BDA ASSIGNMENT 8

Name Devansh Thakkar
Srn no: 202201187
Roll no: 25
TY-A

Problem Statement: Filter rows and columns of a Spark Dataframe

CODE & OUTPUT:

1) **Step 1: Set up SparkSession**

First, start by importing the necessary libraries and creating a SparkSession.

Sudo su->pyspark

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

```
>>> spark = SparkSession.builder.appName("FilterEXample").getOrCreate()
24/11/14 10:00:52 WARN SparkSession: Using an existing Spark session; only runti
me SQL configurations will take effect.
```

2) Step 2: Create a DataFrame

```
>>> data = [
... ("Alice", 29, "Data Scientist"),
... ("Bob",34 ,"Software Engineer"),
... ("Cathy",30, "Analyst"),
... ("David",35,"Data Scientist")
... ]
>>> columns = ["Name", "Age", "Occupation"]
>>> df = spark.createDataFrame(data,columns)
>>> df.show
<bound method DataFrame.show of DataFrame[Name: string, Age: bigint, Occupation:
 string]>
>>> df.show()
+-----+---+----------------+
| Name|Age|      Occupation|
+-----+---+----------------+
|Alice| 29|  Data Scientist|
|  Bob| 34|Software Engineer|
|Cathy| 30|         Analyst|
|David| 35|  Data Scientist|
+-----+---+----------------+
```

3) **Step 3: Filter Rows**

To filter rows, use the filter method (or where as an alternative):

```
>>> filtered_df = df.filter(df.Age > 30)
>>> filtere_df.show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'filtere_df' is not defined. Did you mean: 'filtered_df'?
>>> filterde_df.show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'filterde_df' is not defined. Did you mean: 'filtered_df'?
>>> filtered_df.show()
+-----+---+----------------+
| Name|Age|      Occupation|
+-----+---+----------------+
|  Bob| 34|Software Engineer|
|David| 35|  Data Scientist|
+-----+---+----------------+
```

4) **Step 4: Select Columns**

To select specific columns, use the select method:

```
>>> selected_df=df.select("Name","Occupation")
>>> selected_df.show()
+-----+----------------+
| Name|      Occupation|
+-----+----------------+
|Alice|  Data Scientist|
|  Bob|Software Engineer|
|Cathy|         Analyst|
|David|  Data Scientist|
+-----+----------------+
```

5) **Step 5: Combine Row and Column Filtering**

You can combine both row filtering and column selection as follows:

```
>>> filteres_selected_df=df.filter(df.Age > 30).select("Name","Occupation")
>>> filteres_selected_df.show()
+-----+----------------+
| Name|      Occupation|
+-----+----------------+
|  Bob|Software Engineer|
|David|  Data Scientist|
+-----+----------------+
```

pip3 install pyspark

sudo apt update

sudo apt install python3-pip

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col

# Create a SparkSession
spark = SparkSession.builder \
    .appName("Filter Rows and Columns Example") \
    .master("local[*]") \
    .getOrCreate()

# Sample data
data = [
    ("Alice", "HR", 5000),
    ("Bob", "Finance", 6000),
    ("Charlie", "IT", 7000),
    ("David", "Finance", 4500),
    ("Eve", "HR", 5500),
    ("Frank", "IT", 7200)
]
columns = ["Name", "Department", "Salary"]

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

# Display the DataFrame
print("Original DataFrame:")
df.show()
```