How to Select PySpark Columns Containing a Specific String

Authored by

stats writer

January 20, 2026

You can use the following syntax to select only columns that contain a specific string in a PySpark DataFrame:

df_new = df.select()

This particular example selects only the columns in the DataFrame that contain 'team' in their name.

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

Example: Select Columns Containing a Specific String in PySpark

Suppose we have the following PySpark DataFrame that contains information about 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_name|team_position|player_points|assists|
+---------+-------------+-------------+-------+
| A| Guard| 11| 4|

A	Forward	8	5
B	Guard	22	6
A	Forward	22	7
C	Guard	14	12
A	Guard	14	8
B	Forward	13	9
B	Center	7	9
+---------+-------------+-------------+-------+

We can use the following syntax to only select the columns that contain 'team' somewhere in their name:

#select columns that contain 'team' in the name
df_new = df.select()

#view new DataFrame
df_new.show()

+---------+-------------+
|team_name|team_position|
+---------+-------------+
A	Guard
A	Forward
B	Guard
A	Forward
C	Guard
A	Guard
B	Forward
B	Center
+---------+-------------+

The resulting DataFrame only contains the two columns that contain 'team' in the column name.

Note that if you'd like to select an additional column by name, you can use a plus sign (**+**) to do so.

For example, you can use the following syntax to select all columns with 'team' in their name along with the assists column:

#select columns that contain 'team' in the name and the 'assists' column

df_new = df.select(+)

#view new DataFrame
df_new.show()

```
+---------+-------------+-------+
|team_name|team_position|assists|
+---------+-------------+-------+
| A| Guard| 4|
| A| Forward| 5|
| B| Guard| 6|
| A| Forward| 7|
| C| Guard| 12|
| A| Guard| 8|
| B| Forward| 9|
| B| Center| 9|
+---------+-------------+-------+
```

The resulting DataFrame contains all columns with 'team' in their name along with the assists column.

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