How can an empty DataFrame and RDD be created in PySpark?

Authored by

stats writer

June 24, 2024

An empty DataFrame and RDD can be created in PySpark by using the "spark.createDataFrame()" and "spark.emptyRDD()" functions respectively. These functions allow for the creation of a DataFrame or RDD with no data, but with defined column names and data types. This can be useful for initializing a data structure before populating it with data, or for creating a placeholder for future data processing operations.

In this article, I will explain how to create an empty PySpark DataFrame/RDD manually with or without schema (column names) in different ways. Below I have explained one of the many scenarios where we need to create an empty DataFrame.

While working with files, sometimes we may not receive a file for processing, however, we still need to create a DataFrame manually with the same schema we expect. If we don't create with the same schema, our operations/transformations (like union's) on DataFrame fail as we refer to the columns that may not present.

To handle situations similar to these, we always need to create a DataFrame with the same schema, which means the same column names and datatypes regardless of the file exists or empty file processing.

1. Create Empty RDD in PySpark

Create an empty RDD by using `emptyRDD()` of SparkContext for example `spark.sparkContext.emptyRDD()`.

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()

#Creates Empty RDD
emptyRDD = spark.sparkContext.emptyRDD()
print(emptyRDD)

#Diplays
#EmptyRDD at emptyRDD
```

Alternatively you can also get empty RDD by using `spark.sparkContext.parallelize()`.

```
#Creates Empty RDD using parallelize
rdd2= spark.sparkContext.parallelize()
print(rdd2)
```

#EmptyRDD at emptyRDD at NativeMethodAccessorImpl.java:0
#ParallelCollectionRDD at readRDDFromFile at PythonRDD.scala:262

Note: If you try to perform operations on empty RDD you going to get `ValueError("RDD is empty")`.

2. Create Empty DataFrame with Schema (StructType)

In order to create an empty PySpark DataFrame manually with schema (column names & data types) first, Create a schema using StructType and StructField .

```
#Create Schema
from pyspark.sql.types import StructType,StructField, StringType
schema = StructType()
```

Now use the empty RDD created above and pass it to `createDataFrame()` of SparkSession along with the schema for column names & data types.

```
#Create empty DataFrame from empty RDD
df = spark.createDataFrame(emptyRDD,schema)
df.printSchema()
```

This yields below schema of the empty DataFrame.

```
root
|-- firstname: string (nullable = true)
|-- middlename: string (nullable = true)
|-- lastname: string (nullable = true)
```

3. Convert Empty RDD to DataFrame

You can also create empty DataFrame by converting empty RDD to DataFrame using `toDF()`.

```
#Convert empty RDD to Dataframe
df1 = emptyRDD.toDF(schema)
df1.printSchema()
```

4. Create Empty DataFrame with Schema.

So far I have covered creating an empty DataFrame from RDD, but here will create it manually with schema and without RDD.

```
#Create empty DataFrame directly.
df2 = spark.createDataFrame(, schema)
df2.printSchema()
```

5. Create Empty DataFrame without Schema (no columns)

To create empty DataFrame with out schema (no columns) just create a empty schema and use it while creating PySpark DataFrame.

```
#Create empty DatFrame with no schema (no columns)
df3 = spark.createDataFrame(, StructType())
df3.printSchema()
```

#print below empty schema
#root

Happy Learning !!

Related Articles