Spark by {Examples\$panktps://https://https://parkbyexamples.com/n/)

PySpark Tutorial PySpark _(https://sparkbyexamples.com/pyspark-**PySpark Tutorial For Beginners** (https://sparkbyexamples.com/py spark-tutorial/) tutorial/) PySpark - Features (https://sparkbyexamples.com/py Hive (https://sparkbyexamples.com/apache-hivespark-tutorial/#features) PySpark - Advantages (https://sparkbyexamples.com/py spark-tutorial/#advantages) PySpark - Modules & Packages (https://sparkbyexamples.com/py HBase phttps://sparkbyexamples.com/apache-hbase-spark-tutorial/#modulespackages) Column Values in tutorial/) DataFrame PySpark - Cluster Managers (https://sparkbyexamples.com/py spark-tutorial/#cluster-manager) Kafka (https://sparkbyexamples.com/apache-kafka-PySpark - Install on Windows (https://sparkbyexamples.com/author/admin/) -(https://sparkbyexamples.com/py spark-tutorial/#pyspark-PySpark installation) tutorials-with-examples/amples.com/category/pyspark/) PySpark - Web/Application UI (https://sparkbyexamples.com/sp ark/spark-web-ui-understanding/) FAQ's Youtcan/reptace gelwamplatues of RySparkuestions/) DataFrame by using SQL string functions PySpark - SparkSession regexp_replace(), translate(), and overlay() (https://sparkbyexamples.com/py More with Phythonsexamples.com/) spark/pyspark-what-issparksession/) PySpark - RDD (https://sparkbyexamples.com/py spark-rdd) PySpark - Parallelize (https://sparkbyexamples.com/py spark/pyspark-parallelize-createrdd/) PySpark - repartition() vs coalesce() (https://sparkbyexamples.com/py spark/pyspark-repartition-vs-In this article, I will cover examples of how coalesce/) to replace part of a string with another PySpark - Broadcast Variables

In this article, I will cover examples of how to replace part of a string with another string, replace all columns, change values conditionally, replace values from a python dictionary, replace column value from another DataFrame column e.t.c

First, let's create a PySpark DataFrame with some addresses and will use this DataFrame to explain how to replace column values.

PySpark DataFrame

<u>PySpark – Create a DataFrame</u> (https://sparkbyexamples.com/py

(https://sparkbyexamples.com/py

(https://sparkbyexamples.com/pyspark/pyspark-accumulator-with-

spark/pyspark-broadcast-

PySpark - Accumulator

variables/)

example/)

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

PySpark - Create an empty
DataFrame
(https://sparkbyexamples.com/py
spark/pyspark-create-an-emptydataframe/)

PySpark – Convert RDD to

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

PySpark – Convert DataFrame to Pandas (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>

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

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

<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> (<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

1. PySpark Replace String

Column Values

By using PySpark SQL function regexp_replace() you can replace a column value with a string for another string/substring. regexp_replace() uses Java regex for matching, if the regex does not match it returns an empty string, the below example replace the street name Rd value with Road string on address column.

2. Replace Column Values

Conditionally

In the above example, we just replaced Rd with Road, but not replaced St and Ave values, let's see how to replace column values conditionally in PySpark Dataframe by using when().otherwise() SQL condition function

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

spark/pyspark-orderby-and-sortexplained/)

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

<u>PySpark – join()</u>
(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>

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

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

<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/)



3. Replace Column Value with Dictionary (map)

You can also replace column values from the **python dictionary (map)**. In the below example, we replace the string value of the state column with the full abbreviated name from a dictionary **key-value pair**, in order to do so I use <u>PySpark map()</u> transformation to loop through each row of <u>DataFrame</u>

 $\underline{\text{(https://sparkbyexamples.com/pyspark/pysp}}\\ \underline{\text{ark-map-transformation/)}}.$

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

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>

<u>PySpark – Date and Timestamp</u> <u>Functions</u> (https://sparkbyexamples.com/pys

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

<u>PySpark – JSON Functions</u> (<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/)



PySpark Built-In Functions

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

<u>PySpark – expr()</u> (https://sparkbyexamples.com/pysp

4. Replace Column Value

Character by Character

By using translate() string function you can replace character by character of DataFrame column value. In the below example, every character of 1 is replaced with A, 2 replaced with B, and 3 replaced with C on the address column.

```
#Using translate to replace charact
from pyspark.sql.functions import t
df.withColumn('address', translate(
    .show(truncate=False)

#+---+
#|id |address | state|
#+---+
#|1 |A485A Jeffrey Rd |DE |
#|2 |4C4BA Margarita St|NY |
#|3 |ACAAA Siemon Ave | CA |
#+---+
```

i. Replace Column with

Inother Column Value

by using expr() and regexp_replace() you can replace column value with a value from another DataFrame column. In the below example, we match the value from col2 in col1 and replace with col3 to create new_column. Use expr() to provide SQL like expressions
(https://sparkbyexamples.com/pyspark/pysp

ark/pyspark-sql-expr-expressionfunction/)

PySpark - lit()

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

PySpark - split()

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

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

<u>Pyspark – substring()</u>
(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/pysp ark/pyspark-replace-columnvalues/#regexp_replace-replacestring-columns)

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

<u>PySpark - to_timestamp()</u>
(https://sparkbyexamples.com/spark
/pyspark-to_timestamp-convertstring-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/)

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> $\underline{\text{ark-sql-expr-expression-function/})}$ and is used to refer to another column to perform operations.

6. Replace All or Multiple

Column Values

If you want to replace values on all or selected DataFrame columns, refer to How to Replace NULL/None values on all column in PySpark

(https://sparkbyexamples.com/pyspark/pyspark-fillna-fill-replace-null-values/) or How to replace empty string with NULL/None value (https://sparkbyexamples.com/pyspark/pyspark-replace-empty-value-with-none-ondataframe/)

7. Using overlay() Function

Replace column value with a string value from another column.



| PRICE DROP | PRICE DROP |
|------------|------------|
| PRICE DROP | PRICE DROP |

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

PySpark - array()

(https://sparkbyexamples.com/pysp ark/pyspark-arraytype-column-withexamples/#array)

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

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

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

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

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

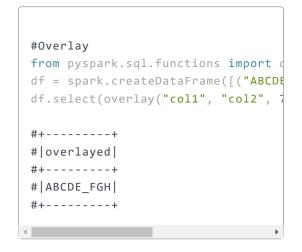
<u>PySpark - row_number()</u>
(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row_number)

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

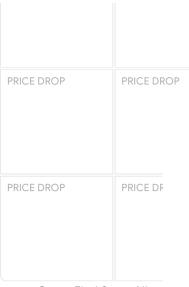
PySpark - dense_rank() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#dense_rank)

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

<u>PySpark – typedLit()</u> (https://sparkbyexamples.com/pysp



8. Complete Example



Come Find Some Nice Khankudi

ark/pyspark-lit-add-literalconstant/#typedlit)

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

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

<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)

```
from pyspark.sql import SparkSessic
spark = SparkSession.builder.master
address = [(1,"14851 Jeffrey Rd","[
    (2,"43421 Margarita St","NY"),
    (3,"13111 Siemon Ave","CA")]
df =spark.createDataFrame(address,
df.show()
#Replace string
from pyspark.sql.functions import
df.withColumn('address', regexp reg
  .show(truncate=False)
#Replace string
from pyspark.sql.functions import v
df.withColumn('address',
      when(df.address.endswith('Rd'
     .when(df.address.endswith('St
     .when(df.address.endswith('Ave
     .otherwise(df.address)) \
     .show(truncate=False)
#Replace values from Dictionary
stateDic={'CA':'California','NY':'N
df2=df.rdd.map(lambda x:
    (x.id,x.address,stateDic[x.stat
    ).toDF(["id","address","state"
df2.show()
#Using translate
from pyspark.sql.functions import
df.withColumn('address', translate(
  .show(truncate=False)
#Replace column with another column
from pyspark.sql.functions import
df = spark.createDataFrame([("ABCDE")
df.withColumn("new_column",
              expr("regexp replace(
              .alias("replaced_valu
              ).show()
#Overlay
from pyspark.sql.functions import
df = spark.createDataFrame([("ABCDE")
df.select(overlay("col1", "col2",
```

Conclusion

In conclusion regexp_replace() function is used to replace a string in a DataFrame column with another value, translate() function to replace character by character of column values, overlay() function to

overlay string with another column string from start position and number of characters. Finally, you have also learned how to replace column values from a dictionary using Python examples.

Happy Learning !!

References

 https://kb.databricks.com/data/nullempty-strings.html (https://kb.databricks.com/data/nullempty-strings.html)

Share this:

- (https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/?share=facebook&nb=1)
- (https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/?share=reddit&nb=1)
- φ (https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/?share=pinterest&nb=1)
- t (https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/?share=tumblr&nb=1)
- (https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/?share=pocket&nb=1)
- (https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/?share=linkedin&nb=1)
- (https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/?share=twitter&nb=1)

TAGS: OVERLAY()

(HTTPS://SPARKBYEXAMPLES.COM/TAG/OVERLAY/), REGEXP REPLACE()

(HTTPS://SPARKBYEXAMPLES.COM/TAG/REGEXP_REP LACE/)



NNK

(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...





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/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/p yspark/)

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/)

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/)

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

(https: (https:

//www. //www.

(https: facebo linkedi (https:

//twitte ok.co n.com/ //githu

r.com/ m/spar in/n- b.com/

sparkb kbyex nk- spark-

yexam ample b860a examp

ples) s/) 8193/) les/)

Copyright sparkbyexamples.com