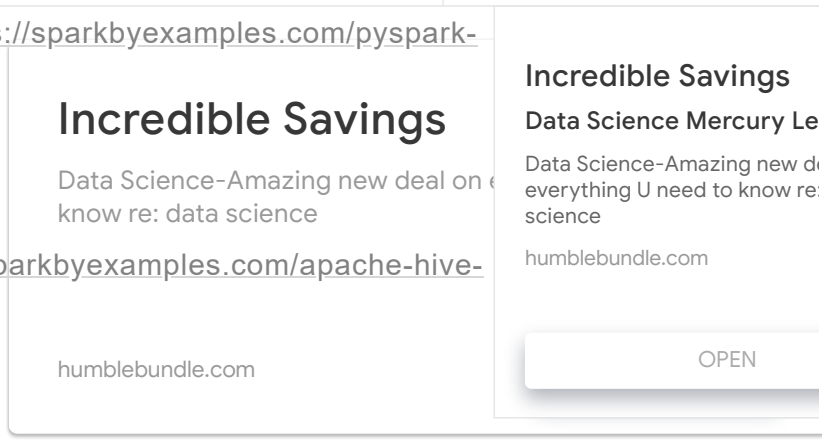
PySpark Tutorial

- PySpark Tutorial For Beginners (https://sparkbyexamples.com/pyspark-tutorial/)
- PySpark – Features (https://sparkbyexamples.com/pyspark-tutorial/#features)
- PySpark – Advantages (https://sparkbyexamples.com/pyspark-tutorial/#advantages)
- PySpark – Modules & Packages (https://sparkbyexamples.com/pyspark-tutorial/#modules-packages)
- PySpark – Cluster Managers (https://sparkbyexamples.com/pyspark-tutorial/#cluster-manager)
- PySpark – Install on Windows (https://sparkbyexamples.com/pyspark-tutorial/#pyspark-installation)
- PySpark – Web/Application UI (https://sparkbyexamples.com/spark/spark-web-ui-understanding/)
- PySpark – SparkSession (https://sparkbyexamples.com/pyspark/pyspark-what-is-sparksession/)
- PySpark – RDD (https://sparkbyexamples.com/pyspark-rdd)
- PySpark – Parallelize (https://sparkbyexamples.com/pyspark/pyspark-parallelize-create-rdd/)
- PySpark – repartition() vs coalesce() (https://sparkbyexamples.com/pyspark/pyspark-repartition-vs-coalesce/)
- PySpark – Broadcast Variables (https://sparkbyexamples.com/pyspark/pyspark-broadcast-variables/)



PySpark flatMap() Transformation

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

PySpark `flatMap()` is a transformation operation that flattens the RDD/DataFrame (array/map DataFrame columns) after applying the function on every element and returns a new PySpark RDD/DataFrame. In this article, you will learn the syntax and usage of the PySpark flatMap() with an example.

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

PySpark – Accumulator (https://sparkbyexamples.com/pyspark/pyspark-accumulator-with-example/)

PySpark DataFrame

PySpark – Create a DataFrame (https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/)

PySpark – Create an empty DataFrame (https://sparkbyexamples.com/pyspark/pyspark-create-an-empty-dataframe/)

PySpark – Convert RDD to DataFrame (https://sparkbyexamples.com/pyspark/convert-pyspark-rdd-to-dataframe/)

PySpark – Convert DataFrame to Pandas (https://sparkbyexamples.com/pyspark/convert-pyspark-dataframe-to-pandas/)

PySpark – show() (https://sparkbyexamples.com/pyspark/pyspark-show-display-dataframe-contents-in-table/)

PySpark – StructType & StructField (https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/)

PySpark – Row Class (https://sparkbyexamples.com/pyspark/pyspark-row-using-rdd-dataframe/)

PySpark – Column Class (https://sparkbyexamples.com/pyspark/pyspark-column-functions/)

PySpark – select() (https://sparkbyexamples.com/pyspark/select-columns-from-pyspark-dataframe/)

PySpark – collect() (https://sparkbyexamples.com/pyspark/pyspark-collect/)

PySpark – withColumn() (https://sparkbyexamples.com/pyspark/pyspark-withcolumn/)

```
data = ["Project Gutenberg's",
        "Alice's Adventures in
        "Project Gutenberg's",
        "Adventures in Wonderla
        "Project Gutenberg's"]
rdd=spark.sparkContext.parallel
for element in rdd.collect():
    print(element)
```

This yields the below output

```
Project Gutenberg's
Alice's Adventures in Wonderland
Project Gutenberg's
Adventures in Wonderland
Project Gutenberg's
```

flatMap() Syntax

```
flatMap(f, preservesPartitionin
```

flatMap() Example

Now, let's see with an example of how to apply a flatMap() transformation on RDD. On the below example, first, it splits each record by space in an RDD and finally flattens it. Resulting RDD consists of a single word on each record.

```
rdd2=rdd.flatMap(lambda x: x.sp
for element in rdd2.collect():
    print(element)
```

This yields below output.

PySpark – withColumnRenamed() (https://sparkbyexamples.com/pyspark/pyspark-rename-dataframe-column/)

PySpark – where() & filter() (https://sparkbyexamples.com/pyspark/pyspark-where-filter/)

PySpark – drop() & dropDuplicates() (https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/)

PySpark – orderBy() and sort() (https://sparkbyexamples.com/pyspark/pyspark-orderby-and-sort-explained/)

PySpark – groupBy() (https://sparkbyexamples.com/pyspark/pyspark-groupby-explained-with-example/)

PySpark – join() (https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/)

PySpark – union() & unionAll() (https://sparkbyexamples.com/pyspark/pyspark-union-and-unionall/)

PySpark – unionByName() (https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/)

PySpark – UDF (User Defined Function) (https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/)

PySpark – map() (https://sparkbyexamples.com/pyspark/pyspark-map-transformation/)

PySpark – flatMap() (https://sparkbyexamples.com/pyspark/pyspark-flatmap-transformation/)

pyspark – foreach() (https://sparkbyexamples.com/pyspark/pyspark-loop-iterate-through-rows-in-dataframe/#use-foreach-loop-through-dataframe)

PySpark – sample() vs sampleBy()

```
Project
Gutenberg’s
Alice’s
Adventures
in
Wonderland
Project
Gutenberg’s
Adventures
in
Wonderland
Project
Gutenberg’s
```

Complete PySpark flatMap() example

Below is the complete example of flatMap() function that works with RDD.

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

data = ["Project Gutenberg’s",
        "Alice’s Adventures in
        "Project Gutenberg’s",
        "Adventures in Wonderla
        "Project Gutenberg’s"]
rdd=spark.sparkContext.parallel
for element in rdd.collect():
    print(element)

#Flatmap
rdd2=rdd.flatMap(lambda x: x.sp
for element in rdd2.collect():
    print(element)
```

Using flatMap() transformation on DataFrame

Unfortunately, PySpark DataFame doesn’t have flatMap() transformation however, DataFrame has explode() SQL function that is used to flatten the column (https://sparkbyexamples.com/pyspark/

(https://sparkbyexamples.com/pyspark/pyspark-sampling-example/)

PySpark – fillna() & fill() (https://sparkbyexamples.com/pyspark/pyspark-fillna-fill-replace-null-values/)

PySpark – pivot() (Row to Column) (https://sparkbyexamples.com/pyspark/pyspark-pivot-and-unpivot-dataframe/)

PySpark – partitionBy() (https://sparkbyexamples.com/pyspark/pyspark-partitionby-example/)

PySpark – ArrayType Column (Array) (https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/)

PySpark – MapType (Map/Dict) (https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/)

PySpark SQL Functions

PySpark – Aggregate Functions (https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/)

PySpark – Window Functions (https://sparkbyexamples.com/pyspark/pyspark-window-functions/)

PySpark – Date and Timestamp Functions (https://sparkbyexamples.com/pyspark/pyspark-sql-date-and-timestamp-functions/)

PySpark – JSON Functions (https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/)

PySpark Datasources

PySpark – Read & Write CSV File (https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/)

PySpark – Read & Write Parquet File (https://sparkbyexamples.com/

pyspark-explode-array-and-map-columns-to-rows/). Below is a complete example.

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

arrayData = [
        ('James',['Java','Scala
        ('Michael',['Spark','Ja
        ('Robert',['CSharp',''
        ('Washington',None,None
        ('Jefferson',['1','2'],
df = spark.createDataFrame(data

from pyspark.sql.functions impo
df2 = df.select(df.name,explode
df2.printSchema()
df2.show()
```

This example flattens the array column “knownLanguages” and yields below output

```
root
 |-- name: string (nullable = t
 |-- col: string (nullable = tr

+---------+------+
|     name|   col|
+---------+------+
|    James|  Java|
|    James| Scala|
|  Michael| Spark|
|  Michael|  Java|
|  Michael|  null|
|   Robert|CSharp|
|   Robert|      |
|Jefferson|     1|
|Jefferson|     2|
+---------+------+
```

Conclusion

In conclusion, you have learned how to apply a PySpark flatMap() transformation to flattens the array or map columns and also learned how to use alternatives for DataFrame.

pyspark/pyspark-read-and-write-parquet-file/)

PySpark – Read & Write JSON file (https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/)

Reference

ttps://docs.azuredatabricks.net/_static/notebooks/flat-map-transformer-xample.html
ıttps://docs.azuredatabricks.net/_static/notebooks/flat-map-transformer-example.html)

ⓘ ×

py Learning !!

SSN

JOIN SSN
LEARN DATA SCIENCE
TAKE YOUR CAREER TO THE NE

APPLY NOW

e this:

ps://sparkbyexamples.com/pyspark/pyspark-nap-transformation/?share=facebook&nb=1)

ps://sparkbyexamples.com/pyspark/pyspark-nap-transformation/?share=reddit&nb=1)

ps://sparkbyexamples.com/pyspark/pyspark-nap-transformation/?share=pinterest&nb=1)

ps://sparkbyexamples.com/pyspark/pyspark-nap-transformation/?share=tumblr&nb=1)

ps://sparkbyexamples.com/pyspark/pyspark-nap-transformation/?share=pocket&nb=1)

ps://sparkbyexamples.com/pyspark/pyspark-nap-transformation/?share=linkedin&nb=1)

ps://sparkbyexamples.com/pyspark/pyspark-nap-transformation/?share=twitter&nb=1)

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

PySpark Built-In Functions

PySpark – when() (https://sparkbyexamples.com/pyspark/pyspark-when-otherwise/)

PySpark – expr() (https://sparkbyexamples.com/pyspark/pyspark-sql-expr-expression-function/)

(https://sparkbyexa

NNK (Https://Sparkbyexamples.Com/Author/Admin/)

SparkByExamples.com is a Big Data and Spark examples community page, all examples are simple and

PySpark – lit() (https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/)

PySpark – split() (https://sparkbyexamples.com/pyspark/pyspark-convert-string-to-array-column/)

PySpark – concat_ws() (https://sparkbyexamples.com/pyspark/pyspark-convert-array-column-to-string-column/)

Pyspark – substring() (https://sparkbyexamples.com/pyspark/pyspark-substring-from-a-column/)

PySpark – translate() (https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#translate-replace-character-by-character)

PySpark – regexp_replace() (https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#regexp_replace-replace-string-columns)

PySpark – overlay() (https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#overlay-function)

PySpark – to_timestamp() (https://sparkbyexamples.com/spark/pyspark-to_timestamp-convert-string-to-timestamp-type/)

PySpark – to_date() (https://sparkbyexamples.com/pyspark/pyspark-to_date-convert-timestamp-to-date/)

PySpark – date_format() (https://sparkbyexamples.com/pyspark/pyspark-date_format-convert-date-to-string-format/)

PySpark – datediff() (https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#datediff)

PySpark – months_between() (https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#months_between())

mples.com/author/admin/) easy to understand and well tested in our development environment Read more .. (https://sparkbyexamples.com/about-sparkbyexamples/)

Leave a Reply

Enter your comment here...

PySpark – explode() (https://sparkbyexamples.com/pyspark/pyspark-explode-nested-array-into-rows/)

PySpark – array_contains() (https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array_contains)

PySpark – array() (https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array)

PySpark – collect_list() (https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-list)

PySpark – collect_set() (https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-set)

PySpark – create_map() (https://sparkbyexamples.com/pyspark/pyspark-convert-dataframe-columns-to-maptype-dict/)

PySpark – map_keys() (https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#map_keys)

PySpark – map_values() (https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#map_values)

PySpark – struct() (https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/#update-struct-function)

PySpark – countDistinct() (https://sparkbyexamples.com/pyspark/pyspark-count-distinct-from-dataframe/)

PySpark – sum(), avg() (https://sparkbyexamples.com/pyspark/pyspark-dataframe-groupby-and-sort-by-descending-order/)

PySpark – row_number() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row_number)

PySpark – rank() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#rank)

PySpark – dense_rank() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#dense_rank)

PySpark – percent_rank() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#percent_rank)

PySpark – typedLit() (https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/#typedlit)

PySpark – from_json() (https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#from_json)

PySpark – to_json() (https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#to_json)

PySpark – json_tuple() (https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#json_tuple)

PySpark – get_json_object() (https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#get_json_object)

PySpark – schema_of_json() (https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema_of_json)

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/catego

Recent Posts

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/)

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/cassandra/)

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

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

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

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/)

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/)

(https:/ /www.f acebo ok.co m/spar kbyexa mples/) (https:/ /www.li nkedin .com/i n/n-nk- b860a 8193/) (https:/ /github .com/s parkby examp les) (https:/ park- examp les/)

Copyright sparkbyexamples.com