How to Create a Crosstab in PySpark with a Practical Example

Authored by

stats writer

February 2, 2026

Creating a crosstab in PySpark involves using the crosstab function from the pyspark.sql module. This function takes two columns of a Spark DataFrame and computes the frequency of their unique combinations. An example of creating a crosstab in PySpark would be as follows:

df = spark.createDataFrame(,)

df.crosstab("category", "value").show()

This will create a new DataFrame with the categories as rows, values as columns, and the frequency of their combinations as values. This function is useful for analyzing categorical data and identifying any patterns or relationships between the two variables.

Create a Crosstab in PySpark (With Example)

A crosstab is a table that summarizes the counts of two categorical variables.

You can use the following syntax to create a crosstab in PySpark:

df.crosstab(col1='team', col2='position').show()

This particular example creates a crosstab using the team column in the DataFrame along the rows and the position column along the columns of the crosstab.

The following example shows how to use this syntax in practice.

Example: How to Create a Crosstab in PySpark

Suppose we have the following PySpark DataFrame that contains information about the points scored by various basketball players:

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

#define data
data = ,
,
,
,
,
,
,
,
]

#define column names
columns =

#create dataframe using data and column names
```

```
df = spark.createDataFrame(data, columns)

#view dataframe
df.show()

+----+--------+------+
|team|position|points|
+----+--------+------+
| A| Guard| 11|
| A| Guard| 8|
| A| Forward| 22|
| A| Forward| 22|
| B| Guard| 14|
| B| Forward| 14|
| B| Forward| 13|
| B| Forward| 7|
| C| Forward| 11|
| C| Guard| 10|
+----+--------+------+
```


We can use the following syntax to create a crosstab using team as the rows and position as the columns:

```
#create crosstab using 'team' and 'points' columns
```

```
df.crosstab(col1='team', col2='position').show()

+-------------+-------+-----+
|team_position|Forward|Guard|
+-------------+-------+-----+
|            B|      3|    1|
|            C|      1|    1|
|            A|      2|    2|
+-------------+-------+-----+
```

The resulting crosstab shows the count of each team and position in the DataFrame.

For example, we can see:

There are 3 players who are Forwards on team B.There is 1 player who is a Guard on team B.There is 1 player who is a Forward on team C.There is 1 player who is a Guard on team C.There are 2 players who are Forwards on team A.There are 2 players who are Guards on team A.

Note: You can find the complete documentation for the PySpark crosstab function .

The following tutorials explain how to perform other common tasks in PySpark: