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 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 SQL – How to Remove Duplicate Rows

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

Duplicate rows could be remove or drop from Spark SQL DataFrame using `distinct()` and `dropDuplicates()` functions, distinct() can be used to remove rows that have the same values on all columns whereas dropDuplicates() can be used to remove rows that have the same values on multiple selected columns.

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/)

Before we start, first let's create a DataFrame (https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/) with some duplicate rows and duplicate values on a few columns.

```
import spark.implicits._

val simpleData = Seq(("James", "
  ("Michael", "Sales", 4600),
  ("Robert", "Sales", 4100),
  ("Maria", "Finance", 3000),
  ("James", "Sales", 3000),
  ("Scott", "Finance", 3300),
  ("Jen", "Finance", 3900),
  ("Jeff", "Marketing", 3000),
  ("Kumar", "Marketing", 2000),
  ("Saif", "Sales", 4100)
)
val df = simpleData.toDF("employ
df.show()
```

Yields below output

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

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

[Spark RDD – Convert RDD to DataFrame (https://sparkbyexamples.com/apache-spark-rdd/convert-spark-rdd-to-dataframe-dataset/)](https://sparkbyexamples.com/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/)](https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/)

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

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

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

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

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

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

```
+-------------+----------+------
|employee_name|department|salary
+-------------+----------+------
|        James|     Sales|  3000
|      Michael|     Sales|  4600
|       Robert|     Sales|  4100
|        Maria|   Finance|  3000
|        James|     Sales|  3000
|        Scott|   Finance|  3300
|          Jen|   Finance|  3900
|         Jeff| Marketing|  3000
|        Kumar| Marketing|  2000
|         Saif|     Sales|  4100
+-------------+----------+------
```

On the above table, I've highlighted all duplicate rows, As you notice we have 2 rows that have duplicate values on all columns and we have 4 rows that have duplicate values on "department" and "salary" columns.

1. Use distinct() – Remove Duplicate Rows on DataFrame

On the above dataset, we have a total of 10 rows and one row with all values duplicated, performing distinct on this DataFrame should get us 9 as we have one duplicate.

- 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/)
- Spark SQL – StructType | StructField (https://sparkbyexamples.com/spark/spark-sql-structtype-on-dataframe/)
- Spark SQL – Schema (https://sparkbyexamples.com/spark/spark-schema-explained-with-examples/)
- Spark SQL – Groupby (https://sparkbyexamples.com/spark/using-groupby-on-dataframe/)
- Spark SQL – Sort DataFrame (https://sparkbyexamples.com/spark/spark-how-to-sort-dataframe-column-explained/)
- Spark SQL – Join Types (https://sparkbyexamples.com/spark/spark-sql-dataframe-join/)
- Spark SQL – Union and UnionAll (https://sparkbyexamples.com/spark/spark-dataframe-union-and-union-all/)
- Spark SQL – map() vs mapPartitions() (https://sparkbyexamples.com/spark/spark-map-vs-mappartitions-transformation/)

```
//Distinct all columns
val distinctDF = df.distinct()
println("Distinct count: "+disti
distinctDF.show(false)
```

`distinct()` function on DataFrame returns a new DataFrame after removing the duplicate records. This example yields the below output.

```
Distinct count: 9
+-------------+----------+------
|employee_name|department|salary
+-------------+----------+------
|James        |Sales     |3000
|Michael      |Sales     |4600
|Maria        |Finance   |3000
|Robert       |Sales     |4100
|Saif         |Sales     |4100
|Scott        |Finance   |3300
|Jeff         |Marketing |3000
|Jen          |Finance   |3900
|Kumar        |Marketing |2000
+-------------+----------+------
```

Alternatively, you can also run `dropDuplicates()` function which return a new DataFrame with duplicate rows removed.

```
val df2 = df.dropDuplicates()
println("Distinct count: "+df2.c
df2.show(false)
```

2. Use dropDuplicate() – Remove Duplicate Rows on DataFrame

[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/)

[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/)

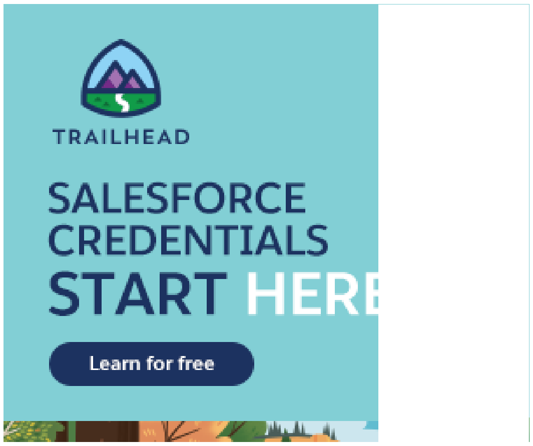
[Spark SQL – Sampling (https://sparkbyexamples.com/spark/spark-sampling-with-examples/)](https://sparkbyexamples.com/spark/spark-sampling-with-examples/)

Spark doesn't have a distinct method that takes columns that should run distinct on however, Spark provides another signature of `dropDuplicates()` function which takes multiple columns to eliminate duplicates.

Note that calling dropDuplicates() on DataFrame returns a new DataFrame with duplicate rows removed.

```
//Distinct using dropDuplicates
val dropDisDF = df.dropDuplicate
println("Distinct count of depar
dropDisDF.show(false)
```

Yields below output. If you notice the output, It dropped 2 records that are duplicate.



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

```
Distinct count of department & s
+-------------+----------+------
|employee_name|department|salary
+-------------+----------+------
|Jen          |Finance   |3900
|Maria        |Finance   |3000
|Scott        |Finance   |3300
|Michael      |Sales     |4600
|Kumar        |Marketing |2000
|Robert       |Sales     |4100
|James        |Sales     |3000
|Jeff         |Marketing |3000
+-------------+----------+------
```

3. Source code – Remove Duplicate Rows

spark/spark-read-and-write-json-file/)

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

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

Spark – Read & Write Avro files (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/)

Spark – Read & Write HBase 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/)

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

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/)

```
package com.sparkbyexamples.spar

import org.apache.spark.sql.Spar
import org.apache.spark.sql.func

object SQLDistinct extends App {

  val spark: SparkSession = Spar
    .master("local[1]")
    .appName("SparkByExamples.co
    .getOrCreate()

  spark.sparkContext.setLogLevel

  import spark.implicits._

  val simpleData = Seq(("James",
    ("Michael", "Sales", 4600),
    ("Robert", "Sales", 4100),
    ("Maria", "Finance", 3000),
    ("James", "Sales", 3000),
    ("Scott", "Finance", 3300),
    ("Jen", "Finance", 3900),
    ("Jeff", "Marketing", 3000),
    ("Kumar", "Marketing", 2000)
    ("Saif", "Sales", 4100)
  )
  val df = simpleData.toDF("empl
  df.show()

  //Distinct all columns
  val distinctDF = df.distinct()
  println("Distinct count: "+dis
  distinctDF.show(false)

  val df2 = df.dropDuplicates()
  println("Distinct count: "+df2
  df2.show(false)

  //Distinct using dropDuplicate
  val dropDisDF = df.dropDuplica
  println("Distinct count of dep
  dropDisDF.show(false)

}
```

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/)

The complete example is available at GitHub (https://github.com/spark-examples/spark-scala-examples/blob/master/src/main/scala/com/sparkbyexamples/spark/dataframe/functions/aggregate/DistinctCount.scala) for reference.

4. Conclusion

In this Spark article, you have learned how to remove DataFrame rows that are exact duplicates using `distinct()` and learned how to remove duplicate rows based on multiple columns using `dropDuplicate()` function with Scala example.

Share this:

(https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/?share=facebook&nb=1)

(https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/?share=reddit&nb=1)

(https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/?share=pinterest&nb=1)

(https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/?share=tumblr&nb=1)

(https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/?share=pocket&nb=1)

(https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/?share=linkedin&nb=1)

(https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/?share=twitter&nb=1)

TAGS: DISTINCT() (HTTPS://SPARKBYEXAMPLES.COM/TAG/DISTINCT/) , DROPDUPLICATES() (HTTPS://SPARKBYEXAMPLES.COM/TAG/DROPDUPLICATES/)

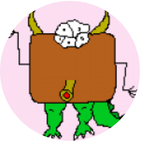
(https://sp

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

arkbyexamples.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/)

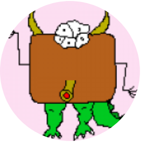
› THIS POST HAS 4 COMMENTS



Anonymous

10 APR 2021 REPLY

Awesome



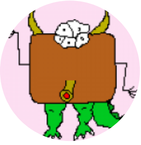
Anonymous

13 MAR 2021 REPLY

These tutorials are realy helpful, it would have been very good, if written in pyhton

NNK 13 MAR 2021

Hi, There are many articles writ Python as-well. Please refer to Remove duplicate rows in PySp (Spark with Python) (https://sparkbyexamples.com/p k/pyspark-distinct-to-drop-duplicates/)



Anonymous

8 MAR 2021 REPLY

Thank you, that helped.

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/categor

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

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

y/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/)

(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/)

(https://twitter.com/sparkbyexamples)

(https://www.facebook.com/sparkbyexamples/)

(https://www.linkedin.com/in/n-nk-b860a8193/)

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