06/02/2023

- Transformations are applied on a RDD to give another RDD
- while Actions are performed on a RDD to give a non-RDD value.

Action in Pyspark RDD:-

① collect() Action:-

Returns list of all elements of the RDD

```
collect_rdd = sc.parallelize([1,2,3,4,5])
print (collect_rdd.collect())
```

② count() Action:-

count() action of RDD is an ~~application~~ operation that returns the number of elements of our RDD.

```
count_rdd = sc.parallelize([1,2,3])
print (count_rdd.count())
```

[∵ o/p = 3]

③ The first() Action:-

It returns the first element of our RDD

```
first_rdd = sc.Parallelize([1,2,3])
print (first_rdd.first())
```

[∵ o/p = 1]

④ take() Action:-

take(n) action on RDD returns n no. of elements from RDD

```
take_rdd = sc.parallelize([1,2,3])
Print (take_rdd.take(2))
```

[∴ o/p = [1,2]]

⑤ reduce() Action:-

It takes two elements from the given RDD and operators

```
reduce_rdd = sc.parallelize([1,3,4,6])
print (reduce_rdd.reduce(lambda x,y: x+y))
```

[∴ o/p = 14]

⑥ Save As text file():-

It is used to save the resultant RDD as text file.

Transformation in Pyspark:-

① The map() Transformation:-

It maps a value to the element of an RDD

```
my-rdd = sc.parallelize([1,2,3,4])
Print (my-rdd.map(lambda x: x+10).collect())   [:. o/p:-[1,12,13,14]]
```

② The filter() Transformation:-

It filtering the elements from a Pyspark RDD

```
filter-rdd = sc.parallelize([2,3,4,6])
Print(filter-rdd.filter(lambda x: x%2==0).collect())   [:. o/p=[2,4,6]]
```

Renaming Column in Pandas Dataframe:-

METHOD-1:- Column Renamed()

SYNTAX:- Dataframe.withcolumnRenamed(existing, new)

CODE :- df.withcolumnRenamed("DOB", "Date of Birth").show()

METHOD-2:- Using selectExp()

syntax:- Dataframe.selectExpr(expression)

code :-
```
data = df.selectExpr("Name as name", "DOB", "Gender", "salary").
data.show()
```

METHOD-3:- Using Select() method.

Syntax:- Dataframe.select(cols)

code :-
```
data.select(col("Name"), col("DOB"), col("Gender"), col("salary").alias("Amount"
```

METHOD-4:- Using toDF()

Code Syntax :- Datalist = ["Emp Name", "Date of Birth", "Gender", "Salary"]

Code :-
```
new-df = df.toDF(*Data-list)
new-df.show()
```

Selecting, Ren | Renaming col | Inbox - saisac | ChatGPT | How to Integr | (1) WhatsApp | what is union

community.cloud.databricks.com/?o=1969337358064883#notebook/3867848309277915/command/3867848309277916

databricks

saisachin1616@gmail.com

Renaming columns in Dataframes

Python

File Edit View Run Help Last edit was yesterday New cell UI: OFF

Run all | Terminated | Share | Publish

```python
# Importing necessary libraries
from pyspark.sql import SparkSession

# Create a spark session
spark = SparkSession.builder.appName('pyspark - example join').getOrCreate()

# Create data in dataframe
data = [(('Ram'), '1991-04-01', 'M', 3000),
        (('Mike'), '2000-05-19', 'M', 4000),
        (('Rohini'), '1978-09-05', 'M', 4000),
        (('Maria'), '1967-12-01', 'F', 4000),
        (('Jenis'), '1980-02-17', 'F', 1200)]

# Column names in dataframe
columns = ["Name", "DOB", "Gender", "salary"]

# Create the spark dataframe
df = spark.createDataFrame(data=data,
                           schema=columns)

# Print the dataframe
df.show()
```

(3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [Name: string, DOB: string ... 2 more fields]

```
+------+----------+------+------+
|  Name|       DOB|Gender|salary|
+------+----------+------+------+
|   Ram|1991-04-01|     M|  3000|
|  Mike|2000-05-19|     M|  4000|
|Rohini|1978-09-05|     M|  4000|
| Maria|1967-12-01|     F|  4000|
| Jenis|1980-02-17|     F|  1200|
+------+----------+------+------+
```

91°F Sunny | Search | ENG IN | 17:40 07-02-2024

Renaming columns in Dataframes

Python

File Edit View Run Help Last edit was yesterday New cell UI: OFF

Run all | Terminated | Share | Publish

Command took 2.65 seconds -- by saisachin1616@gmail.com at 2/6/2024, 4:59:02 PM on My Cluster

Cmd 2

Python

```
1  #USING withColumnRenamed()
2
3  # Rename the column name from DOB to DateOfBirth
4  # Print the dataframe
5  df.withColumnRenamed("DOB","DateOfBirth").show()
```

(3) Spark Jobs

```
+------+-----------+------+------+
|  Name|DateOfBirth|Gender|salary|
+------+-----------+------+------+
|   Ram| 1991-04-01|     M|  3000|
|  Mike| 2000-05-19|     M|  4000|
|Rohini| 1978-09-05|     M|  4000|
| Maria| 1967-12-01|     F|  4000|
| Jenis| 1980-02-17|     F|  1200|
+------+-----------+------+------+
```

Command took 1.40 seconds -- by saisachin1616@gmail.com at 2/6/2024, 5:00:46 PM on My Cluster

Cmd 3

Renaming columns in Dataframes

Python

File Edit View Run Help Last edit was yesterday New cell UI: OFF

Run all | Terminated | Share | Publish

Command took 1.40 seconds -- by saisachin1616@gmail.com at 2/6/2024, 5:00:46 PM on My Cluster

Cmd 3

```
1  # Renaming multiple column names
2
3  # Rename the column name 'Gender' to 'Sex'
4  # Then for the returning dataframe
5  # again rename the 'salary' to 'Amount'
6  df.withColumnRenamed("Gender","Sex").withColumnRenamed("salary","Amount").show()
```

(3) Spark Jobs

```
+------+----------+---+------+
|  Name|       DOB|Sex|Amount|
+------+----------+---+------+
|   Ram|1991-04-01|  M|  3000|
|  Mike|2000-05-19|  M|  4000|
|Rohini|1978-09-05|  M|  4000|
| Maria|1967-12-01|  F|  4000|
| Jenis|1980-02-17|  F|  1200|
+------+----------+---+------+
```

Command took 0.85 seconds -- by saisachin1616@gmail.com at 2/6/2024, 5:01:55 PM on My Cluster

Cmd 4

Renaming columns in Dataframes

Python

File Edit View Run Help Last edit was yesterday New cell UI: OFF

Run all | Terminated | Share | Publish

```python
# USING selectExpr()

# Select the 'Name' as 'name'
# Select remaining with their original name
data = df.selectExpr("Name as name","DOB","Gender","salary")

# Print the dataframe
data.show()
```

(3) Spark Jobs

data: pyspark.sql.dataframe.DataFrame = [name: string, DOB: string ... 2 more fields]

```
+------+----------+------+------+
|  name|       DOB|Gender|salary|
+------+----------+------+------+
|   Ram|1991-04-01|     M|  3000|
|  Mike|2000-05-19|     M|  4000|
|Rohini|1978-09-05|     M|  4000|
| Maria|1967-12-01|     F|  4000|
| Jenis|1980-02-17|     F|  1200|
+------+----------+------+------+
```

Renaming columns in Dataframes

Python

File Edit View Run Help Last edit was yesterday New cell UI: OFF

Run all | Terminated | Share | Publish

```
1  # USING select() method
2
3  # Import col method from pyspark.sql.functions
4  from pyspark.sql.functions import col
5
6  # Select the 'salary' as 'Amount' using aliasing
7  # Select remaining with their original name
8  data = df.select(col("Name"),col("DOB"),col("Gender"),col("salary").alias('Amount'))
9
10 # Print the dataframe
11 data.show()
```

(3) Spark Jobs

data: pyspark.sql.dataframe.DataFrame = [Name: string, DOB: string ... 2 more fields]

```
+------+----------+------+------+
|  Name|       DOB|Gender|Amount|
+------+----------+------+------+
|   Ram|1991-04-01|     M|  3000|
|  Mike|2000-05-19|     M|  4000|
|Rohini|1978-09-05|     M|  4000|
| Maria|1967-12-01|     F|  4000|
| Jenis|1980-02-17|     F|  1200|
+------+----------+------+------+
```

Selecting, Ren | Renaming col | Inbox - saisac | ChatGPT | How to Integr | (1) WhatsApp | what is union

community.cloud.databricks.com/?o=1969337358064883#notebook/3867848309277915/command/3867848309277917

databricks

saisachin1616@gmail.com

Renaming columns in Dataframes

Python

File Edit View Run Help Last edit was yesterday New cell UI: OFF

Run all | Terminated | Share | Publish

Cmd 6

```
1  # USING toDF()
2
3  Data_list = ["Emp Name","Date of Birth"," Gender-m/f","Paid salary"]
4  new_df = df.toDF(*Data_list)
5  new_df.show()
```

(3) Spark Jobs

new_df: pyspark.sql.dataframe.DataFrame = [Emp Name: string, Date of Birth: string ... 2 more fields]

```
+--------+-------------+-----------+-----------+
|Emp Name|Date of Birth| Gender-m/f|Paid salary|
+--------+-------------+-----------+-----------+
|     Ram|   1991-04-01|          M|       3000|
|    Mike|   2000-05-19|          M|       4000|
|  Rohini|   1978-09-05|          M|       4000|
|   Maria|   1967-12-01|          F|       4000|
|   Jenis|   1980-02-17|          F|       1200|
+--------+-------------+-----------+-----------+
```

Command took 1.11 seconds -- by saisachin1616@gmail.com at 2/6/2024, 5:07:02 PM on My Cluster

Cmd 7

91°F Sunny | Search | ENG IN | 17:41 07-02-2024

Transformations in PySpark RDDs

Python

File Edit View Run Help Last edit was 6 minutes ago New cell UI: OFF

Run all | Demo | Share | Publish

```
1. The .map() Transformation
```

Cmd 2

```
my_rdd = sc.parallelize([1,2,3,4])
print(my_rdd.map(lambda x: x+ 10).collect())
```

(1) Spark Jobs

```
[11, 12, 13, 14]
```

Command took 8.56 seconds -- by saisachin1616@gmail.com at 2/7/2024, 5:34:06 PM on Demo

Cmd 3

```
2. The .filter() Transformation
```

Cmd 4

Python

```
filter_rdd = sc.parallelize([2, 3, 4, 5, 6, 7])
print(filter_rdd.filter(lambda x: x%2 == 0).collect())
```

(1) Spark Jobs

```
[2, 4, 6]
```

Transformations in PySpark RDDs

Python

File Edit View Run Help Last edit was 6 minutes ago New cell UI: OFF

Run all | Demo | Share | Publish

```
1  3. The .union() Transformation
```

Cmd 6

```
1  union_inp = sc.parallelize([2,4,5,6,7,8,9])
2  union_rdd_1 = union_inp.filter(lambda x: x % 2 == 0)
3  union_rdd_2 = union_inp.filter(lambda x: x % 3 == 0)
4  print(union_rdd_1.union(union_rdd_2).collect())
```

▸ (1) Spark Jobs

```
[2, 4, 6, 8, 6, 9]
```

Command took 0.77 seconds -- by saisachin1616@gmail.com at 2/7/2024, 5:35:22 PM on Demo

Cmd 7

```
1  4. The .flatMap() Transformation
```

Cmd 8

```
1  flatmap_rdd = sc.parallelize(["Hey there", "This is PySpark RDD Transformations"])
2  (flatmap_rdd.flatMap(lambda x: x.split(" ")).collect())
```

▸ (1) Spark Jobs

```
['Hey', 'there', 'This', 'is', 'PySpark', 'RDD', 'Transformations']
```

Command took 1.15 seconds -- by saisachin1616@gmail.com at 2/7/2024, 5:37:49 PM on Demo

Cmd 9

Selecting, Ren | Actions in PyS | Drafts (1) - sai | ChatGPT | How to Integr | WhatsApp | what is union

community.cloud.databricks.com/?o=1969337358064883#notebook/3347010216383462/command/3347010216383463

databricks

saisachin1616@gmail.com

Actions in PySpark RDDs

Python

File Edit View Run Help Last edit was 13 minutes ago New cell UI: OFF

Run all | Demo | Share | Publish

```
1  ACTIONS
```

Cmd 2

```
1  1. The .collect() Action
```

Cmd 3

```
1  collect_rdd = sc.parallelize([1,2,3,4,5])
2  print(collect_rdd.collect())
```

▸ (1) Spark Jobs

```
[1, 2, 3, 4, 5]
```

Command took 4.72 seconds -- by saisachin1616@gmail.com at 2/7/2024, 5:27:05 PM on Demo

Cmd 4

Python

```
1  2. The .count() Action
```

Cmd 5

```
1  from pyspark import SparkContext
2  sc = SparkContext.getOrCreate()
3  count_rdd = sc.parallelize([1,2,3,4,5,5,6,7,8,9])
4  print(count_rdd.count())
```

▸ (1) Spark Jobs

```
10
```

91°F Sunny | Search | ENG IN | 17:45 07-02-2024

Actions in PySpark RDDs

Python

File Edit View Run Help Last edit was 13 minutes ago New cell UI: OFF

Run all | Demo | Share | Publish

```
3. The .first() Action
```

Cmd 7

```
from pyspark import SparkContext
sc = SparkContext.getOrCreate()
count_rdd = sc.parallelize([1,2,3,4,5,5,6,7,8,9])
print(count_rdd.count())
first_rdd = sc.parallelize([1,2,3,4,5,6,7,8,9,10])
print(first_rdd.first())
```

(2) Spark Jobs

```
10
1
```

Command took 1.42 seconds -- by saisachin1616@gmail.com at 2/7/2024, 5:28:16 PM on Demo

Cmd 8

```
4. The .take() Action
```

Cmd 9

```
take_rdd = sc.parallelize([1,2,3,4,5])
print(take_rdd.take(3))
```

(2) Spark Jobs

```
[1, 2, 3]
```

Command took 0.66 seconds -- by saisachin1616@gmail.com at 2/7/2024, 5:28:46 PM on Demo

Selecting, Ren | Actions in PyS | Drafts (1) - sai | ChatGPT | How to Integr | WhatsApp | what is union

community.cloud.databricks.com/?o=1969337358064883#notebook/3347010216383462/command/3347010216383463

databricks

saisachin1616@gmail.com

Actions in PySpark RDDs

Python

File Edit View Run Help Last edit was 13 minutes ago New cell UI: OFF

Run all | Demo | Share | Publish

```
5. The .reduce() Action
```

Cmd 11

```
from pyspark import SparkContext
sc = SparkContext.getOrCreate()
reduce_rdd = sc.parallelize([1,3,4,6])
print(reduce_rdd.reduce(lambda x, y : x + y))
```

(1) Spark Jobs

14

Command took 0.61 seconds -- by saisachin1616@gmail.com at 2/7/2024, 5:29:11 PM on Demo

Cmd 12

```
6. The .saveAsTextFile() Action
```

Cmd 13

```
from pyspark import SparkContext
sc = SparkContext.getOrCreate()
save_rdd = sc.parallelize([1, 2, 3, 4, 5, 6])
save_rdd.saveAsTextFile('dbfs:/output_directory/file.txt')
sc.stop()
```

(1) Spark Jobs

The spark context has stopped and the driver is restarting. Your notebook will be automatically reattached.

Cmd 14

91°F Sunny | Search | ENG IN | 17:46 07-02-2024