and map(), where [map\(\) transformation](#) always returns the same number of elements as in input  
(<https://sparkbyexamples.com/spark/spark-map-transformation/>).

[Spark SQL – Map \(MapType\) column](#)  
(<https://sparkbyexamples.com/spark/spark-dataframe-map-maptype-column/>).

[Spark SQL – Flatten Nested Struct Column](#)  
(<https://sparkbyexamples.com/spark/spark-flatten-nested-struct-column/>).

[Spark SQL – Flatten Nested Array Column](#)  
(<https://sparkbyexamples.com/spark/spark-flatten-nested-array-column-to-single-column/>).

[Spark SQL – Explode Array & Map Columns](#)  
(<https://sparkbyexamples.com/spark/explode-spark-array-and-map-dataframe-column/>).

[Spark SQL – Sampling](#)  
(<https://sparkbyexamples.com/spark/spark-sampling-with-examples/>).

[Spark SQL – Partitioning](#)  
(<https://sparkbyexamples.com/spark/spark-partitioning-understanding/>).

## Spark SQL Functions

[Spark SQL String Functions](#)  
(<https://sparkbyexamples.com/spark/usage-of-spark-sql-string-functions/>).

[Spark SQL Date and Timestamp Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-date-and-time-functions/>).

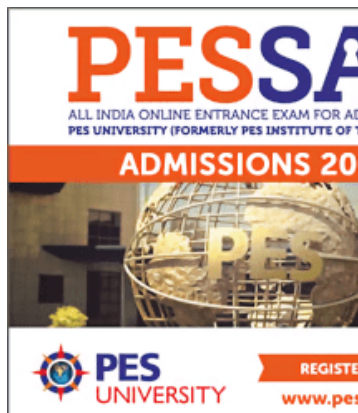
[Spark SQL Array Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-array-functions/>).

[Spark SQL Map Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-map-functions/>).

[Spark SQL Sort Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-sort-functions/>).

[Spark SQL Aggregate Functions](#)  
(<https://sparkbyexamples.com/>).

## Syntax:



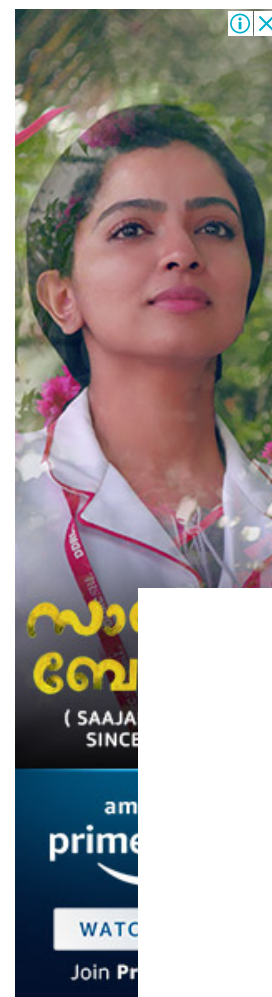
```
flatMap[U](f : scala.Function1[  
    org.apache.spark.rdd.R
```

**Example 1:** Like on `map()` example, on `flatMap()` also, I have used the `split()` function and it returns an Array. However, `flatMap()` converted the array into a row. hence you end up with more records than in input DataFrame.

```
//Flat Map Transformation  
val flatMapDF=df.flatMap(fun=>  
  {  
    fun.getString(0).split(" ")  
  })  
flatMapDF.show()
```

```
//Output  
+-----+  
|      value|  
+-----+  
|   Project|  
|Gutenberg's|  
|   Alice's|  
|Adventures|  
|      in|  
|Wonderland|  
|   Project|  
|Gutenberg's|  
|Adventures|  
|      in|  
|Wonderland|  
|   Project|  
|Gutenberg's|  
+-----+
```

**Example 2:**



[spark/spark-sql-aggregate-functions/](#)

[Spark SQL Window Functions \(https://sparkbyexamples.com/spark/spark-sql-window-functions/\)](#)

[Spark SQL JSON Functions \(https://sparkbyexamples.com/spark/spark-most-used-json-functions-with-examples/\)](#)

## Spark Data Source API

[Spark – Read & Write CSV file \(https://sparkbyexamples.com/spark/spark-read-csv-file-into-dataframe/\)](#)

[Spark – Read and Write JSON file \(https://sparkbyexamples.com/spark/spark-read-and-write-json-file/\)](#)

[Spark – Read & Write Parquet file \(https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/\)](#)

[Spark – Read & Write XML file \(https://sparkbyexamples.com/spark/spark-read-write-xml/\)](#)

[Spark – Read & Write Avro files \(https://sparkbyexamples.com/spark/read-write-avro-file-spark-dataframe/\)](#)

[Spark – Read & Write Avro files \(Spark version 2.3.x or earlier\) \(https://sparkbyexamples.com/spark/using-avro-data-files-from-spark-sql-2-3-x/\)](#)

[Spark – Read & Write HBase using “hbase-spark” Connector \(https://sparkbyexamples.com/spark/spark-read-write-using-hbase-spark-connector/\)](#)

[Spark – Read & Write from HBase using Hortonworks \(https://sparkbyexamples.com/spark/create-spark-dataframe-from-hbase-using-hortonworks/\)](#)

[Spark – Read & Write ORC file \(https://sparkbyexamples.com/spark/spark-read-orc-file-into-dataframe/\)](#)

```
val arrayStructureData = Seq(
  Row("James,,Smith",List("Java","Scala","C++")),
  Row("Michael,Rose,",List("Spark","Java","C++")),
  Row("Robert,,Williams",List("CSharp","VB","R"))
)

val arrayStructureSchema = new StructType()
  .add("name",StringType)
  .add("languagesAtSchool", ArrayType(StringType))
  .add("currentState", StringType)

val df = spark.createDataFrame(
  spark.sparkContext.parallelize(arrayStructureData),
  arrayStructureSchema
)

import spark.implicits._

//flatMap() Usage
val df2=df.flatMap(f => {
  val lang=f.getSeq[String](1)
  lang.map((f.getString(0),_))
})

val df3=df2.toDF("Name","Language","State")
df3.show(false)
```

#Outputs

Name	Language	State
James,,Smith	Java	CA
James,,Smith	Scala	CA
James,,Smith	C++	CA
Michael,Rose,	Spark	NJ
Michael,Rose,	Java	NJ
Michael,Rose,	C++	NJ
Robert,,Williams	CSharp	NV
Robert,,Williams	VB	NV
Robert,,Williams	R	NV

## Conclusion of Map() vs flatMap()

In this article, you have learned map() and flatMap() are transformations that exists in both RDD and DataFrame. map() transformation is used to transform the data into different values, types by returning the same number of records. flatMap() transformation is used to transform from one record to multiple records.



[Spark – Read Binary File](https://sparkbyexamples.com/spark/spark-read-binary-file-into-dataframe/)  
(<https://sparkbyexamples.com/spark/spark-read-binary-file-into-dataframe/>).

## Spark Streaming & Kafka

[Spark Streaming – OutputModes](https://sparkbyexamples.com/spark/spark-streaming-outputmode/)  
(<https://sparkbyexamples.com/spark/spark-streaming-outputmode/>).

[Spark Streaming – Reading Files From Directory](https://sparkbyexamples.com/spark/spark-streaming-read-json-files-from-directory/)  
(<https://sparkbyexamples.com/spark/spark-streaming-read-json-files-from-directory/>).

[Spark Streaming – Reading Data From TCP Socket](https://sparkbyexamples.com/spark/spark-streaming-from-tcp-socket/)  
(<https://sparkbyexamples.com/spark/spark-streaming-from-tcp-socket/>).

[Spark Streaming – Processing Kafka Messages in JSON Format](https://sparkbyexamples.com/spark/spark-streaming-with-kafka/)  
(<https://sparkbyexamples.com/spark/spark-streaming-with-kafka/>).


[Spark Streaming – Processing Kafka messages in AVRO Format](https://sparkbyexamples.com/spark/spark-streaming-consume-and-produce-kafka-messages-in-avro-format/)  
(<https://sparkbyexamples.com/spark/spark-streaming-consume-and-produce-kafka-messages-in-avro-format/>).


[Spark SQL Batch – Consume & Produce Kafka Message](https://sparkbyexamples.com/spark/spark-batch-processing-produce-consume-kafka-topic/)  
(<https://sparkbyexamples.com/spark/spark-batch-processing-produce-consume-kafka-topic/>).





Happy Learning !!


Share this:


 (<https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/?share=facebook&nb=1>)


 (<https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/?share=reddit&nb=1>)

 (<https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/?share=pinterest&nb=1>)

 (<https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/?share=tumblr&nb=1>)

 (<https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/?share=pocket&nb=1>)

 (<https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/?share=linkedin&nb=1>)

 (<https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/?share=twitter&nb=1>)

TAGS: [FLATMAP\(\)](https://sparkbyexamples.com/tag/flatmap/)  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/FLATMAP-2/](https://sparkbyexamples.com/tag/flatmap/)), [MAP\(\)](https://sparkbyexamples.com/tag/map/)  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/MAP-2/](https://sparkbyexamples.com/tag/map-2/)).



**NNK**

(<https://sparkbyexamples.com/author/admin/>)

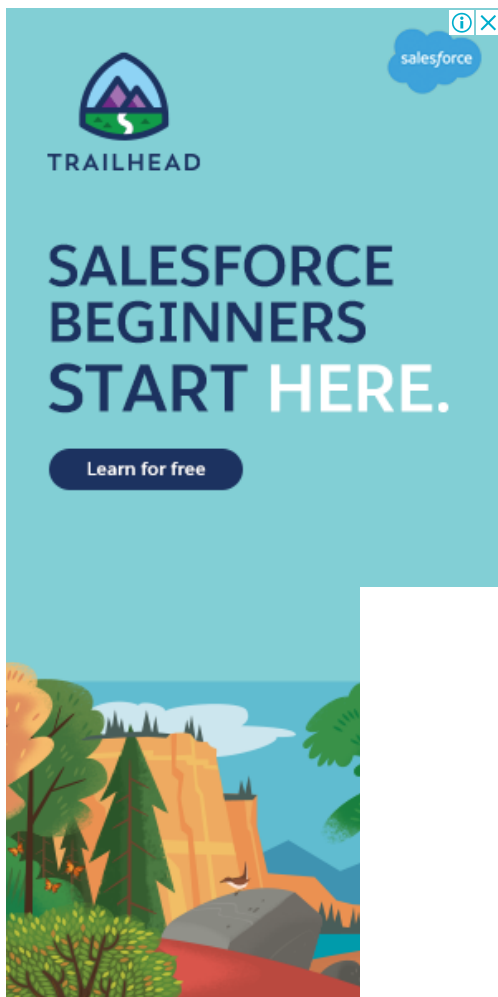
(<https://sparkbyexamples.com/author/admin/>).

SparkByExamples.com is a Big Data and Spark examples community page, all examples are simple and easy to understand and well tested in our development environment [Read more...](https://sparkbyexamples.com/about-sparkbyexamples/)  
(<https://sparkbyexamples.com/about-sparkbyexamples/>).

ave a Reply

ter your comment here...





## Personalized Tissue

Custom Tissue Holder  
With Name & Charm in  
Wine, Tan, Black &  
Many More Variants.  
Shop Now!

## The Messy Corner

## Categories

Apache Hadoop  
(<https://sparkbyexamples.com/category/hadoop/>)

Apache Spark  
(<https://sparkbyexamples.com/category/spark/>)

Apache Spark Streaming  
(<https://sparkbyexamples.com/category/spark/apache-spark-streaming/>)

Apache Kafka  
(<https://sparkbyexamples.com/category/kafka/>)

Apache HBase  
(<https://sparkbyexamples.com/category/hbase/>)

Apache Cassandra  
(<https://sparkbyexamples.com/category/cassandra/>)

Snowflake Database  
(<https://sparkbyexamples.com/category/snowflake/>)

H2O Sparkling Water  
(<https://sparkbyexamples.com/category/h2o-sparkling-water/>)

## Recent Posts

## Spark regexp\_replace() – Replace String Value

([https://sparkbyexamples.com/spark/spark-regexp\\_replace-replace-string-value/](https://sparkbyexamples.com/spark/spark-regexp_replace-replace-string-value/))

## How to Run a PySpark Script from Python?

(<https://sparkbyexamples.com/pyspark/run-pyspark-script-from-python-subprocess/>)

Spark SQL like() Using Wildcard  
Example  
(<https://sparkbyexamples.com/spark/spark-sql-like-using-wildcard-example/>)

## Spark isin() & IS NOT IN Operator Example

(<https://sparkbyexamples.com/spark/spark-isin-is-not-in-operator-example/>)

Spark – Get Size/Length of Array & Map Column  
(<https://sparkbyexamples.com/spark/spark-get-size-length-of-array-map-column/>)

## Spark Using Length/Size Of a DataFrame Column

(<https://sparkbyexamples.com/spark/spark-using-length-size-of-a-dataframe-column/>)

■ About SparkByExamples.Com

SparkByExamples.com is a Big Data and Spark examples community page, all examples are simple and easy to understand, and well tested in our development environment Read more .. (<https://sparkbyexamples.com/about-sparkbyexamples/>)

Follow Us

(https:/

/www.f (https:/

(https:/ acebo /www.li (https:/

/twitter ok.co nkedin /github

.com/s m/spar .com/i .com/s

parkby kbyexa n/n-nk- park-

examp mples/ b860a examp

les) ) 8193/) les/)





PySpark  
(<https://sparkbyexamples.com/category/pyspark/>)

---

Spark rlike() Working with Regex  
Matching Examples  
(<https://sparkbyexamples.com/spark/spark-rlike-regex-matching-examples/>)

---

Spark Check String Column Has  
Numeric Values  
(<https://sparkbyexamples.com/spark/spark-check-string-column-has-numeric-values/>)

---

Spark Check Column Data Type is  
Integer or String  
(<https://sparkbyexamples.com/spark/spark-check-column-data-type-is-integer-or-string/>)

---

Copyright sparkbyexamples.com

