Home (/) About (https://sparkbyexamples.com/about-sparkbyexamples/) | { One stop for all Spark Examples }

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

Spark (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 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 UDF (User Defined Function)

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

PYSPARK UDF

PySpark

- How to convert python function to PySpark SQL UDF
- Use UDF on DataFrame withColumn() and select()
- Register PySpark UDF and use it on SQL
- Handling null checks
- Performance & optimization considerations

PySpark UDF Example

PySpark UDF (a.k.a User Defined Function) is the most useful feature of Spark SQL & DataFrame that is used to extend the PySpark build in capabilities. In this article, I will explain what is UDF? why do we need it and how to create and use it on DataFrame `select()`, withColumn() (https://sparkbyexamples.com/pyspark/pyspark-dataframe-withcolumn/) and SQL using PySpark (Spark with Python) examples.

[PySpark – Accumulator (https://sparkbyexamples.com/pyspark/pyspark-accumulator-with-example/)](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/)](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/)](https://sparkbyexamples.com/pyspark/pyspark-create-an-empty-dataframe/)

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

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

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

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

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

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

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

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

Note: UDF's are the most expensive operations hence use them only you have no choice and when essential. In the later section of the article, I will explain why using UDF's is an expensive operation in detail.

Table of contents

1. PySpark UDF Introduction

1.1 What is UDF?

UDF's a.k.a User Defined Functions, If you are coming from SQL background, UDF's are nothing new to you as most of the traditional RDBMS databases support User Defined Functions, these

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/)

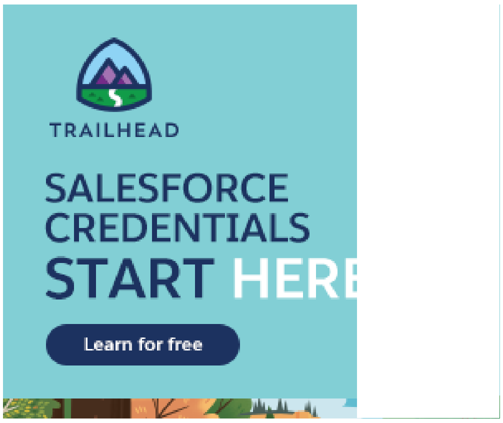
PySpark – flatMap() (https://sparkbyexamples.com/pyspark/pyspark-flatmap-transformation/)

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

PySpark – sample() vs sampleBy()

functions need to register in the database library and use them on SQL as regular functions.

PySpark UDF's are similar to UDF on traditional databases. In PySpark, you create a function in a Python syntax and wrap it with PySpark SQL `udf()` or register it as udf and use it on DataFrame and SQL respectively.



1.2 Why do we need a UDF?

UDF's are used to extend the functions of the framework and re-use these functions on multiple DataFrame's. For example, you wanted to convert every first letter of a word in a name string to a capital case; PySpark build-in features don't have this function hence you can create it a UDF and reuse this as needed on many Data Frames. UDF's are once created they can be re-used on several DataFrame's and SQL expressions.

Before you create any UDF, do your research to check if the similar function you wanted is already available in Spark SQL Functions (https://sparkbyexamples.com/spark/spark-sql-functions-understanding/). PySpark SQL provides several predefined common functions and many more new functions are added with every release. hence, It is best to check before you reinventing the wheel.

When you creating UDF's you need to design them very carefully otherwise you will come across optimization &

(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/

performance issues.

2. Create PySpark UDF

2.1 Create a DataFrame

Before we jump in creating a UDF, first let's create a PySpark DataFrame (https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/).

```
spark = SparkSession.builder.ap

columns = ["Seqno","Name"]
data = [("1", "john jones"),
    ("2", "tracey smith"),
    ("3", "amy sanders")]

df = spark.createDataFrame(data

df.show(truncate=False)
```

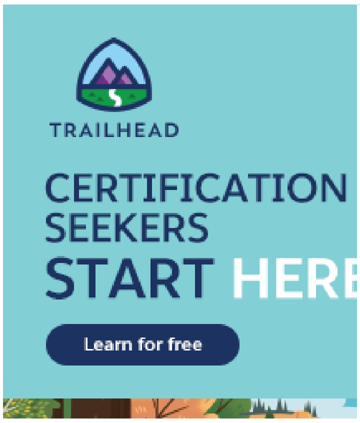
Yields below output.

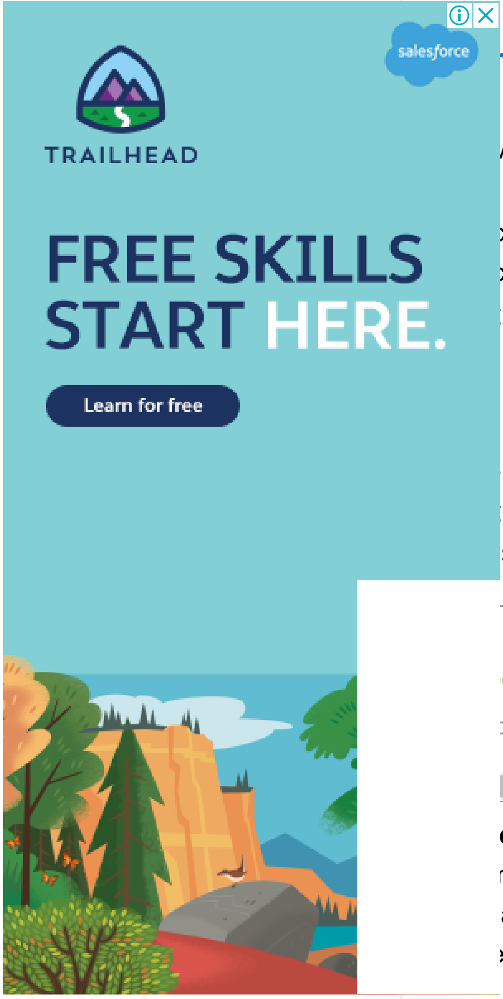
```
+-----+------------+
|Seqno|Names       |
+-----+------------+
|1    |john jones  |
|2    |tracey smith|
|3    |amy sanders |
+-----+------------+
```

2.2 Create a Python Function

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

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





PySpark Built-In Functions

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

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

The first step in creating a UDF is creating a Python function. Below snippet creates a function `convertCase()` which takes a string parameter and converts the first letter of every word to capital letter. UDF's take parameters of your choice and returns a value.

```
ef convertCase(str):
    resStr=""
    arr = str.split(" ")
    for x in arr:
       resStr= resStr + x[0:1].
    return resStr
```

2.3 Convert a Python
nction to PySpark UDF

/ convert this function
vertCase() to UDF by passing the
ction to PySpark SQL udf(), this
ction is available at
.apache.spark.sql.functions.
 package. Make sure you import
 package before using it.

park SQL udf() function returns
.apache.spark.sql.expression
serDefinedFunction class object.

```
'" Converting function to UDF
onvertUDF = udf(lambda z: conv
```

e: The default type of the udf() is
ngType hence, you can also write
above statement without return
e.

```
""" Converting function to UDF
StringType() is by default henc
convertUDF = udf(lambda z: conv
```

- 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())

3. Using UDF with DataFrame

3.1 Using UDF with PySpark DataFrame select()

Now you can use `convertUDF()` on a DataFrame column as a regular build-in function.

```
df.select(col("Seqno"), \
    convertUDF(col("Name")).ali
   .show(truncate=False)
```

This results below output.

```
+-----+-------------+
|Seqno|Name         |
+-----+-------------+
|1    |John Jones   |
|2    |Tracey Smith |
|3    |Amy Sanders  |
+-----+-------------+
```

3.2 Using UDF with PySpark DataFrame withColumn()

You could also use udf on DataFrame `withColumn()` function, to explain this I will create another `upperCase()` function which converts the input string to upper case.

```
def upperCase(str):
    return str.upper()
```

Let’s convert `upperCase()` python function to UDF and then use it with DataFrame `withColumn()`. Below

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)

example converts the values of “Name” column to upper case and creates a new column “Curated Name”

```
upperCaseUDF = udf(lambda z:uppe

df.withColumn("Cureated Name", (
  .show(truncate=False)
```

This yields below output.

```
+-----+-------------+----------
|Seqno|Name         |Cureated Na
+-----+-------------+----------
|1    |john jones   |JOHN JONES
|2    |tracey smith|TRACEY SMIT
|3    |amy sanders  |AMY SANDERS
+-----+-------------+----------
```

3.3 Registering PySpark UDF & use it on SQL

In order to use `convertCase()` function on PySpark SQL, you need to register the function with PySpark by using `spark.udf.register()`.

```
""" Using UDF on SQL """
spark.udf.register("convertUDF"
df.createOrReplaceTempView("NAM
spark.sql("select Seqno, conver
     .show(truncate=False)
```

This yields the same output as 3.1 example.

4. Creating UDF using annotation

In the previous sections, you have learned creating a UDF is a 2 step process, first, you need to create a Python function, second convert function to UDF using SQL `udf()`

- 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)

function, however, you can avoid these two steps and create it with just a single step by using annotations.

```
@udf(returnType=StringType())
def upperCase(str):
    return str.upper()

df.withColumn("Cureated Name", 
.show(truncate=False)
```

This results same output as section 3.2

5. Special Handling

5.1 Execution order

One thing to aware is in PySpark/Spark does not guarantee the order of evaluation of subexpressions meaning expressions are not guarantee to evaluated left-to-right or in any other fixed order. PySpark reorders the execution for query optimization and planning hence, AND, OR, WHERE and HAVING expression will have side effects.

So when you are designing and using UDF, you have to be very careful especially with null handling as these results runtime exceptions.

```
"""
No guarantee Name is not null w
If convertUDF(Name) like '%John
you will get runtime error
"""
spark.sql("select Seqno, conver
         "where Name is not nul
     .show(truncate=False)
```

5.2 Handling null check

UDF’s are error-prone when not designed carefully. for example, when you have a column that contains the value null on some records

```
""" null check """

columns = ["Seqno","Name"]
data = [("1", "john jones"),
    ("2", "tracey smith"),
    ("3", "amy sanders"),
    ('4',None)]

df2 = spark.createDataFrame(dat
df2.show(truncate=False)
df2.createOrReplaceTempView("NA

spark.sql("select convertUDF(Na
     .show(truncate=False)
```

Note that from the above snippet, record with “Seqno 4” has value “None” for “name” column. Since we are not handling null with UDF function, using this on DataFrame returns below error. Note that in Python None is considered null.

```
AttributeError: 'NoneType' obje

    at org.apache.spark.api.pyth
    at org.apache.spark.sql.exec
    at org.apache.spark.sql.exec
    at org.apache.spark.api.pyth
    at org.apache.spark.Interrup
    at scala.collection.Iterator
```

Below points to remember

- Its always best practice to check for null inside a UDF function rather than checking for null outside.
- In any case, if you can’t do a null check in UDF at lease use IF or CASE WHEN to check for null and call UDF conditionally.

```
spark.udf.register("_nullsafeUD

spark.sql("select _nullsafeUDF(
     .show(truncate=False)

spark.sql("select Seqno, _nulls
          " where Name is not n
     .show(truncate=False)
```

This executes successfully without errors as we are checking for null/none while registering UDF.

5.3 Performance concern using UDF

UDF's are a black box to PySpark hence it can't apply optimization and you will lose all the optimization PySpark does on Dataframe/Dataset. When possible you should use Spark SQL built-in functions (https://sparkbyexamples.com/spark/spark-sql-functions-understanding/) as these functions provide optimization. Consider creating UDF only when existing built-in SQL function doesn't have it.

6. Complete PySpark UDF Example

Below is complete UDF function example in Scala

```
import pyspark
from pyspark.sql import SparkSe
from pyspark.sql.functions impo
from pyspark.sql.types import S

spark = SparkSession.builder.ap

columns = ["Seqno","Name"]
data = [("1", "john jones"),
    ("2", "tracey smith"),
    ("3", "amy sanders")]

df = spark.createDataFrame(data

df.show(truncate=False)

def convertCase(str):
    resStr=""
    arr = str.split(" ")
    for x in arr:
       resStr= resStr + x[0:1].
    return resStr

""" Converting function to UDF
convertUDF = udf(lambda z: conv

df.select(col("Seqno"), \
    convertUDF(col("Name")).ali
.show(truncate=False)

def upperCase(str):
    return str.upper()

upperCaseUDF = udf(lambda z:upp

df.withColumn("Cureated Name",
.show(truncate=False)

""" Using UDF on SQL """
spark.udf.register("convertUDF"
df.createOrReplaceTempView("NAM
spark.sql("select Seqno, conver
     .show(truncate=False)

spark.sql("select Seqno, conver
          "where Name is not nu
     .show(truncate=False)

""" null check """

columns = ["Seqno","Name"]
data = [("1", "john jones"),
    ("2", "tracey smith"),
    ("3", "amy sanders"),
    ('4',None)]
```

```
df2 = spark.createDataFrame(dat
df2.show(truncate=False)
df2.createOrReplaceTempView("NA

spark.udf.register("_nullsafeUD

spark.sql("select _nullsafeUDF(
     .show(truncate=False)

spark.sql("select Seqno, _nulls
          " where Name is not n
     .show(truncate=False)
```

This example is also available at [Spark GitHub project (https://github.com/spark-examples/pyspark-examples/blob/master/pyspark-udf.py)](https://github.com/spark-examples/pyspark-examples/blob/master/pyspark-udf.py) for reference.

Conclusion

In this article, you have learned the following

- PySpark UDF is a User Defined Function that is used to create a reusable function in Spark.
- Once UDF created, that can be re-used on multiple DataFrames and SQL (after registering).
- The default type of the udf() is StringType.
- You need to handle nulls explicitly otherwise you will see side-effects.

Reference

- [https://docs.databricks.com/spark/latest/spark-sql/udf-python.html (https://docs.databricks.com/spark/latest/spark-sql/udf-python.html)](https://docs.databricks.com/spark/latest/spark-sql/udf-python.html)
- [http://spark.apache.org/docs/latest/api/python/_modules/pyspark/sql/udf.html (https://spark.apache.org/docs/latest/api/python/_modules/pyspark/sql/udf.html)](https://spark.apache.org/docs/latest/api/python/_modules/pyspark/sql/udf.html)

Share this:

(https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/?share=facebook&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/?share=reddit&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/?share=pinterest&nb=1) 3

(https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/?share=tumblr&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/?share=pocket&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/?share=linkedin&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/?share=twitter&nb=1)

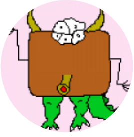
TAGS: PYSPARK UDF (HTTPS://SPARKBYEXAMPLES.COM/TAG/PYSPARK-UDF/), SPARK SQL UDF (HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARK-SQL-UDF/), UDF (HTTPS://SPARKBYEXAMPLES.COM/TAG/UDF/)

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

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/)

THIS POST HAS 4 COMMENTS



Anonymous

30 JAN 2021 REPLY

Why are you showing the whole example in Scala? This is PySpark… it should be in Python!

NNK **31 JAN 2021** REPLY

Hi, Complete example is in PySpark however, the Github link was pointing to Scala which I corrected now. Thanks for pointing it out.

Anonymous

7 OCT 2020 REPLY

Do you know to make a UDF globally, means can a notebook calls the UDF defined in another notebook?

Anonymous

5 DEC 2020 REPLY

If you use Zeppelin notebooks you can use the same interpreter in the several notebooks (change it in Intergpreter menu). Or search for precode option of Interpreter — in this optionn you can define any udf which will be created when the Interpreter started. You can setup the precode option in the same Interpreter menu

Leave a Reply

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/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/)

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

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

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

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://twitter.com/sparkbyexamples) (https://www.facebook.com/sparkbyexamples) (https://www.linkedin.com/in/n-nk-b860a8193/) (https://github.com/spark-examples/)