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

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

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

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

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

FAQ's (https://sparkbyexamples.com/spark-questions/)

More ✔ (https://sparkbyexamples.com/) 🔍

PySpark map() Transformation

👤 NNK (https://sparkbyexamples.com/author/admin/) -
📁 PySpark (https://sparkbyexamples.com/category/pyspark/)

PySpark map (`map()`) is an RDD transformation that is used to apply the transformation function (lambda) on every element of RDD/DataFrame and returns a new RDD. In this article, you will learn the syntax and usage of the RDD map() transformation with an example and how to use it with DataFrame.

RDD `map()` transformation is used to apply any complex operations like adding a column, updating a column, transforming the data e.t.c, the output

of map transformations would always
have the same number of records as
input.

- **Note1:** DataFrame doesn't have
  map() transformation to use with
  DataFrame hence you need to
  DataFrame to RDD first.
- **Note2:** If you have a heavy
  initialization use PySpark
  mapPartitions() transformation
  instead of map(), as with
  mapPartitions() heavy initialization
  executes only once for each partition
  instead of every record.

Related: Spark map() vs
mapPartitions() Explained with
Examples
(https://sparkbyexamples.com/spark/sp
ark-map-vs-mappartitions-
transformation/)

First, let's create an RDD from the list.

```
from pyspark.sql import SparkSe
spark = SparkSession.builder.ma
    .appName("SparkByExamples.c

data = ["Project","Gutenberg's"
"in","Wonderland","Project","Gu
"in","Wonderland","Project","Gu

rdd=spark.sparkContext.parallel
```

map() Syntax

```
map(f, preservesPartitioning=Fa
```

PySpark map() Example
with RDD

In this PySpark map() example, we are
adding a new element with value 1 for
each element, the result of the RDD is
PairRDDFunctions which contains
key-value pairs, word of type String as
Key and 1 of type Int as value.

```
rdd2=rdd.map(lambda x: (x,1))
for element in rdd2.collect():
    print(element)
```

This yields below output.

```
('Project', 1)
('Gutenberg's', 1)
('Alice's', 1)
('Adventures', 1)
('in', 1)
('Wonderland', 1)
('Project', 1)
('Gutenberg's', 1)
('Adventures', 1)
('in', 1)
('Wonderland', 1)
('Project', 1)
('Gutenberg's', 1)
```

PySpark map() Example
with DataFrame

PySpark DataFrame doesn't have
`map()` transformation to apply the
lambda function, when you wanted to
apply the custom transformation, you
need to convert the DataFrame to RDD
and apply the map() transformation.
Let's use another dataset to explain
this.

PySpark SQL Functions

PySpark Datasources

```
data = [('James','Smith','M',30
    ('Anna','Rose','F',41),
    ('Robert','Williams','M',62),
]

columns = ["firstname","lastname
df = spark.createDataFrame(data
df.show()
+---------+--------+------+----
|firstname|lastname|gender|sala
+---------+--------+------+----
|    James|   Smith|     M|
|     Anna|    Rose|     F|
|   Robert|Williams|     M|
+---------+--------+------+----
```

```
# Refering columns by index.
rdd2=df.rdd.map(lambda x:
    (x[0]+","+x[1],x[2],x[3]*2)
    )
df2=rdd2.toDF(["name","gender",
df2.show()
+---------------+------+-------
|           name|gender|new_sal
+---------------+------+-------
|    James,Smith|     M|
|     Anna,Rose|     F|
|Robert,Williams|     M|
+---------------+------+-------
```

Note that abovel have used index to get
the column values, alternatively, you
can also refer to the DataFrame column
names while iterating.

```
# Referring Column Names
rdd2=df.rdd.map(lambda x:
    (x["firstname"]+","+x["last
    )
```

Another alternative

```
# Referring Column Names
rdd2=df.rdd.map(lambda x:
    (x.firstname+","+x.lastname
    )
```

You can also create a custom function to perform an operation. Below `func1()` function executes for every DataFrame row from the lambda function.

```
# By Calling function
def func1(x):
    firstName=x.firstname
    lastName=x.lastname
    name=firstName+","+lastName
    gender=x.gender.lower()
    salary=x.salary*2
    return (name,gender,salary)

rdd2=df.rdd.map(lambda x: func1
```

Complete PySpark map() example

Below is complete example of PySpark map() transformation.

```python
from pyspark.sql import SparkSe
spark = SparkSession.builder.ap

data = ["Project",
"Gutenberg's",
"Alice's",
"Adventures",
"in",
"Wonderland",
"Project",
"Gutenberg's",
"Adventures",
"in",
"Wonderland",
"Project",
"Gutenberg's"]

rdd=spark.sparkContext.parallel

rdd2=rdd.map(lambda x: (x,1))
for element in rdd2.collect():
    print(element)

data = [('James','Smith','M',30
  ('Anna','Rose','F',41),
  ('Robert','Williams','M',62),
]

columns = ["firstname","lastnam
df = spark.createDataFrame(data
df.show()

rdd2=df.rdd.map(lambda x:
    (x[0]+","+x[1],x[2],x[3]*2)
    )
df2=rdd2.toDF(["name","gender",
df2.show()

#Referring Column Names
rdd2=df.rdd.map(lambda x:
    (x["firstname"]+","+x["last
    )

#Referring Column Names
rdd2=df.rdd.map(lambda x:
    (x.firstname+","+x.lastname
    )

def func1(x):
    firstName=x.firstname
    lastName=x.lastname
    name=firstName+","+lastName
    gender=x.gender.lower()
    salary=x.salary*2
    return (name,gender,salary)
```

```
rdd2=df.rdd.map(lambda x: func1
```

In conclusion, you have learned how to apply a `map()` transformation on every element of PySpark RDD and learned it returns the same number of elements as input RDD. This is one of the differences between `map()` and `flatMap()` (https://sparkbyexamples.com/pys park/pyspark-rdd-flatmap-transformation/) transformations. And you have also learned how to use map() on DataFrame by converting DataFrame to RDD.

Happy Learning !!

Reference:
https://spark.apache.org/docs/latest/api /python/pyspark.sql.html (https://spark.apache.org/docs/latest/ap i/python/pyspark.sql.html)

Share this:

(https://sparkbyexamples.com/pyspark/pyspark-
map-transformation/?share=facebook&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-
map-transformation/?share=reddit&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-
map-transformation/?share=pinterest&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-
map-transformation/?share=tumblr&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-
map-transformation/?share=pocket&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-
map-transformation/?share=linkedin&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-
map-transformation/?share=twitter&nb=1)

**TAGS: MAP()
(HTTPS://SPARKBYEXAMPLES.COM/TAG/MAP-2/)**

NNK (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/)

Leave a Reply

Enter your comment here...

Categories

Apache Hadoop
(https://sparkbyexamples.com/catego
ry/hadoop/)

Apache Spark
(https://sparkbyexamples.com/catego
ry/spark/)

Apache Spark Streaming
(https://sparkbyexamples.com/catego
ry/spark/apache-spark-streaming/)

Apache Kafka
(https://sparkbyexamples.com/catego
ry/kafka/)

Apache HBase
(https://sparkbyexamples.com/catego
ry/hbase/)

Apache Cassandra
(https://sparkbyexamples.com/catego
ry/cassandra/)

Snowflake Database
(https://sparkbyexamples.com/catego
ry/snowflake/)

H2O Sparkling Water
(https://sparkbyexamples.com/catego

Recent Posts

Spark regexp_replace() – Replace
String Value
(https://sparkbyexamples.com/spark/sp
ark-regexp_replace-replace-string-
value/)

How to Run a PySpark Script from
Python?
(https://sparkbyexamples.com/pyspark/r
un-pyspark-script-from-python-
subprocess/)

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

Spark isin() & IS NOT IN Operator
Example
(https://sparkbyexamples.com/spark/sp
ark-isin-is-not-in-operator-example/)

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

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

ry/h2o-sparkling-water/)

PySpark
(https://sparkbyexamples.com/catego
ry/pyspark/)

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

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

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

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