PySpark Where Filter Function | Multiple Conditions

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

PySpark `filter()` function is used to filter the rows from RDD/DataFrame based on the given condition or SQL expression, you can also use `where()` clause instead of the filter() if you are coming from an SQL background, both these functions operate exactly the same.

In this PySpark article, you will learn how to apply a filter on DataFrame columns of string, arrays, struct types by using single and multiple conditions and also applying filter using `isin()` with PySpark (Python Spark) examples.

Related Article:

- How to Filter Rows with NULL/NONE (IS NULL & IS NOT NULL) in PySpark (https://sparkbyexamples.com/pyspark/pyspark-filter-rows-with-null-values/)

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/)

- Spark Filter – startsWith(), endsWith() Examples (https://sparkbyexamples.com/spark/spark-filter-startswith-endswith-examples/)
- Spark Filter – contains(), like(), rlike() Examples (https://sparkbyexamples.com/spark/spark-filter-contains-like-rlike-examples/)

Note: PySpark Column Functions (https://sparkbyexamples.com/pyspark/pyspark-column-functions/) provides several options that can be used with filter().

1. PySpark DataFrame filter() Syntax

Below is syntax of the filter function. condition would be an expression you wanted to filter.

```
filter(condition)
```

Before we start with examples, first let's create a DataFrame (https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/). Here, I am using a DataFrame with StructType (https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/) and ArrayType (https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/) columns as I will also be covering examples with struct and array types as-well.

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

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

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

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

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

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

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

[PySpark – unionByName() (https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/)](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/)](https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/)

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

[PySpark – flatMap() (https://sparkbyexamples.com/pyspark/pyspark-flatmap-transformation/)](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)](https://sparkbyexamples.com/pyspark/pyspark-loop-iterate-through-rows-in-dataframe/#use-foreach-loop-through-dataframe)

[PySpark – sample() vs sampleBy()]

```
from pyspark.sql.types import S
from pyspark.sql.types import S
data = [
    (("James","","Smith"),["Jav
    (("Anna","Rose",""),["Spark
    (("Julia","","Williams"),["
    (("Maria","Anne","Jones"),[
    (("Jen","Mary","Brown"),["C
    (("Mike","Mary","Williams")
 ]

schema = StructType([
     StructField('name', Struct
        StructField('firstname'
        StructField('middlename
         StructField('lastname'
     ])),
     StructField('languages', A
     StructField('state', Strin
     StructField('gender', Stri
 ])

df = spark.createDataFrame(data
df.printSchema()
df.show(truncate=False)
```

This yields below schema and DataFrame results.

```
root
 |-- name: struct (nullable = t
 |    |-- firstname: string (nu
 |    |-- middlename: string (n
 |    |-- lastname: string (nul
 |-- languages: array (nullable
 |    |-- element: string (cont
 |-- state: string (nullable =
 |-- gender: string (nullable =

+----------------------+-------
|name                  |languag
+----------------------+-------
|[James, , Smith]      |[Java,
|[Anna, Rose, ]        |[Spark,
|[Julia, , Williams]   |[CSharp
|[Maria, Anne, Jones]  |[CSharp
|[Jen, Mary, Brown]    |[CSharp
|[Mike, Mary, Williams]|[Python
+----------------------+-------
```

(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/

2. DataFrame filter() with Column Condition

Use Column with the condition to filter the rows from DataFrame, using this you can express complex condition by referring column names using `dfObject.colname`

```
# Using equals condition
df.filter(df.state == "OH").sho

+----------------------+-------
|name                  |languag
+----------------------+-------
|[James, , Smith]      |[Java, 
|[Julia, , Williams]   |[CSharp
|[Mike, Mary, Williams]|[Python
+----------------------+-------

# not equals condition
df.filter(df.state != "OH") \
    .show(truncate=False)
df.filter(~(df.state == "OH"))
    .show(truncate=False)
```

Same example can also written as below. In order to use this first you need to import `from pyspark.sql.functions import col`

```
#Using SQL col() function
from pyspark.sql.functions impo
df.filter(col("state") == "OH")
    .show(truncate=False)
```

3. DataFrame filter() with SQL Expression

If you are coming from SQL background, you can use that knowledge in PySpark to filter DataFrame rows with SQL expressions.

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/)

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-

```
Jsing SQL Expression
f.filter("gender == 'M'").show
For not equal
f.filter("gender != 'M'").show
f.filter("gender <> 'M'").show
```

4. PySpark Filter with Multiple Conditions

In PySpark, to `filter()` rows on DataFrame based on multiple conditions, you case use either `Column` with a condition or SQL expression. Below is just a simple example using AND (&) condition, you can extend this with OR(|), and NOT(!) conditional expressions as needed.

```
//Filter multiple condition
df.filter( (df.state  == "OH")
    .show(truncate=False)
```

This yields below DataFrame results.

```
+----------------------+-------
|name                  |languag
+----------------------+-------
|[James, , Smith]      |[Java,
|[Mike, Mary, Williams]|[Python
+----------------------+-------
```

5. Filter Based on List Values

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())

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/)

If you have a list of elements and you wanted to filter that is not in the list or in the list, use `isin()` function of `Column` class (https://sparkbyexamples.com/pyspark/pyspark-column-functions/) and it doesn’t have isnotin() function but you do the same using not operator (~)

```
#Filter IS IN List values
li=["OH","CA","DE"]
df.filter(df.state.isin(li)).sh
+--------------------+---------
|                name|         
+--------------------+---------
|    [James, , Smith]|[Java, Sc
| [Julia, , Williams]|      [CS
|[Mike, Mary, Will...|      [Py
+--------------------+---------


# Filter NOT IS IN List values
#These show all records with NY
df.filter(~df.state.isin(li)).s
df.filter(df.state.isin(li)==Fa
```

6. Filter Based on Starts With, Ends With, Contains

You can also filter DataFrame rows by using `startswith()`, `endswith()` and `contains()` methods of Column class. For more examples on Column class, refer to PySpark Column Functions (https://sparkbyexamples.com/pyspark/pyspark-column-functions/).

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)

```
# Using startswith
df.filter(df.state.startswith("
+--------------------+---------
|                name|
+--------------------+---------
|      [Anna, Rose, ]|[Spark, J
|[Maria, Anne, Jones]|      [CS
|  [Jen, Mary, Brown]|      [CS
+--------------------+---------

#using endswith
df.filter(df.state.endswith("H"

#contains
df.filter(df.state.contains("H"
```

7. PySpark Filter like and rlike

If you have SQL background you must be familiar with `like` and `rlike` (regex like), PySpark also provides similar methods in Column class to filter similar values using wildcard characters. You can use rlike() to filter by checking values case insensitive.

```
data2 = [(2,"Michael Rose"),(3,
     (4,"Rames Rose"),(5,"Rames
  ]
df2 = spark.createDataFrame(dat

# like - SQL LIKE pattern
df2.filter(df2.name.like("%rose
+---+----------+
| id|      name|
+---+----------+
|  5|Rames rose|
+---+----------+

# rlike - SQL RLIKE pattern (LI
#This check case insensitive
df2.filter(df2.name.rlike("(?i)
+---+------------+
| id|        name|
+---+------------+
|  2|Michael Rose|
|  4|  Rames Rose|
|  5|  Rames rose|
```

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)

8. Filter on an Array column

When you want to filter rows from DataFrame based on value present in an array collection column, you can use the first syntax. The below example uses `array_contains()` from Pyspark SQL functions (https://sparkbyexamples.com/spark/spark-sql-functions/) which checks if a value contains in an array if present it returns true otherwise false.

```
from pyspark.sql.functions impo
df.filter(array_contains(df.lan
    .show(truncate=False)
```

This yields below DataFrame results.

```
+----------------+-------------
|name            |languages
+----------------+-------------
|[James, , Smith]|[Java, Scala,
|[Anna, Rose, ]  |[Spark, Java,
+----------------+-------------
```

9. Filtering on Nested Struct columns

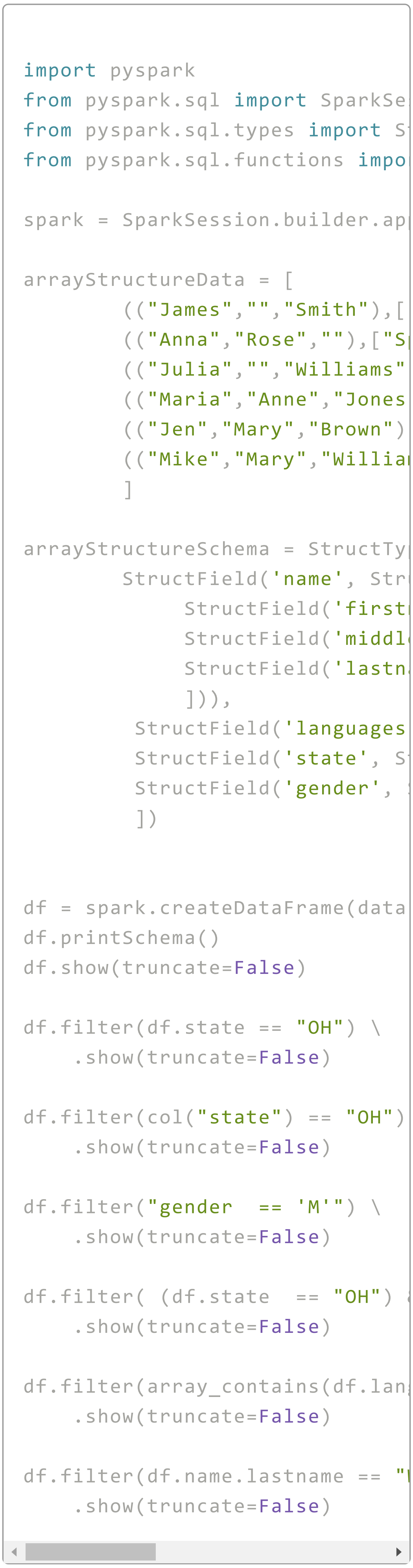
If your DataFrame consists of nested struct columns, you can use any of the above syntaxes to filter the rows based on the nested column.

```
  //Struct condition
df.filter(df.name.lastname == "
    .show(truncate=False)
```

This yields below DataFrame results

```
+----------------------+-------
|name                  |languag
+----------------------+-------
|[Julia, , Williams]   |[CSharp
|[Mike, Mary, Williams]|[Python
+----------------------+-------
```

10. Source code of PySpark where filter



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

spark = SparkSession.builder.ap

arrayStructureData = [
        (("James","","Smith"),[
        (("Anna","Rose",""),["S
        (("Julia","","Williams"
        (("Maria","Anne","Jones
        (("Jen","Mary","Brown")
        (("Mike","Mary","Willia
        ]

arrayStructureSchema = StructTy
        StructField('name', Str
             StructField('first
             StructField('middl
             StructField('lastn
             ])),
         StructField('languages
         StructField('state', S
         StructField('gender',
         ])


df = spark.createDataFrame(data
df.printSchema()
df.show(truncate=False)

df.filter(df.state == "OH") \
    .show(truncate=False)

df.filter(col("state") == "OH")
    .show(truncate=False)

df.filter("gender  == 'M'") \
    .show(truncate=False)

df.filter( (df.state  == "OH")
    .show(truncate=False)

df.filter(array_contains(df.lan
    .show(truncate=False)

df.filter(df.name.lastname == "
    .show(truncate=False)
```

Examples explained here are also available at PySpark examples GitHub (https://github.com/spark-examples/pyspark-examples/blob/master/pyspark-filter.py) project for reference.

11. Conclusion

In this tutorial, I’ve explained how to filter rows from PySpark DataFrame based on single or multiple conditions and SQL expression, also learned filtering rows by providing conditions on the array and struct column with Spark with Python examples.

Alternatively, you can also use `where()` function to filter the rows on PySpark DataFrame.

Happy Learning !!

Share this:

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

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

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

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

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

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

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

TAGS: FILTER() (HTTPS://SPARKBYEXAMPLES.COM/TAG/FILTER/), WHERE() (HTTPS://SPARKBYEXAMPLES.COM/TAG/WHERE/)

(https://sparkbyexamples.co

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 m/author/admin/)

environment Read more .. (https://sparkbyexamples.com/about-sparkbyexamples/)

THIS POST HAS ONE COMMENT

Anonymous

24 MAR 2021 REPLY

I am new to pyspark and this blog was extremely helpful to understand the concept. Thank you!!

Leave a Reply

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/)

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

ark-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/)

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/)

understand, and well tested in our development environment Read more .. (https://sparkbyexamples.com/about-sparkbyexamples/)

Follow Us

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