

```
In [14]: 1 from pyspark.sql import SparkSession
2 spark = SparkSession.builder.appName('Filter_operations').getOrCreate()
3 spark
```

Out[14]: **SparkSession - in-memory**
SparkContext

[Spark UI \(http://DESKTOP-5NN8I6U:4041\)](http://DESKTOP-5NN8I6U:4041)

Version

v3.5.1

Master

local[*]

AppName

Filter_operations

```
In [15]: 1 df=spark.read.csv(r'C:\Users\Neha\Downloads\NARESH IT DATA SCIENCE\MAY\
◀────────────────────────────────────────────────────────────────────────────────▶▶
```

```
In [16]: 1 df.show()
```

```
+-----+-----+-----+
|   Name| Departments|salary|
+-----+-----+-----+
|   jack|Data Science| 10000|
|   alex|           ML|   5000|
|caroline|           AI|   4000|
|   jack|           AI|   4000|
|  sandra|Data Science|   3000|
|   jack|           NLP|  20000|
|   alex|           AI|  10000|
|   kathy|           ML|   5000|
|  andrew|Data Science|  10000|
|   jack|           NLP|   2000|
+-----+-----+-----+
```

```
In [17]: 1 df.printSchema()
```

```
root
 |-- Name: string (nullable = true)
 |-- Departments: string (nullable = true)
 |-- salary: integer (nullable = true)
```

```
In [24]: 1 df.groupby('Name')
```

Out[24]: GroupedData[grouping expressions: [Name], value: [Name: string, Department s: string ... 1 more field], type: GroupBy]

In [25]: 1 df.groupby('Name').show()

```
-----
-
AttributeError                                Traceback (most recent call last)
Cell In[25], line 1
----> 1 df.groupby('Name').show()

AttributeError: 'GroupedData' object has no attribute 'show'
```

In [22]: 1 df.groupby('Name').sum().show()

```
+-----+-----+
|   Name|sum(salary)|
+-----+-----+
|   jack|      36000|
| andrew|      10000|
| sandra|       3000|
|   alex|      15000|
|  kathy|       5000|
|caroline|       4000|
+-----+-----+
```

In [26]: 1 df.show()

```
+-----+-----+-----+
|   Name|Departments|salary|
+-----+-----+-----+
|   jack|Data Science| 10000|
|   alex|          ML|   5000|
|caroline|          AI|   4000|
|   jack|          AI|   4000|
| sandra|Data Science|   3000|
|   jack|          NLP|  20000|
|   alex|          AI|  10000|
|  kathy|          ML|   5000|
| andrew|Data Science|  10000|
|   jack|          NLP|   2000|
+-----+-----+-----+
```

In [27]: 1 df.groupby('Name').sum().show()

```
+-----+-----+
|   Name|sum(salary)|
+-----+-----+
|   jack|      36000|
| andrew|      10000|
| sandra|       3000|
|   alex|      15000|
|  kathy|       5000|
|caroline|       4000|
+-----+-----+
```

```
In [28]: 1 df.groupby('Name').avg().show()
```

```
+-----+-----+
|   Name|avg(salary)|
+-----+-----+
|   jack|    9000.0|
| andrew|   10000.0|
| sandra|    3000.0|
|   alex|    7500.0|
|  kathy|    5000.0|
|caroline|    4000.0|
+-----+-----+
```

```
In [30]: 1 df.groupby('Departments').sum().show()
```

```
+-----+-----+
|Departments|sum(salary)|
+-----+-----+
|         NLP|      22000|
|         AI|      18000|
|         ML|      10000|
|Data Science|      23000|
+-----+-----+
```

```
In [31]: 1 df.groupby('Departments').mean().show()
```

```
+-----+-----+
|Departments|    avg(salary)|
+-----+-----+
|         NLP|      11000.0|
|         AI|       6000.0|
|         ML|       5000.0|
|Data Science| 7666.666666666667|
+-----+-----+
```

```
In [32]: 1 df.groupby('Departments').min().show()
```

```
+-----+-----+
|Departments|min(salary)|
+-----+-----+
|         NLP|       2000|
|         AI|       4000|
|         ML|       5000|
|Data Science|       3000|
+-----+-----+
```

```
In [33]: 1 df.groupby('Departments').max().show()
```

```
+-----+-----+
| Departments|max(salary)|
+-----+-----+
|          NLP|      20000|
|           AI|      10000|
|           ML|       5000|
|Data Science|      10000|
+-----+-----+
```

```
In [34]: 1 df.groupby('Departments').count().show()
```

```
+-----+-----+
| Departments|count|
+-----+-----+
|          NLP|     2|
|           AI|     3|
|           ML|     2|
|Data Science|     3|
+-----+-----+
```

```
In [35]: 1 df.agg({'Salary': 'sum'}).show()
```

```
+-----+
|sum(Salary)|
+-----+
|       73000|
+-----+
```

```
In [36]: 1 df.agg({'Salary': 'min'}).show()
```

```
+-----+
|min(Salary)|
+-----+
|        2000|
+-----+
```

```
In [37]: 1 df.agg({'Salary': 'max'}).show()
```

```
+-----+
|max(Salary)|
+-----+
|       20000|
+-----+
```

```
In [38]: 1 df.agg({'Salary': 'count'}).show()
```

```
+-----+
|count(Salary)|
+-----+
|              10|
+-----+
```

```
In [39]: 1 df.agg({'Salary': 'median'}).show()
```

```
+-----+
|median(Salary)|
+-----+
|           5000.0|
+-----+
```

```
In [40]: 1 df.agg({'Salary': 'mode'}).show()
```

```
+-----+
|mode(Salary)|
+-----+
|          10000|
+-----+
```

IPYNP Folder/PySpark/1 MAY 24- Filter Operations in PySparkFilter_operations - Executors

Not secure | desktop-5nn86gu4041/executors/Welcome to Studen...New tabDellSearch for jobsWelcome to Studen...FacebookWord to PDF - Con...YouTubeScriptina | Font & T...GmailGooglePlagiarism Checker...

Executors

Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	68.9 KiB / 366.3 MiB	0.0 B	8	0	0	39	39	8.6 min (2 s)	5 KiB	3.3 KiB	3.3 KiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	68.9 KiB / 366.3 MiB	0.0 B	8	0	0	39	39	8.6 min (2 s)	5 KiB	3.3 KiB	3.3 KiB	0

Executors

Show 20 entries

Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump	Add Time	Remove Time
driver	DESKTOP-5NN86GU50112	Active	0	68.9 KiB / 366.3 MiB	0.0 B	8	0	0	39	39	8.6 min (2 s)	5 KiB	3.3 KiB	3.3 KiB	Thread Dump	2024-05-04 21:35:56	-

Showing 1 to 1 of 1 entries

Previous1Next

```
In [ ]: 1
```