What is the usage of foreach() in PySpark and can you provide some examples?

Authored by

stats writer

June 24, 2024

RECOMMENDED CITATION

stats writer (2024). *What is the usage of foreach() in PySpark and can you provide some examples?*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=150813

ARABPSYCHOLOGY.COM

The "foreach()" function in PySpark is used to apply a specific action or operation to each element in a distributed collection, such as a DataFrame or RDD. It eliminates the need for explicit iteration and allows for efficient parallel processing. Some examples of using "foreach()" include printing each element in a DataFrame, saving data to an external database, or performing custom transformations on each element. Overall, "foreach()" is a useful tool for performing distributed operations on large datasets in PySpark.

PySpark `foreach()` is an action operation that is available in RDD, DataFram to iterate/loop over each element in the DataFrmae, It is similar to `for` with advanced concepts. This is different than other actions as `foreach()` function doesn't return a value instead it executes the input function on each element of an RDD, DataFrame

1. PySpark DataFrame foreach()

1.1 foreach() Syntax

Following is the syntax of the foreach() function

```
# Syntax
DataFrame.foreach(f)
```

1.2 PySpark foreach() Usage

When `foreach()` applied on PySpark DataFrame, it executes a function specified in for each element of DataFrame. This operation is mainly used if you wanted to manipulate accumulators, save the DataFrame results to RDBMS tables, Kafka topics, and other external sources.

In this example, to make it simple we just print the DataFrame to the console.

```
# Import
from pyspark.sql import SparkSession

# Create SparkSession
spark = SparkSession.builder.appName('SparkByExamples.com')
.getOrCreate()

# Prepare Data
columns =
data =
```

ARABPSYCHOLOGY.COM

```
# Create DataFrame
df = spark.createDataFrame(data=data,schema=columns)
df.show()

# foreach() Example
def f(df):
print(df.Seqno)
df.foreach(f)
```

Using foreach() to update the accumulator shared variable.

```
# foreach() with accumulator Example
accum=spark.sparkContext.accumulator(0)
df.foreach(lambda x:accum.add(int(x.Seqno)))
print(accum.value) #Accessed by driver
```

2. PySpark RDD foreach() Usage

The foreach() on RDD behaves similarly to DataFrame equivalent, hence the same syntax and it is also used to manipulate accumulators from RDD, and write external data sources.

2.1 Syntax

```
# Syntax
RDD.foreach(f: Callable, None]) → None
```

2.2 RDD foreach() Example

```
# foreach() with RDD example
accum=spark.sparkContext.accumulator(0)
rdd=spark.sparkContext.parallelize()
rdd.foreach(lambda x:accum.add(x))
print(accum.value) #Accessed by driver
```

Conclusion

In conclusion, PySpark foreach() is an action operation of RDD and DataFrame which doesn't have

ARABPSYCHOLOGY.COM

any return type and is used to manipulate the accumulator and write any external data sources.

Related Articles

Happy Learning !!

ARABPSYCHOLOGY.COM