How can I use PySpark to explode a nested array into rows?

Authored by

stats writer

June 24, 2024

PySpark is a Python-based framework used for large-scale data processing. It provides a convenient way to handle nested arrays in data by using the "explode" function. This function allows users to expand nested arrays into separate rows, making it easier to analyze and manipulate the data. By using PySpark's "explode" function, users can efficiently handle complex data structures and extract valuable insights from their datasets. This feature is particularly useful for data scientists and analysts working with large and diverse datasets.

Problem: How to explode & flatten nested array (Array of Array) DataFrame columns into rows using PySpark.

Solution: PySpark explode function can be used to explode an Array of Array (nested Array) `ArrayType(ArrayType(StringType))` columns to rows on PySpark DataFrame using python example.

Before we start, let's create a DataFrame with a nested array column. From below example column "subjects" is an array of ArraType which holds subjects learned.

```
import pyspark
from pyspark.sql import SparkSession
```

spark = SparkSession.builder.appName('pyspark-by-examples').getOrCreate()

arrayArrayData = ,]),
("Michael",,]),
("Robert",,])
]

df = spark.createDataFrame(data=arrayArrayData, schema =)
df.printSchema()
df.show(truncate=False)

df.printSchema() and df.show() returns the following schema and table.

```
root
|-- name: string (nullable = true)
|-- subjects: array (nullable = true)
| |-- element: array (containsNull = true)
| | |-- element: string (containsNull = true)
```

+-------+----------------------------------+
|name |subjects |

```
+-------+---------------------------------+
|James |, ]|
|Michael|, ]|
|Robert |, ] |
+-------+---------------------------------+
```

Now, let's explode "subjects" array column to array rows. after exploding, it creates a new column 'col' with rows represents an array.

```
from pyspark.sql.functions import explode
df.select(df.name,explode(df.subjects)).show(truncate=False)
```

Outputs:

```
+-------+------------------+
|name |col |
+-------+------------------+
|James ||
|James | |
|Michael||
|Michael| |
|Robert | |
|Robert | |
+-------+------------------+
```

If you want to flatten the arrays, use flatten function which converts array of array columns to a single array on DataFrame.

```
from pyspark.sql.functions import flatten
df.select(df.name,flatten(df.subjects)).show(truncate=False)
```

Outputs:

```
+-------+------------------------------+
|name |flatten(subjects) |
+-------+------------------------------+
|James ||
```

```
|Michael||
|Robert | |
+-------+------------------------------+
```

Happy Learning !!

Related Articles