-30% -50% -30%

shop.adidas.co.in

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 withColumn() Usage with Examples

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

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

PySpark withColumn() is a transformation function of DataFrame which is used to change the value, convert the datatype of an existing column, create a new column, and many more. In this post, I will walk you through commonly used PySpark DataFrame column operations using withColumn() examples.

Learn more

- PySpark withColumn – To change column DataType
- Transform/change value of an existing column

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/)

- Derive new column from an existing column
- Add a column with the literal value
- Rename column name
- Drop DataFrame column

First, let's create a DataFrame (https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/) to work with.

```
data = [('James','','Smith','19
  ('Michael','Rose','','2000-05
  ('Robert','','Williams','1978
  ('Maria','Anne','Jones','1967
  ('Jen','Mary','Brown','1980-0
]

columns = ["firstname","middlen
from pyspark.sql import SparkSe
spark = SparkSession.builder.ap
df = spark.createDataFrame(data
```

1. Change DataType using PySpark withColumn()

By using PySpark `withColumn()` on a DataFrame, we can cast or change the data type of a column. In order to change data type (https://sparkbyexamples.com/pyspark/pyspark-cast-column-type/), you would also need to use `cast()` function along with withColumn(). The below statement changes the datatype from `String` to `Integer` (https://sparkbyexamples.com/spark/spark-cast-string-type-to-integer-type-int/) for the `salary` column.

```
df.withColumn("salary",col("sal
```

2. Update The Value of an Existing Column

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/)

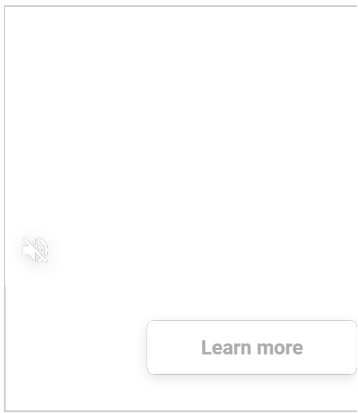
pyspark – foreach() (https://sparkbyexamples.com/pyspark/pyspark-loop-iterate-through-rows-in-dataframe/#use-foreach-loop-through-dataframe)

PySpark – sample() vs sampleBy()

PySpark `withColumn()` function of DataFrame can also be used to change the value of an existing column. In order to change the value, pass an existing column name as a first argument and a value to be assigned as a second argument to the withColumn() function. Note that the second argument should be `Column` type . Also, see Different Ways to Update PySpark DataFrame Column (https://sparkbyexamples.com/pyspark/pyspark-update-a-column-with-value/).

```
df.withColumn("salary",col("sal
```

This snippet multiplies the value of “salary” with 100 and updates the value back to “salary” column.



3. Create a Column from an Existing

To add/create a new column, specify the first argument with a name you want your new column to be and use the second argument to assign a value by applying an operation on an existing column. Also, see Different Ways to Add New Column to PySpark DataFrame (https://sparkbyexamples.com/pyspark/pyspark-add-new-column-to-dataframe/).

```
df.withColumn("CopiedColumn",co
```

(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/

This snippet creates a new column “CopiedColumn” by multiplying “salary” column with value -1.

4. Add a New Column using withColumn()

In order to create a new column, pass the column name you wanted to the first argument of `withColumn()` transformation function. Make sure this new column not already present on DataFrame, if it presents it updates the value of that column.

On below snippet, `PySpark lit()` function is used to add a constant value to a DataFrame column (https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/). We can also chain in order to add multiple columns.

```
df.withColumn("Country", lit("U
df.withColumn("Country", lit("U
  .withColumn("anotherColumn",l
  .show()
```

5. Rename Column Name

Though you cannot rename a column using withColumn, still I wanted to cover this as renaming is one of the common operations we perform on DataFrame. To rename an existing column use `withColumnRenamed()` (https://sparkbyexamples.com/pyspark/pyspark-rename-dataframe-column/) function on DataFrame.

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

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

Setu Nutrition
Setu Nutrition

Setu Lean Lite
₹479 ~~₹1,100~~
Checkout the latest products
OPEN

```
F.withColumnRenamed("gender","
 .show(truncate=False)
```

Drop Column From Spark DataFrame

"drop" function to drop a specific mn from the DataFrame os://sparkbyexamples.com/pyspark/ ark-drop-column-from-dataframe/).

```
F.drop("salary") \
 .show()
```

e: Note that all of these functions rn the new DataFrame after lying the functions instead of ating DataFrame.

7. PySpark withColumn() Complete Example

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-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)

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

spark = SparkSession.builder.ap

data = [('James','','Smith','19
  ('Michael','Rose','','2000-05
  ('Robert','','Williams','1978
  ('Maria','Anne','Jones','1967
  ('Jen','Mary','Brown','1980-0
]

columns = ["firstname","middlen
df = spark.createDataFrame(data
df.printSchema()
df.show(truncate=False)

df2 = df.withColumn("salary",co
df2.printSchema()
df2.show(truncate=False)

df3 = df.withColumn("salary",co
df3.printSchema()
df3.show(truncate=False)

df4 = df.withColumn("CopiedColu
df4.printSchema()

df5 = df.withColumn("Country",
df5.printSchema()

df6 = df.withColumn("Country",
   .withColumn("anotherColumn",
df6.printSchema()

df.withColumnRenamed("gender","
  .show(truncate=False)

df4.drop("CopiedColumn") \
.show(truncate=False)
```

The complete code can be downloaded from PySpark withColumn GitHub project (https://github.com/spark-examples/pyspark-examples/blob/master/pyspark-withcolumn.py)

Happy Learning !!

Share this:

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)

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

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

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

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

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

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

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

TAGS: WITHCOLUMN (HTTPS://SPARKBYEXAMPLES.COM/TAG/WITHCOLUMN/), WITHCOLUMNRENAMED (HTTPS://SPARKBYEXAMPLES.COM/TAG/WITHCOLUMNRENAMED/)

(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 3 COMMENTS

raghav 21 DEC 2020 REPLY

Hi
df2 = df.withColumn("salary",col("salary").cast("Integer"))
df2.printSchema()
I dont want to create a new dataframe if I am changing the datatype of existing dataframe.
Is there a way I can change column datatype in existing dataframe without creating a new dataframe ?

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)

NNK 25 DEC 2020 REPLY

DataFrames are immutable hence you cannot change anything directly on it. every operation on DataFrame results in a new DataFrame. If you want to change the DataFrame, I would recommend using the Schema at the time of creating the DataFrame.

Anonymous

16 SEP 2020 REPLY

Can you please explain Split column to multiple columns from Scala example into python

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/category/snowflake/)

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

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

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

Copyright sparkbyexamples.com