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/)

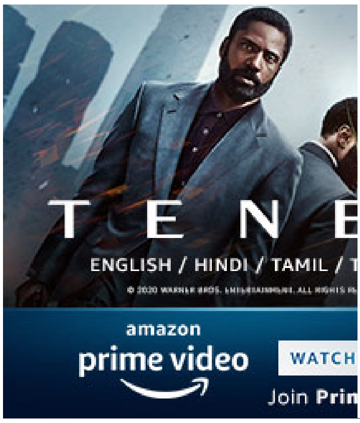
FAQ's (https://sparkbyexamples.com/spark-questions/)

More (https://sparkbyexamples.com/)

Spark Merge Two DataFrames with Different Columns or Schema

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

In Spark or PySpark let's see how to merge/unione two DataFrames with a different number of columns (different schema). In Spark 3.1, you can easily achieve this using unionByName() transformation by passing allowMissingColumns with the value true. In order version, this property is not available



Spark Tutorial

Spark – Installation on Windows (https://sparkbyexamples.com/spark/apache-spark-installation-on-windows/)

Spark – Installation on Linux | Ubuntu (https://sparkbyexamples.com/spark/spark-installation-on-linux-ubuntu/)

Spark – Cluster Setup with Hadoop Yarn (https://sparkbyexamples.com/spark/spark-setup-on-hadoop-yarn/)

Spark – Web/Application UI (https://sparkbyexamples.com/spark/spark-web-ui-understanding/)

Spark – Setup with Scala and IntelliJ (https://sparkbyexamples.com/spark/spark-setup-run-with-scala-intellij/)

Spark – How to Run Examples From this Site on IntelliJ IDEA (https://sparkbyexamples.com/spark/how-to-run-spark-examples-from-intellij/)

Spark – SparkSession (https://sparkbyexamples.com/spark/sparksession-explained-with-examples/)

Spark – SparkContext (https://sparkbyexamples.com/spark/spark-sparkcontext/)

Spark RDD Tutorial

Spark RDD – Parallelize (https://sparkbyexamples.com/apache-spark-rdd/how-to-create-an-rdd-using-parallelize/)

Spark RDD – Read text file (https://sparkbyexamples.com/apache-spark-rdd/spark-read-multiple-text-files-into-a-single-rdd/)

Spark RDD – Read CSV (https://sparkbyexamples.com/apache-spark-rdd/spark-load-csv-file-into-rdd/)

Spark RDD – Create RDD (https://sparkbyexamples.com/apache-spark-rdd/different-ways-to-create-spark-rdd/)

Spark RDD – Create Empty RDD (https://sparkbyexamples.com/apache-spark-rdd/spark-how-to-create-an-empty-rdd/)

Spark RDD – Transformations (https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-transformations/)

Spark RDD – Actions (https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-actions/)

Spark RDD – Pair Functions (https://sparkbyexamples.com/apache-spark-rdd/spark-pair-rdd-functions/)

Spark RDD – Repartition and Coalesce (https://sparkbyexamples.com/spark/spark-repartition-vs-coalesce/)

Spark RDD – Shuffle Partitions (https://sparkbyexamples.com/spark/spark-shuffle-partitions/)

Spark RDD – Cache vs Persist (https://sparkbyexamples.com/spark/spark-difference-between-cache-and-persist/)

Spark RDD – Persistance Storage Levels (https://sparkbyexamples.com/spark/spark-persistence-storage-levels/)

Spark RDD – Broadcast Variables (https://sparkbyexamples.com/spark/spark-broadcast-variables/)

Spark RDD – Accumulator Variables (https://sparkbyexamples.com/spark/spark-accumulators/)

Spark RDD – Convert RDD to DataFrame (https://sparkbyexamples.com/

```
//Scala
merged_df = df1.unionByName(df2

#PySpark
merged_df = df1.unionByName(df2
```

The difference between `unionByName()` function and `union()` is that this function resolves columns by name (not by position). In other words, unionByName() is used to merge two DataFrame's by column names instead of by position.

In case if you are using older than Spark 3.1 version, use below approach to merge DataFrame's with different column names.

- Spark Merge DataFrames with Different Columns (Scala Example)
- PySpark Merge DataFrames with Different Columns (Python Example)

Spark Merge Two DataFrames with Different Columns

In this section I will cover Spark with Scala example of how to merge two different DataFrames, first let's create DataFrames with different number of columns. DataFrame `df1` missing column `state` and `salary` and `df2` missing column `age`.

[apache-spark-rdd/convert-spark-rdd-to-dataframe-dataset/)]

Spark SQL Tutorial

[Spark SQL – Create DataFrame (https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/)]

[Spark SQL – Select Columns (https://sparkbyexamples.com/spark/spark-select-columns-from-dataframe/)]

[Spark SQL – Add and Update Column (withColumn) (https://sparkbyexamples.com/spark/spark-dataframe-withcolumn/)]

[Spark SQL – Rename Nested Column (https://sparkbyexamples.com/spark/rename-a-column-on-spark-dataframes/)]

[Spark SQL – Drop column (https://sparkbyexamples.com/spark/spark-drop-column-from-dataframe-dataset/)]

[Spark SQL – Where | Filter (https://sparkbyexamples.com/spark/spark-dataframe-where-filter/)]

[Spark SQL – When Otherwise (https://sparkbyexamples.com/spark/spark-case-when-otherwise-example/)]

[Spark SQL – Collect data to Driver (https://sparkbyexamples.com/spark/spark-dataframe-collect/)]

[Spark SQL – Distinct (https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/)]

[Spark SQL- Pivot Table DataFrame (https://sparkbyexamples.com/spark/how-to-pivot-table-and-unpivot-a-spark-dataframe/)]

[Spark SQL – Data Types (https://sparkbyexamples.com/spark/spark-sql-dataframe-data-types/)]

```
//Create DataFrame df1 with col
val data = Seq(("James","Sales"
              ("Robert","Sales
import spark.implicits._
val df1 = data.toDF("name","dep
df1.printSchema()

//root
// |-- name: string (nullable =
// |-- dept: string (nullable =
// |-- age: long (nullable = tr
```

Second DataFrame

```
//Create DataFrame df1 with col
val data2=Seq(("James","Sales",
             ("Jen","Finance",
val df2 = data2.toDF("name","de
df2.printSchema()

//root
// |-- name: string (nullable =
// |-- dept: string (nullable =
// |-- state: string (nullable
// |-- salary: long (nullable =
```

Now create a new DataFrames from existing after adding missing columns. newly added columns contains `null` values and we [add constant column using lit() function (https://sparkbyexamples.com/spark/using-lit-and-typedlit-to-add-a-literal-or-constant-to-spark-dataframe/)].

[Spark SQL – StructType | StructField (https://sparkbyexamples.com/spark/spark-sql-structtype-on-dataframe/)](https://sparkbyexamples.com/spark/spark-sql-structtype-on-dataframe/)

[Spark SQL – Schema (https://sparkbyexamples.com/spark/spark-schema-explained-with-examples/)](https://sparkbyexamples.com/spark/spark-schema-explained-with-examples/)

[Spark SQL – Groupby (https://sparkbyexamples.com/spark/using-groupby-on-dataframe/)](https://sparkbyexamples.com/spark/using-groupby-on-dataframe/)

[Spark SQL – Sort DataFrame (https://sparkbyexamples.com/spark/spark-how-to-sort-dataframe-column-explained/)](https://sparkbyexamples.com/spark/spark-how-to-sort-dataframe-column-explained/)

[Spark SQL – Join Types (https://sparkbyexamples.com/spark/spark-sql-dataframe-join/)](https://sparkbyexamples.com/spark/spark-sql-dataframe-join/)

[Spark SQL – Union and UnionAll (https://sparkbyexamples.com/spark/spark-dataframe-union-and-union-all/)](https://sparkbyexamples.com/spark/spark-dataframe-union-and-union-all/)

[Spark SQL – map() vs mapPartitions() (https://sparkbyexamples.com/spark/spark-map-vs-mappartitions-transformation/)](https://sparkbyexamples.com/spark/spark-map-vs-mappartitions-transformation/)

[Spark SQL – foreach() vs foreachPartition() (https://sparkbyexamples.com/spark/spark-foreachpartition-vs-foreach-explained/)](https://sparkbyexamples.com/spark/spark-foreachpartition-vs-foreach-explained/)

[Spark SQL – map() vs flatMap() (https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/)](https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/)

[Spark SQL – Persist and Cache (https://sparkbyexamples.com/spark/spark-dataframe-cache-and-persist-explained/)](https://sparkbyexamples.com/spark/spark-dataframe-cache-and-persist-explained/)

[Spark SQL – UDF (User Defined Functions) (https://sparkbyexamples.com/spark/spark-sql-udf/)](https://sparkbyexamples.com/spark/spark-sql-udf/)

[Spark SQL – Array (ArrayType) Column (https://sparkbyexamples.com/spark/spark-array-arraytype-dataframe-column/)](https://sparkbyexamples.com/spark/spark-array-arraytype-dataframe-column/)

```
val merged_cols = df1.columns.t
import org.apache.spark.sql.fun
def getNewColumns(column: Set[S
    merged_cols.toList.map(x =>
      case x if column.contains
      case _ => lit(null).as(x)
    })
}
val new_df1=df1.select(getNewCo
val new_df2=df2.select(getNewCo
```

Finally merge two DataFrame's by using column names

```
//Finally join two dataframe's 
val merged_df=new_df1.unionByNa
merged_df.show()

//+-------+---------+----+-----
//|   name|     dept| age|state
//+-------+---------+----+-----
//|  James|    Sales|  34| null
//|Michael|    Sales|  56| null
//| Robert|    Sales|  30| null
//|  Maria|  Finance|  24| null
//|  James|    Sales|null|   NY
//|  Maria|  Finance|null|   CA
//|    Jen|  Finance|null|   NY
//|   Jeff|Marketing|null|   CA
//+-------+---------+----+-----
```

PySpark Merge Two DataFrames with Different Columns

In PySpark to merge two DataFrames with different columns, will use the similar approach explain above and uses `unionByName()` transformation. First let's create DataFrame's with different number of columns.

[Spark SQL – Map (MapType) column (https://sparkbyexamples.com/spark/spark-dataframe-map-maptype-column/)](https://sparkbyexamples.com/spark/spark-dataframe-map-maptype-column/)

[Spark SQL – Flatten Nested Struct Column (https://sparkbyexamples.com/spark/spark-flatten-nested-struct-column/)](https://sparkbyexamples.com/spark/spark-flatten-nested-struct-column/)

[Spark SQL – Flatten Nested Array Column (https://sparkbyexamples.com/spark/spark-flatten-nested-array-column-to-single-column/)](https://sparkbyexamples.com/spark/spark-flatten-nested-array-column-to-single-column/)

[Spark SQL – Explode Array & Map Columns (https://sparkbyexamples.com/spark/explode-spark-array-and-map-dataframe-column/)](https://sparkbyexamples.com/spark/explode-spark-array-and-map-dataframe-column/)

[Spark SQL – Sampling (https://sparkbyexamples.com/spark/spark-sampling-with-examples/)](https://sparkbyexamples.com/spark/spark-sampling-with-examples/)

[Spark SQL – Partitioning (https://sparkbyexamples.com/spark/spark-partitioning-understanding/)](https://sparkbyexamples.com/spark/spark-partitioning-understanding/)

Spark SQL Functions

[Spark SQL String Functions (https://sparkbyexamples.com/spark/usage-of-spark-sql-string-functions/)](https://sparkbyexamples.com/spark/usage-of-spark-sql-string-functions/)

[Spark SQL Date and Timestamp Functions (https://sparkbyexamples.com/spark/spark-sql-date-and-time-functions/)](https://sparkbyexamples.com/spark/spark-sql-date-and-time-functions/)

[Spark SQL Array Functions (https://sparkbyexamples.com/spark/spark-sql-array-functions/)](https://sparkbyexamples.com/spark/spark-sql-array-functions/)

[Spark SQL Map Functions (https://sparkbyexamples.com/spark/spark-sql-map-functions/)](https://sparkbyexamples.com/spark/spark-sql-map-functions/)

[Spark SQL Sort Functions (https://sparkbyexamples.com/spark/spark-sql-sort-functions/)](https://sparkbyexamples.com/spark/spark-sql-sort-functions/)

[Spark SQL Aggregate Functions (https://sparkbyexamples.com/](https://sparkbyexamples.com/)

```
from pyspark.sql import SparkSe
spark = SparkSession.builder.ap

#Create DataFrame df1 with colu
data = [("James","Sales",34), (
    ("Robert","Sales",30), ("Ma
columns= ["name","dept","age"]
df1 = spark.createDataFrame(dat
df1.printSchema()

#Create DataFrame df1 with colu
data2=[("James","Sales","NY",90
    ("Jen","Finance","NY",7900)
columns2= ["name","dept","state
df2 = spark.createDataFrame(dat
df2.printSchema()
```

Now add missing columns ‘state‘ and ‘salary‘ to df1 and ‘age‘ to df2 with null values.

```
#Add missing columns 'state' &
from pyspark.sql.functions impo
for column in [column for colum
    df1 = df1.withColumn(column

#Add missing column 'age' to df
for column in [column for colum
    df2 = df2.withColumn(column
```

Now merge/union the DataFrames using `unionByName()`. The difference between `unionByName()` function and `union()` is that this function resolves columns by name (not by position). In other words, unionByName() is used to merge two DataFrame’s by column names instead of by position.

[spark/spark-sql-aggregate-functions/)](https://sparkbyexamples.com/spark/spark-sql-aggregate-functions/)

[Spark SQL Window Functions (https://sparkbyexamples.com/spark/spark-sql-window-functions/)](https://sparkbyexamples.com/spark/spark-sql-window-functions/)

[Spark SQL JSON Functions (https://sparkbyexamples.com/spark/spark-most-used-json-functions-with-examples/)](https://sparkbyexamples.com/spark/spark-most-used-json-functions-with-examples/)

Spark Data Source API

[Spark – Read & Write CSV file (https://sparkbyexamples.com/spark/spark-read-csv-file-into-dataframe/)](https://sparkbyexamples.com/spark/spark-read-csv-file-into-dataframe/)

[Spark – Read and Write JSON file (https://sparkbyexamples.com/spark/spark-read-and-write-json-file/)](https://sparkbyexamples.com/spark/spark-read-and-write-json-file/)

[Spark – Read & Write Parquet file (https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/)](https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/)

[Spark – Read & Write XML file (https://sparkbyexamples.com/spark/spark-read-write-xml/)](https://sparkbyexamples.com/spark/spark-read-write-xml/)

[Spark – Read & Write Avro files (https://sparkbyexamples.com/spark/read-write-avro-file-spark-dataframe/)](https://sparkbyexamples.com/spark/read-write-avro-file-spark-dataframe/)

[Spark – Read & Write Avro files (Spark version 2.3.x or earlier) (https://sparkbyexamples.com/spark/using-avro-data-files-from-spark-sql-2-3-x/)](https://sparkbyexamples.com/spark/using-avro-data-files-from-spark-sql-2-3-x/)

[Spark – Read & Write HBase using "hbase-spark" Connector (https://sparkbyexamples.com/spark/spark-read-write-using-hbase-spark-connector/)](https://sparkbyexamples.com/spark/spark-read-write-using-hbase-spark-connector/)

[Spark – Read & Write from HBase using Hortonworks (https://sparkbyexamples.com/spark/create-spark-dataframe-from-hbase-using-hortonworks/)](https://sparkbyexamples.com/spark/create-spark-dataframe-from-hbase-using-hortonworks/)

[Spark – Read & Write ORC file (https://sparkbyexamples.com/spark/spark-read-orc-file-into-dataframe/)](https://sparkbyexamples.com/spark/spark-read-orc-file-into-dataframe/)

```
#Finally join two dataframe's d
merged_df=df1.unionByName(df2)
merged_df.show()
```

Conclusion

In this article, you have learned with spark & PySpark examples of how to merge two DataFrames with different columns can be done by adding missing columns to the DataFrame's and finally union them using unionByName().

Happy Learning !!

Share this:

(https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/?share=facebook&nb=1)

(https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/?share=reddit&nb=1)

(https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/?share=pinterest&nb=1)

(https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/?share=tumblr&nb=1)

(https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/?share=pocket&nb=1)

(https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/?share=linkedin&nb=1)

(https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/?share=twitter&nb=1)

[(https://sparkbyexamples.com/author/admin/)](https://sparkbyexamples.com/author/admin/)

[NNK (Https://Sparkbyexamples.Com/Author/Admin/)](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/)](https://sparkbyexamples.com/about-sparkbyexamples/)

Leave a Reply

Spark – Read Binary File (https://sparkbyexamples.com/spark/spark-read-binary-file-into-dataframe/)

Spark Streaming & Kafka

Spark Streaming – OutputModes (https://sparkbyexamples.com/spark/spark-streaming-outputmode/)

Spark Streaming – Reading Files From Directory (https://sparkbyexamples.com/spark/spark-streaming-read-json-files-from-directory/)

Spark Streaming – Reading Data From TCP Socket (https://sparkbyexamples.com/spark/spark-streaming-from-tcp-socket/)

Spark Streaming – Processing Kafka Messages in JSON Format (https://sparkbyexamples.com/spark/spark-streaming-with-kafka/)

Spark Streaming – Processing Kafka messages in AVRO Format (https://sparkbyexamples.com/spark/spark-streaming-consume-and-produce-kafka-messages-in-avro-format/)

Spark SQL Batch – Consume & Produce Kafka Message (https://sparkbyexamples.com/spark/spark-batch-processing-produce-consume-kafka-topic/)

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 – 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/)

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() (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/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-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/)

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)

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/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-column-data-type-is-integer-or-string/)