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

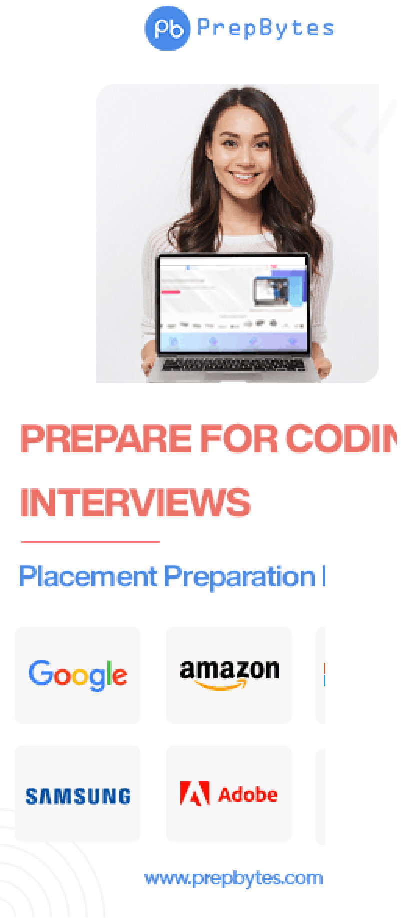
- PySpark Tutorial For Beginners (https://sparkbyexamples.com/pyspark-tutorial/)
- PySpark – Features (https://sparkbyexamples.com/pyspark-tutorial/#features)
- PySpark – Advantages (https://sparkbyexamples.com/pyspark-tutorial/#advantages)
- PySpark – Modules & Packages (https://sparkbyexamples.com/pyspark-tutorial/#modules-packages)
- 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 Join Types | Join Two DataFrames

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

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

PySpark Join is used to combine two DataFrames and by chaining these you can join multiple DataFrames; it supports all basic join type operations available in traditional SQL like `INNER`, `LEFT OUTER`, `RIGHT OUTER`, `LEFT ANTI`, `LEFT SEMI`, `CROSS`, `SELF JOIN`. PySpark Joins are wider transformations that involve data shuffling across the network (https://sparkbyexamples.com/spark/spark-shuffle-partitions/).



[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/)

PySpark SQL Joins comes with more optimization by default (thanks to DataFrames) however still there would be some performance issues to consider while using.

In this **PySpark SQL Join** tutorial, you will learn different Join syntaxes and using different Join types on two or more DataFrames and Datasets using examples.

1. PySpark Join Syntax

PySpark SQL join has a below syntax and it can be accessed directly from DataFrame.

```
join(self, other, on=None, how=
```

`join()` operation takes parameters as below and returns DataFrame.

- param other: Right side of the join
- param on: a string for the join column name
- param how: default `inner`. Must be one of `inner`, `cross`, `outer`,`full`,

[PySpark – withColumnRenamed() (https://sparkbyexamples.com/pyspark/pyspark-rename-dataframe-column/)](https://sparkbyexamples.com/pyspark/pyspark-rename-dataframe-column/)

[PySpark – where() & filter() (https://sparkbyexamples.com/pyspark/pyspark-where-filter/)](https://sparkbyexamples.com/pyspark/pyspark-where-filter/)

[PySpark – drop() & dropDuplicates() (https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/)](https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/)

[PySpark – orderBy() and sort() (https://sparkbyexamples.com/pyspark/pyspark-orderby-and-sort-explained/)](https://sparkbyexamples.com/pyspark/pyspark-orderby-and-sort-explained/)

[PySpark – groupBy() (https://sparkbyexamples.com/pyspark/pyspark-groupby-explained-with-example/)](https://sparkbyexamples.com/pyspark/pyspark-groupby-explained-with-example/)

[PySpark – join() (https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/)](https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/)

[PySpark – union() & unionAll() (https://sparkbyexamples.com/pyspark/pyspark-union-and-unionall/)](https://sparkbyexamples.com/pyspark/pyspark-union-and-unionall/)

[PySpark – unionByName() (https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/)](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/)](https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/)

[PySpark – map() (https://sparkbyexamples.com/pyspark/pyspark-map-transformation/)](https://sparkbyexamples.com/pyspark/pyspark-map-transformation/)

[PySpark – flatMap() (https://sparkbyexamples.com/pyspark/pyspark-flatmap-transformation/)](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)](https://sparkbyexamples.com/pyspark/pyspark-loop-iterate-through-rows-in-dataframe/#use-foreach-loop-through-dataframe)

PySpark – sample() vs sampleBy()

`full_outer`, `left`, `left_outer`, `right`, `right_outer`,`left_semi`, and `left_anti`.

You can also write Join expression by adding [where() (https://sparkbyexamples.com/pyspark/pyspark-dataframe-filter/)](https://sparkbyexamples.com/pyspark/pyspark-dataframe-filter/) and [filter() (https://sparkbyexamples.com/pyspark/pyspark-dataframe-filter/)](https://sparkbyexamples.com/pyspark/pyspark-dataframe-filter/) methods on DataFrame and can have Join on multiple columns.

2. PySpark Join Types

Below are the different Join Types PySpark supports.

Join String	Equivalent SQL Join
inner	INNER JOIN
outer, full, fullouter, full_outer	FULL OUTER JOIN
left, leftouter, left_outer	LEFT JOIN
right, rightouter, right_outer	RIGHT JOIN
cross	
anti, leftanti, left_anti	
semi, leftsemi, left_semi	

PySpark Join Types

Before we jump into PySpark SQL Join examples, first, let's create an `"emp"` and `"dept"` [DataFrame's (https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/)](https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/). here, column `"emp_id"` is unique on emp and `"dept_id"` is unique on the dept dataset's and emp_dept_id from emp has a reference to dept_id on dept dataset.

(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/

```
emp = [(1,"Smith",-1,"2018","10
    (2,"Rose",1,"2010","20","M"
    (3,"Williams",1,"2010","10"
    (4,"Jones",2,"2005","10","F
    (5,"Brown",2,"2010","40",""
      (6,"Brown",2,"2010","50",
  ]
empColumns = ["emp_id","name","
       "emp_dept_id","gender","

empDF = spark.createDataFrame(d
empDF.printSchema()
empDF.show(truncate=False)

dept = [("Finance",10), \
    ("Marketing",20), \
    ("Sales",30), \
    ("IT",40) \
  ]
deptColumns = ["dept_name","dep
deptDF = spark.createDataFrame(
deptDF.printSchema()
deptDF.show(truncate=False)
```

This prints “emp” and “dept” DataFrame to the console. Refer complete example below on how to create `spark` object.

```
Emp Dataset
+------+--------+--------------
|emp_id|name    |superior_emp_i
+------+--------+--------------
|1     |Smith   |-1
|2     |Rose    |1
|3     |Williams|1
|4     |Jones   |2
|5     |Brown   |2
|6     |Brown   |2
+------+--------+--------------

Dept Dataset
+---------+-------+
|dept_name|dept_id|
+---------+-------+
|Finance  |10     |
|Marketing|20     |
|Sales    |30     |
|IT       |40     |
+---------+-------+
```

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-

3. PySpark Inner Join DataFrame

`Inner` join is the default join in PySpark and it's mostly used. This joins two datasets on key columns, where keys don't match the rows get dropped from both datasets (`emp` & `dept`).

```
empDF.join(deptDF,empDF.emp_dep
     .show(truncate=False)
```

When we apply Inner join on our datasets, It drops "`emp_dept_id`" 50 from "`emp`" and "`dept_id`" 30 from "`dept`" datasets. Below is the result of the above Join expression.

```
+------+--------+--------------
|emp_id|name    |superior_emp_i
+------+--------+--------------
|1     |Smith   |-1
|2     |Rose    |1
|3     |Williams|1
|4     |Jones   |2
|5     |Brown   |2
+------+--------+--------------
```

4. PySpark Full Outer Join

`Outer` a.k.a `full`, `fullouter` join returns all rows from both datasets, where join expression doesn't match it

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)

returns null on respective record columns.

```
empDF.join(deptDF,empDF.emp_dep
    .show(truncate=False)
empDF.join(deptDF,empDF.emp_dep
    .show(truncate=False)
empDF.join(deptDF,empDF.emp_dep
    .show(truncate=False)
```

From our "emp" dataset's "emp_dept_id" with value 50 doesn't have a record on "dept" hence dept columns have null and "dept_id" 30 doesn't have a record in "emp" hence you see null's on emp columns. Below is the result of the above Join expression.

```
+------+--------+-------------
|emp_id|name    |superior_emp_i
+------+--------+-------------
|2     |Rose    |1
|5     |Brown   |2
|1     |Smith   |-1
|3     |Williams|1
|4     |Jones   |2
|6     |Brown   |2
|null  |null    |null
+------+--------+-------------
```

5. PySpark Left Outer Join

Left a.k.a Leftouter join returns all rows from the left dataset regardless of match found on the right dataset when join expression doesn't match, it assigns null for that record and drops records from right where match not found.

```
  empDF.join(deptDF,empDF("emp_
    .show(false)
  empDF.join(deptDF,empDF("emp_
    .show(false)
```

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)

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)

From our dataset, "emp_dept_id" 5o doesn't have a record on "dept" dataset hence, this record contains null on "dept" columns (dept_name & dept_id). and "dept_id" 30 from "dept" dataset dropped from the results. Below is the result of the above Join expression.

```
+------+--------+--------------
|emp_id|name    |superior_emp_i
+------+--------+--------------
|1     |Smith   |-1
|2     |Rose    |1
|3     |Williams|1
|4     |Jones   |2
|5     |Brown   |2
|6     |Brown   |2
+------+--------+--------------
```

6. Right Outer Join

`Right` a.k.a `Rightouter` join is opposite of `left` join, here it returns all rows from the right dataset regardless of math found on the left dataset, when join expression doesn't match, it assigns null for that record and drops records from left where match not found.

```
empDF.join(deptDF,empDF.emp_dep
   .show(truncate=False)
empDF.join(deptDF,empDF.emp_dep
   .show(truncate=False)
```

From our example, the right dataset "dept_id" 30 doesn't have it on the left dataset "emp" hence, this record contains null on "emp" columns. and "emp_dept_id" 50 dropped as a match not found on left. Below is the result of the above Join expression.

```
+------+--------+--------------
|emp_id|name    |superior_emp_i
+------+--------+--------------
|4     |Jones   |2
|3     |Williams|1
|1     |Smith   |-1
|2     |Rose    |1
|null  |null    |null
|5     |Brown   |2
+------+--------+--------------
```

7. Left Semi Join

`leftsemi` join is similar to `inner` join difference being `leftsemi` join returns all columns from the left dataset and ignores all columns from the right dataset. In other words, this join returns columns from the only left dataset for the records match in the right dataset on join expression, records not matched on join expression are ignored from both left and right datasets.

The same result can be achieved using select on the result of the inner join however, using this join would be efficient.

```
empDF.join(deptDF,empDF.emp_dep
   .show(truncate=False)
```

Below is the result of the above join expression.

```
leftsemi join
+------+--------+--------------
|emp_id|name    |superior_emp_i
+------+--------+--------------
|1     |Smith   |-1
|2     |Rose    |1
|3     |Williams|1
|4     |Jones   |2
|5     |Brown   |2
+------+--------+--------------
```

8. Left Anti Join

leftanti join does the exact opposite of the leftsemi, leftanti join returns only columns from the left dataset for non-matched records.

```
empDF.join(deptDF,empDF.emp_dep
   .show(truncate=False)
```

Yields below output

```
+------+-----+---------------+-
|emp_id|name |superior_emp_id|y
+------+-----+---------------+-
|6     |Brown|2              |2
+------+-----+---------------+-
```

9. PySpark Self Join

Joins are not complete without a self join, Though there is no self-join type available, we can use any of the above-explained join types to join DataFrame to itself. below example use inner self join.

```
empDF.alias("emp1").join(empDF.
    col("emp1.superior_emp_id")
    .select(col("emp1.emp_id"),
      col("emp2.emp_id").alias(
      col("emp2.name").alias("s
   .show(truncate=False)
```

Here, we are joining emp dataset with itself to find out superior emp_id and name for all employees.

```
+------+--------+-------------
|emp_id|name    |superior_emp_i
+------+--------+-------------
|2     |Rose    |1
|3     |Williams|1
|4     |Jones   |2
|5     |Brown   |2
|6     |Brown   |2
+------+--------+-------------
```

4. Using SQL Expression

Since PySpark SQL support native SQL syntax, we can also write join operations after creating temporary tables on DataFrame's and use these tables on `spark.sql()`.

```
empDF.createOrReplaceTempView("
deptDF.createOrReplaceTempView(

joinDF = spark.sql("select * fr
  .show(truncate=False)

joinDF2 = spark.sql("select * f
  .show(truncate=False)
```

5. PySpark SQL Join on multiple DataFrame's

When you need to join more than two tables, you either use SQL expression after creating a temporary view on the DataFrame or use the result of join operation to join with another DataFrame like chaining them. for example

```
df1.join(df2,df1.id1 == df2.id2
   .join(df3,df1.id1 == df3.id3
```

6. PySpark SQL Join Complete Example

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

spark = SparkSession.builder.ap

emp = [(1,"Smith",-1,"2018","10
    (2,"Rose",1,"2010","20","M"
    (3,"Williams",1,"2010","10"
    (4,"Jones",2,"2005","10","F
    (5,"Brown",2,"2010","40",""
      (6,"Brown",2,"2010","50",
  ]
empColumns = ["emp_id","name","
       "emp_dept_id","gender","

empDF = spark.createDataFrame(d
empDF.printSchema()
empDF.show(truncate=False)


dept = [("Finance",10), \
    ("Marketing",20), \
    ("Sales",30), \
    ("IT",40) \
  ]
deptColumns = ["dept_name","dep
deptDF = spark.createDataFrame(
deptDF.printSchema()
deptDF.show(truncate=False)

empDF.join(deptDF,empDF.emp_dep
     .show(truncate=False)

empDF.join(deptDF,empDF.emp_dep
    .show(truncate=False)
empDF.join(deptDF,empDF.emp_dep
    .show(truncate=False)
empDF.join(deptDF,empDF.emp_dep
    .show(truncate=False)

empDF.join(deptDF,empDF.emp_dep
    .show(truncate=False)
empDF.join(deptDF,empDF.emp_dep
   .show(truncate=False)

empDF.join(deptDF,empDF.emp_dep
   .show(truncate=False)
empDF.join(deptDF,empDF.emp_dep
   .show(truncate=False)

empDF.join(deptDF,empDF.emp_dep
   .show(truncate=False)

empDF.join(deptDF,empDF.emp_dep
    .show(truncate=False)

empDF.alias("emp1").join(empDF.
    col("emp1.superior_emp_id")
    .select(col("emp1.emp_id"),
      col("emp2.emp_id").alias(
      col("emp2.name").alias("s
   .show(truncate=False)

empDF.createOrReplaceTempView("
deptDF.createOrReplaceTempView(

joinDF = spark.sql("select * fr
  .show(truncate=False)

joinDF2 = spark.sql("select * f
  .show(truncate=False)
```

Examples explained here are available at the GitHub (https://github.com/spark-examples/pyspark-examples/blob/master/pyspark-join.py) project for reference.

Conclusion

In this PySpark SQL tutorial, you have learned two or more DataFrames can be joined using the `join()` function of the DataFrame, Join types syntax, usage, and examples with PySpark (Spark with Python), I would also recommend reading through Optimizing SQL Joins to know performance impact on joins.

Happy Learning !!

Share this:

(https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/?share=facebook&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/?share=reddit&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/?share=pinterest&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/?share=tumblr&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/?share=pocket&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/?share=linkedin&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/?share=twitter&nb=1)

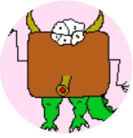
TAGS: CROSS JOIN (HTTPS://SPARKBYEXAMPLES.COM/TAG/CROSS-JOIN/), DATAFRAME JOIN (HTTPS://SPARKBYEXAMPLES.COM/TAG/DATAFRAME-JOIN/), INNER JOIN (HTTPS://SPARKBYEXAMPLES.COM/TAG/INNER-JOIN/), LEFT ANTI SEMI JOIN (HTTPS://SPARKBYEXAMPLES.COM/TAG/LEFT-ANTI-SEMI-JOIN/), LEFT JOIN (HTTPS://SPARKBYEXAMPLES.COM/TAG/LEFT-JOIN/), LEFT SEMI JOIN (HTTPS://SPARKBYEXAMPLES.COM/TAG/LEFT-SEMI-JOIN/), OUTER JOIN (HTTPS://SPARKBYEXAMPLES.COM/TAG/OUTER-JOIN/), RIGHT JOIN (HTTPS://SPARKBYEXAMPLES.COM/TAG/RIGHT-JOIN/), SQL JOIN (HTTPS://SPARKBYEXAMPLES.COM/TAG/SQL-JOIN/)

(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/)

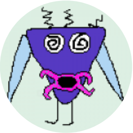
THIS POST HAS 3 COMMENTS



Anonymous

6 MAY 2021 REPLY

Surper content, really helped a lot !!!

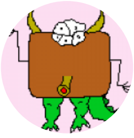


meri 7 MAR 2021 REPLY

there is no any 60 value. I think you meant to write 50 ☺

NNK 14 MAR 2021 REPLY

Thanks for pointing it out. I have corrected it now.



Anonymous

24 NOV 2020 REPLY

Very good job!!

Leave a Reply

Categories

Apache Hadoop (https://sparkbyexamples.com/category/hadoop/)

Apache Spark (https://sparkbyexamples.com/catego

Recent Posts

Spark regexp_replace() – Replace String Value (https://sparkbyexamples.com/spark/spark-regexp_replace-replace-string-value/)

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 ..

ry/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/)

How to Run a PySpark Script from Python? (https://sparkbyexamples.com/pyspark/run-pyspark-script-from-python-subprocess/)

Spark SQL like() Using Wildcard Example (https://sparkbyexamples.com/spark/spark-sql-like-using-wildcard-example/)

Spark isin() & IS NOT IN Operator Example (https://sparkbyexamples.com/spark/spark-isin-is-not-in-operator-example/)

Spark – Get Size/Length of Array & Map Column (https://sparkbyexamples.com/spark/spark-get-size-length-of-array-map-column/)

Spark Using Length/Size Of a DataFrame Column (https://sparkbyexamples.com/spark/spark-using-length-size-of-a-dataframe-column/)

Spark rlike() Working with Regex Matching Examples (https://sparkbyexamples.com/spark/spark-rlike-regex-matching-examples/)

Spark Check String Column Has Numeric Values (https://sparkbyexamples.com/spark/spark-check-string-column-has-numeric-values/)

Spark Check Column Data Type is Integer or String (https://sparkbyexamples.com/spark/spark-check-column-data-type-is-integer-or-string/)

(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/)