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/)

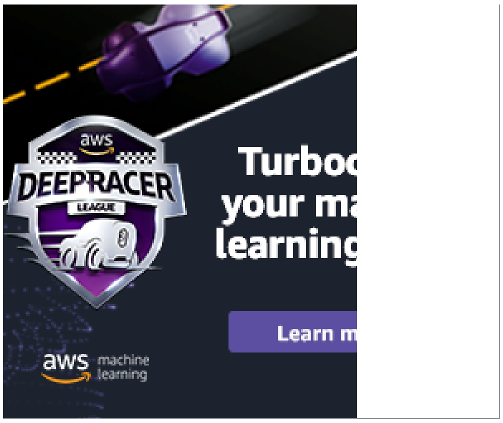
PySpark Tutorial

- 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 – Distinct to Drop Duplicate Rows

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

PySpark `distinct()` function is used to drop/remove the duplicate rows (all columns) from DataFrame and `dropDuplicates()` is used to drop rows based on selected (one or multiple) columns. In this article, you will learn how to use distinct() and dropDuplicates() functions with PySpark example.



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

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/)

values on a few columns. We use this DataFrame to demonstrate how to get distinct multiple columns.

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

data = [("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) \
  ]
columns= ["employee_name", "dep
df = spark.createDataFrame(data
df.printSchema()
df.show(truncate=False)
```

Yields below output

```
+-------------+----------+-----
|employee_name|department|salar
+-------------+----------+-----
|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, record with employer name `Robert` has 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.

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/)

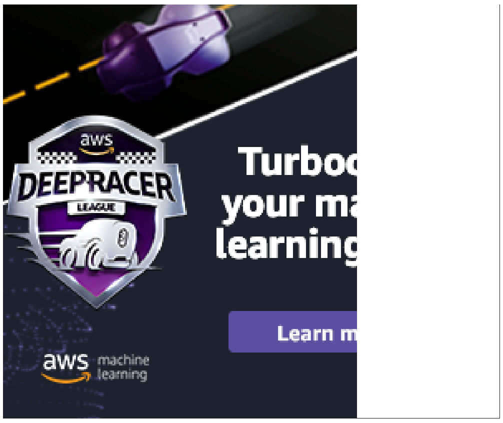
PySpark – flatMap() (https://sparkbyexamples.com/pyspark/pyspark-flatmap-transformation/)

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

PySpark – sample() vs sampleBy()

1. Get Distinct Rows (By Comparing All Columns)

On the above DataFrame, we have a total of 10 rows with 2 rows having all values duplicated, performing distinct on this DataFrame should get us 9 after removing 1 duplicate row.



```
distinctDF = df.distinct()
print("Distinct count: "+str(di
distinctDF.show(truncate=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|salar
+-------------+----------+-----
|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 returns a new DataFrame after removing duplicate rows.

(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/

```
df2 = df.dropDuplicates()
print("Distinct count: "+str(df
df2.show(truncate=False)
```

2. PySpark Distinct of Selected Multiple Columns

PySpark doesn’t have a distinct method which takes columns that should run distinct on (drop duplicate rows on selected multiple columns) however, it 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.

```
dropDisDF = df.dropDuplicates([
print("Distinct count of depart
dropDisDF.show(truncate=False)
```

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

[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/)

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

Source Code to Get Distinct Rows

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())

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

data = [("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) \
  ]
columns= ["employee_name", "dep
df = spark.createDataFrame(data
df.printSchema()
df.show(truncate=False)

#Distinct
distinctDF = df.distinct()
print("Distinct count: "+str(di
distinctDF.show(truncate=False)

#Drop duplicates
df2 = df.dropDuplicates()
print("Distinct count: "+str(df
df2.show(truncate=False)

#Drop duplicates on selected co
dropDisDF = df.dropDuplicates([
print("Distinct count of depart
dropDisDF.show(truncate=False)
}
```

The complete example is available at GitHub (https://github.com/spark-examples/pyspark-examples/blob/master/pyspark-distinct.py) for reference.

Conclusion

In this PySpark SQL article, you have learned `distinct()` method which is used to get the distinct values of rows (all columns) and also learned how to use `dropDuplicates()` to get the distinct and finally learned using dropDuplicates() function to get distinct of multiple columns.

[PySpark – explode() (https://sparkbyexamples.com/pyspark/pyspark-explode-nested-array-into-rows/)](https://sparkbyexamples.com/pyspark/pyspark-explode-nested-array-into-rows/)

[PySpark – array_contains() (https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#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)](https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array)

[PySpark – collect_list() (https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-list)](https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-list)

[PySpark – collect_set() (https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#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/)](https://sparkbyexamples.com/pyspark/pyspark-convert-dataframe-columns-to-maptype-dict/)

[PySpark – map_keys() (https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#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)](https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#map_values)

[PySpark – struct() (https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/#update-struct-function)](https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/#update-struct-function)

[PySpark – countDistinct() (https://sparkbyexamples.com/pyspark/pyspark-count-distinct-from-dataframe/)](https://sparkbyexamples.com/pyspark/pyspark-count-distinct-from-dataframe/)

[PySpark – sum(), avg() (https://sparkbyexamples.com/pyspark/pyspark-dataframe-groupby-and-sort-by-descending-order/)](https://sparkbyexamples.com/pyspark/pyspark-dataframe-groupby-and-sort-by-descending-order/)

[PySpark – row_number() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row_number)](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row_number)

[PySpark – rank() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#rank)](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#rank)

Happy Learning !!

Related Articles:

-

Share this:

(https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/?share=facebook&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/?share=reddit&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/?share=pinterest&nb=1)
2

(https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/?share=tumblr&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/?share=pocket&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/?share=linkedin&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/?share=twitter&nb=1)

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

(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 4 COMMENTS

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

[PySpark – typedLit() (https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/#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)](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)](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)](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)](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)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema_of_json)

abdulsattar

4 APR 2021 REPLY

bro please correct them above if there is no duplicate , i spent alot time on this and after all i came to see if there is someone else having same problem

NNK 4 APR 2021 REPLY

Hi Abdulsattar, I have updated the article when it was pointed out the first time. You should not see duplicate() function used anywhere. Could you please let me know where you are seeing, may be it's caching somewhere.

Sneha 31 OCT 2020 REPLY

I don't see duplicate() method used, is there a confusion between distinct() and duplicate() ? Please check.

NNK 1 NOV 2020 REPLY

Thanks Sneha. Yes, it should be distinct(), there is no duplicate but PySpark also has dropDuplicates().

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/catego

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/)

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

ry/snowflake/)

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

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

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/)