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 – Create DataFrame with Examples

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

You can manually create a **PySpark DataFrame** using `toDF()` and `createDataFrame()` methods, both these function takes different signatures in order to create DataFrame from existing RDD, list, and DataFrame.

You can also create PySpark DataFrame from data sources like TXT, CSV, JSON, ORV, Avro, Parquet, XML formats by reading from HDFS, S3, DBFS, Azure Blob file systems e.t.c.

Related:

- [Fetch More Than 20 Rows & Column Full Value in DataFrame (https://sparkbyexamples.com/spark/spark-fetch-more-than-20-rows-full-column-value/)]
- [Get Current Number of Partitions of Spark DataFrame (https://sparkbyexamples.com/spark/spark-get-current-number-of-partitions-of-dataframe/)]

[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/)

- [How to check if Column Present in Spark DataFrame (https://sparkbyexamples.com/spark/spark-check-column-present-in-dataframe/)](https://sparkbyexamples.com/spark/spark-check-column-present-in-dataframe/)

Finally, PySpark DataFrame also can be created by reading data from RDBMS Databases and NoSQL databases.

In this article, you will learn creating DataFrame by some of these methods with PySpark examples.

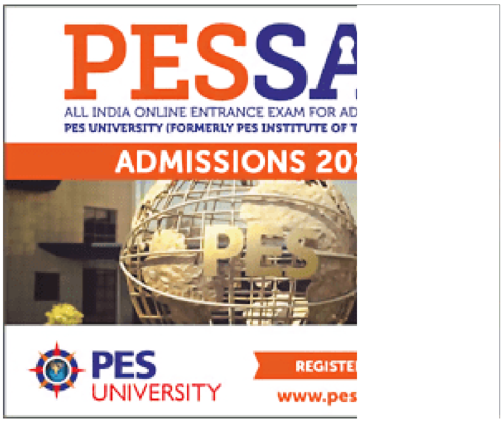


Table of Contents

SPARKSESSION	F
createDataFrame(rdd)	t
createDataFrame(dataList)	t
createDataFrame(rowData,columns)	
createDataFrame(dataList,schema)	

PySpark Create DataFrame matrix

In order to create a DataFrame from a list we need the data hence, first, let’s create the data and the columns that are needed.

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()

```
columns = ["language","users_co
data = [("Java", "20000"), ("Py
```

1. Create DataFrame from RDD

One easy way to manually create PySpark DataFrame is from an existing RDD. first, let’s create a Spark RDD (https://sparkbyexamples.com/spark/different-ways-to-create-spark-rdd/) from a collection List by calling parallelize() (https://sparkbyexamples.com/pyspark/pyspark-parallelize-create-rdd/) function from SparkContext (https://sparkbyexamples.com/spark/spark-sparkcontext/). We would need this **rdd** object for all our examples below.

```
spark = SparkSession.builder.ap
rdd = spark.sparkContext.parall
```

1.1 Using toDF() function

PySpark RDD’s toDF() method is used to create a DataFrame from existing RDD. Since RDD doesn’t have columns, the DataFrame is created with default column names “_1” and “_2” as we have two columns.

```
dfFromRDD1 = rdd.toDF()
dfFromRDD1.printSchema()
```

printschema() yields the below output.

```
root
 |-- _1: string (nullable = tru
 |-- _2: string (nullable = tru
```

If you wanted to provide column names to the DataFrame use `toDF()` method with column names as arguments as shown below.

(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/

```
columns = ["language","users_co
dfFromRDD1 = rdd.toDF(columns)
dfFromRDD1.printSchema()
```

This yields schema of the DataFrame with column names.

```
root
 |-- language: string (nullable
 |-- users: string (nullable = 
```

By default, the datatype of these columns infers to the type of data. We can change this behavior by supplying schema (https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/), where we can specify a column name, data type, and nullable for each field/column.

1.2 Using createDataFrame() from SparkSession

Using createDataFrame() from SparkSession (https://sparkbyexamples.com/tag/sparksession) is another way to create manually and it takes rdd object as an argument. and chain with toDF() to specify name to the columns.

```
dfFromRDD2 = spark.createDataFr
```

[pyspark/pyspark-read-and-write-parquet-file/)](pyspark/pyspark-read-and-write-parquet-file/)

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

PySpark Built-In Functions

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

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

2. Create DataFrame from List Collection

In this section, we will see how to create PySpark DataFrame from a list. These examples would be similar to t we have seen in the above section RDD, but we use the list data ct instead of "rdd" object to create aFrame.

Using eateDataFrame() from arkSession

ing `createDataFrame()` from `rkSession` is another way to ate PySpark DataFrame manually, it s a list object as an argument. and in with `toDF()` to specify names to columns.

```
fFromData2 = spark.createDataF
```

Using eateDataFrame() with the w type

`ateDataFrame()` has another ature in PySpark which takes the ection of Row type and schema for mn names as arguments. To use first we need to convert our "data" ct from the list to list of Row.

```
rowData = map(lambda x: Row(*x)
dfFromData3 = spark.createDataF
```

2.3 Create DataFrame with schema

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())

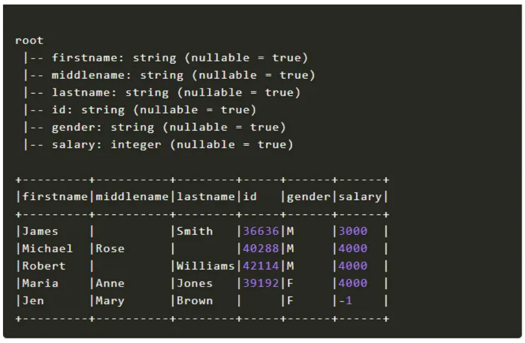
If you wanted to specify the column names along with their data types, you should create the StructType schema first and then assign this while creating a DataFrame.

```
from pyspark.sql.types import S
data2 = [("James","","Smith","3
    ("Michael","Rose","","40288
    ("Robert","","Williams","42
    ("Maria","Anne","Jones","39
    ("Jen","Mary","Brown","","F
  ]

schema = StructType([ \
    StructField("firstname",Str
    StructField("middlename",St
    StructField("lastname",Stri
    StructField("id", StringTyp
    StructField("gender", Strin
    StructField("salary", Integ
  ])

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

This yields below output.



```
root
 |-- firstname: string (nullable = true)
 |-- middlename: string (nullable = true)
 |-- lastname: string (nullable = true)
 |-- id: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- salary: integer (nullable = true)

+---------+----------+--------+-----+------+------+
|firstname|middlename|lastname|id   |gender|salary|
+---------+----------+--------+-----+------+------+
|James    |          |Smith   |36636|M     |3000  |
|Michael  |Rose      |        |40288|M     |4000  |
|Robert   |          |Williams|42114|M     |4000  |
|Maria    |Anne      |Jones   |39192|F     |4000  |
|Jen      |Mary      |Brown   |     |F     |-1    |
+---------+----------+--------+-----+------+------+
```

3. Create DataFrame from Data sources

In real-time mostly you create DataFrame from data source files like CSV, Text, JSON, XML e.t.c.

PySpark by default supports many data formats out of the box without importing any libraries and to create DataFrame you need to use the appropriate method available in `DataFrameReader` class.

- 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)

3.1 Creating DataFrame from CSV

Use `csv()` method of the `DataFrameReader` object to create a DataFrame from CSV file. you can also provide options like what delimiter to use, whether you have quoted data, date formats, infer schema, and many more. Please refer PySpark Read CSV into DataFrame (https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/)

```
df2 = spark.read.csv("/src/reso
```

3.2. Creating from text (TXT) file

Similarly you can also create a DataFrame by reading a from Text file, use `text()` method of the DataFrameReader to do so.

```
df2 = spark.read.text("/src/res
```

3.3. Creating from JSON file

PySpark is also used to process semi-structured data files like JSON format. you can use `json()` method of the DataFrameReader to read JSON file into DataFrame. Below is a simple example.

```
df2 = spark.read.json("/src/res
```

Similarly, we can create DataFrame in PySpark from most of the relational databases which I've not covered here and I will leave this to you to explore.

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)

4. Other sources (Avro, Parquet, ORC, Kafka)

We can also create DataFrame by reading Avro, Parquet, ORC, Binary files and accessing Hive and HBase table, and also reading data from Kafka which I've explained in the below articles, I would recommend reading these when you have time.

- PySpark Read Parquet file into DataFrame (https://sparkbyexamples.com/pyspark/pyspark-read-and-write-parquet-file/)
- DataFrame from Avro source (https://sparkbyexamples.com/spark/using-avro-data-files-from-spark-sql-2-4/)
- DataFrame by Streaming data from Kafka (https://sparkbyexamples.com/spark/spark-streaming-kafka-consumer-example-in-json-format/)

The complete code can be downloaded from GitHub (https://github.com/spark-examples/spark-scala-examples/blob/master/src/main/scala/com/sparkbyexamples/spark/dataframe/CreateDataFrame.scala)

Happy Learning !!

Share this:

(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/?share=facebook&nb=1)

(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/?share=reddit&nb=1)

(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/?share=pinterest&nb=1)

1

(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/?share=tumblr&nb=1)

(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/?share=pocket&nb=1)

(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/?share=linkedin&nb=1)

(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/?share=twitter&nb=1)

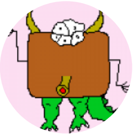
TAGS: DATAFRAME (HTTPS://SPARKBYEXAMPLES.COM/TAG/DATAFRAME/), SPARK.CREATEDATAFRAME (HTTPS://SPARKBYEXAMPLES.COM/TAG/CREATEDATAFRAME/), SPARKCONTEXT (HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARKCONTEXT/), SPARKCONTEXT.PARALLELIZE (HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARKCONTEXT-PARALLELIZE/), SPARKSESSION (HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARKSESSION/), TODF() (HTTPS://SPARKBYEXAMPLES.COM/TAG/TODF/)

(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/)

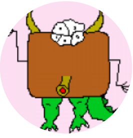
THIS POST HAS 3 COMMENTS



Anonymous

4 FEB 2021 REPLY

* indicates: its passing list as an argument



Anonymous

24 NOV 2020 REPLY

regular expression for arbitrary column names

sagar 30 JUL 2020 REPLY

What is significance of * in below dfFromData2 = spark.createDataFrame(data).toDF(*columns)

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/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/sp

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-nk-b860a8193/)

(https://github.com/spark-examples/)

ark-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/)