PREPARE TO LEAD

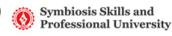
Spark by {Examplessom/n/)

PySpark Tutorial

PySpark (https://sparkbyexamples.

PySpark Tutorial For Beginners (https://sparkbyexamples.com/py spark-tutorial/)

tutorial/)



· Automobile Engineer

- Mechatronics Engine

B.Tech

- Construction Engine Managen
- Compute IT(Cyber

Admission

APPLY I

PySpark - Features

(https://sparkbyexamples.com/py

spark-tutorial/#features)

Hive (https://sparkbyexamples.com/apachenthe world of

PySpark - Advantages

(https://sparkbyexamples.com/py spark-tutorial/#advantages)

PySpark - Modules & Packages

packages)

PySpark - Cluster Managers

(https://sparkbyexamples.com/py spark-tutorial/#cluster-manager)

PySpark - Install on Windows

(https://sparkbyexamples.com/py spark-tutorial/#pyspark-

installation)

(https://sparkbyexamples.com/py HBase phttps://sparkbyexamples.com/apache-hbase-spark-tutorial/#modules-**String to Array**

tutorial/) Column

Kafka (https://sparkbyexamples.com/apache-kafka-(https://sparkbyexamples.com/author/admin/) -

PySpark

tutorials-withsexamples/amples.com/category/pyspark/)

PySpark - Web/Application UI (https://sparkbyexamples.com/sp

PySpark - SparkSession spark/pyspark-what-issparksession/)

PySpark - RDD

(https://sparkbyexamples.com/py spark-rdd)

PySpark - Parallelize (https://sparkbyexamples.com/py spark/pyspark-parallelize-create-

PySpark - repartition() vs coalesce()

(https://sparkbyexamples.com/py spark/pyspark-repartition-vscoalesce/)

PySpark - Broadcast Variables (https://sparkbyexamples.com/py spark/pyspark-broadcastvariables/)

PySpark - Accumulator (https://sparkbyexamples.com/py spark/pyspark-accumulator-withexample/)

PySpark DataFrame

PySpark - Create a DataFrame (https://sparkbyexamples.com/py

ark/spark-web-ui-understanding/) FAQ's P(yStpark/SQtarktoyvidesnsptts to o)mf/spction questions/) convert delimiter separated String to an Array (StringType to ArrayType) column (https://sparkbyexamples.com/py More on (batasidameanthasean belasheoby/) splitting a string column based on a delimiter like space, comma, pipe e.t.c, and converting it into ArrayType (https://sparkbyexamples.com/spark/sparkarray-arraytype-dataframe-column/).



In this article, I will explain converting String to Array column using split() function on DataFrame and SQL query.

Split() function syntax

<u>spark/different-ways-to-create-dataframe-in-pyspark/)</u>

<u>PySpark – Create an empty</u> <u>DataFrame</u>

(https://sparkbyexamples.com/py spark/pyspark-create-an-emptydataframe/)

<u>PySpark – Convert RDD to</u> <u>DataFrame</u>

(https://sparkbyexamples.com/py spark/convert-pyspark-rdd-todataframe/)

<u>PySpark – Convert DataFrame to</u> <u>Pandas</u>

(https://sparkbyexamples.com/py spark/convert-pyspark-dataframeto-pandas/)

<u>PySpark – show()</u>
(https://sparkbyexamples.com/py
spark/pyspark-show-displaydataframe-contents-in-table/)

<u>PySpark – StructType & StructField</u>

(https://sparkbyexamples.com/py spark/pyspark-structtype-andstructfield/)

<u>PySpark – Row Class</u> (https://sparkbyexamples.com/py spark/pyspark-row-using-rdddataframe/)

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

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

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

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

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

<u>PySpark – where() & filter()</u>
(https://sparkbyexamples.com/py
spark/pyspark-where-filter/)</u>

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

<u>PySpark – orderBy() and sort()</u> (https://sparkbyexamples.com/py PySpark SQL split() is grouped under Array Functions

(https://sparkbyexamples.com/spark/sparksql-array-functions/) in PySpark SQL Functions

(https://sparkbyexamples.com/spark/spark-sql-functions-understanding/)class with the below syntax.

The split() function takes the first argument as the DataFrame column of type String and the second argument string delimiter that you want to split on. You can also use the pattern as a delimiter. This function returns pyspark.sql.Column of type Array.

Before we start with usage, first, let's create a DataFrame with a string column with text separated with comma delimiter



This yields the below output. As you notice we have a name column with takens firstname, middle and lastname with comma separated.

<u>spark/pyspark-orderby-and-sort-explained/)</u>

<u>PySpark – groupBy()</u>
(https://sparkbyexamples.com/py
spark/pyspark-groupbyexplained-with-example/)

PySpark - join()

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

PySpark – union() & unionAll() (https://sparkbyexamples.com/py spark/pyspark-union-andunionall/)

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

PySpark – UDF (User Defined Function)

(https://sparkbyexamples.com/py spark/pyspark-udf-user-definedfunction/)

PySpark - map()

(https://sparkbyexamples.com/py spark/pyspark-maptransformation/)

<u>PySpark – flatMap()</u> (https://sparkbyexamples.com/py spark/pyspark-flatmaptransformation/)

pyspark - foreach()

(https://sparkbyexamples.com/py spark/pyspark-loop-iteratethrough-rows-in-dataframe/#useforeach-loop-through-dataframe)

PySpark - sample() vs sampleBy() (https://sparkbyexamples.com/py spark/pyspark-samplingexample/)

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/)

<u>PySpark – partitionBy()</u>
(https://sparkbyexamples.com/py
spark/pyspark-partitionbyexample/)</u>

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

```
root
  |-- name: string (nullable = true)
  |-- dob_year: string (nullable = t
  |-- gender: string (nullable = true)
  |-- salary: integer (nullable = fa
```

PySpark Convert String to Array Column

Below PySpark example snippet splits the String column name on comma delimiter and convert it to an Array. If you do not need the original column, use drop() to remove the column.

This yields below output. As you see below schema NameArray is a array type.

Convert String to Array

Column using SQL Query

Since PySpark provides a way to execute the raw SQL, let's learn how to write the same example using Spark SQL expression.

In order to use raw SQL, first, you need to create a table using createOrReplaceTempView(). This

PySpark – MapType (Map/Dict) (https://sparkbyexamples.com/py spark/pyspark-maptype-dictexamples/)

PySpark SQL Functions

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

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

<u>PySpark – Date and Timestamp</u> <u>Functions</u>

(https://sparkbyexamples.com/pysp ark/pyspark-sql-date-andtimestamp-functions/)

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

PySpark Datasources

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

<u>PySpark – Read & Write Parquet</u> File

(https://sparkbyexamples.com/py spark/pyspark-read-and-writeparquet-file/)

<u>PySpark – Read & Write JSON</u> <u>file</u>

(https://sparkbyexamples.com/py spark/pyspark-read-json-file-intodataframe/)



creates a temporary view from the Dataframe and this view is available lifetime of the current Spark context.

POPULAR

Refills F Persona Leather

If your diary over, worry r always buy a refill pages.

The Messy Corner

df.createOrReplaceTempView("PERSON'
spark.sql("select SPLIT(name,',') a
 .show()

This yields the same output as above example.

Complete Example

Below is the complete example of splitting an String type column based on a delimiter or patterns and converting into ArrayType column.

This example is also available at PySpark-Examples GitHub project
Pyspark-examples/blob/master/pyspark-string-to-array.py) for reference.



POPULAR

Refills F Persona Leather

If your diary over, worry r always buy a refill pages.

The Messy Corner



PySpark Built-In **Functions**

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

PySpark - expr() (https://sparkbyexamples.com/pysp ark/pyspark-sql-expr-expressionfunction/)

PySpark - lit()

(https://sparkbyexamples.com/pysp ark/pyspark-lit-add-literal-constant/)

PySpark - split()

(https://sparkbyexamples.com/pysp ark/pyspark-convert-string-to-arraycolumn/)

PySpark - concat_ws() (https://sparkbyexamples.com/pysp ark/pyspark-convert-array-columnto-string-column/)

Pyspark - substring() (https://sparkbyexamples.com/pysp ark/pyspark-substring-from-acolumn/)

PySpark - translate() (https://sparkbyexamples.com/pysp ark/pyspark-replace-columnvalues/#translate-replacecharacter-by-character)

PySpark - regexp replace() (https://sparkbyexamples.com/pysp ark/pyspark-replace-column-

```
spark = SparkSession.builder \
         .appName('SparkByExamples.
         .getOrCreate()
data = [("James, A, Smith","2018",
            ("Michael, Rose, Jones'
            ("Robert, K, Williams", "2
            ("Maria, Anne, Jones", "20
            ("Jen, Mary, Brown", "2016
columns=["name","dob year","gender'
df=spark.createDataFrame(data,colum
df.printSchema()
df.show(truncate=False)
from pyspark.sql.functions import
df2 = df.select(split(col("name"),
    .drop("name")
df2.printSchema()
df2.show()
df.createOrReplaceTempView("PERSON'
spark.sql("select SPLIT(name,',')
    .show()
```

Conclusion

In this simple article, you have learned how to Convert the string column into an array column by splitting the string by delimiter and also learned how to use the split function on PySpark SQL expression.

Happy Learning !!

Share this:

(https://sparkbyexamples.com/pyspark/pysparkconvert-string-to-array-column/? share=facebook&nb=1)

(https://sparkbyexamples.com/pyspark/pysparkconvert-string-to-array-column/? share=reddit&nb=1)

p (https://sparkbyexamples.com/pyspark/pysparkconvert-string-to-array-column/? share=pinterest&nb=1)

(https://sparkbyexamples.com/pyspark/pysparkconvert-string-to-array-column/? share=tumblr&nb=1)

(https://sparkbyexamples.com/pyspark/pysparkconvert-string-to-array-column/? share=pocket&nb=1)

in (https://sparkbyexamples.com/pyspark/pysparkconvert-string-to-array-column/? share=linkedin&nb=1)

<u>values/#regexp_replace-replace-string-columns)</u>

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

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

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

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

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()).

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

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

<u>PySpark – array()</u> (https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-withexamples/#array)</u>

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

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

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

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

<u>PySpark - map_values()</u> (https://sparkbyexamples.com/pysp_ (https://sparkbyexamples.com/pyspark/pysparkconvert-string-to-array-column/? share=twitter&nb=1)

TAGS: CONVERT STRING TO ARRAY
(HTTPS://SPARKBYEXAMPLES.COM/TAG/CONVERTSTRING-TO-ARRAY!), SPLIT().
(HTTPS://SPARKBYEXAMPLES.COM/TAG/SPLIT!)



<u>NNK</u>

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

(https://sp arkbyexa mples.co m/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...

ark/pyspark-maptype-dictexamples/#map_values)

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

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

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

<u>PySpark - row_number()</u>
(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row_number)

<u>PySpark - rank()</u>
(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#rank)

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

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

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

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

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

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

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

<u>PySpark – schema of json()</u> (https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema of json) Apache Hadoop

(https://sparkbyexamples.com/category/h adoop/)

Apache Spark

(https://sparkbyexamples.com/category/s park/)

Apache Spark Streaming

(https://sparkbyexamples.com/category/s park/apache-spark-streaming/)

Apache Kafka

(https://sparkbyexamples.com/category/k afka/)

Apache HBase

(https://sparkbyexamples.com/category/h base/)

Apache Cassandra

(https://sparkbyexamples.com/category/c assandra/)

Snowflake Database

(https://sparkbyexamples.com/category/s nowflake/)

H2O Sparkling Water

(https://sparkbyexamples.com/category/h 2o-sparkling-water/)

PySpark

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

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

(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/)

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

ples)



//www. //www.

(htt	ps:	facebo	linkedi	(https:
			n.com/	
r.co	om/	m/spar	in/n-	b.com/
spa	rkb	kbyex	nk-	spark-
yex	am	ample	b860a	examp

s/)

8193/)

les/)

Copyright sparkbyexamples.com