**Name -Monika Gariya**

**Email-** [**monikagariya2023@gmail.com**](mailto:monikagariya2023@gmail.com)

**Data Engineering Batch 1**

**Date – 06-02-2024**

**Topic- Transforming Data with PySpark** **RDD’S, Selecting, renaming and filtering data in panda dataframe**

**PySpark RDD Operations**

Resilient Distributed Dataset or RDD in a PySpark is a core data structure of PySpark. PySpark RDD’s is a low-level object and are highly efficient in performing distributed tasks.

PySpark RDD has a set of operations to accomplish any task. These operations are of two types:

1. Transformations

2. Actions

**Transformations**are a kind of operation that takes an RDD as input and produces another RDD as output. Once a transformation is applied to an RDD, it returns a new RDD, the original RDD remains the same and thus are immutable. After applying the transformation, it creates a Directed Acyclic Graph or DAG for computations and ends after applying any actions on it. This is the reason they are called lazy evaluation processes.

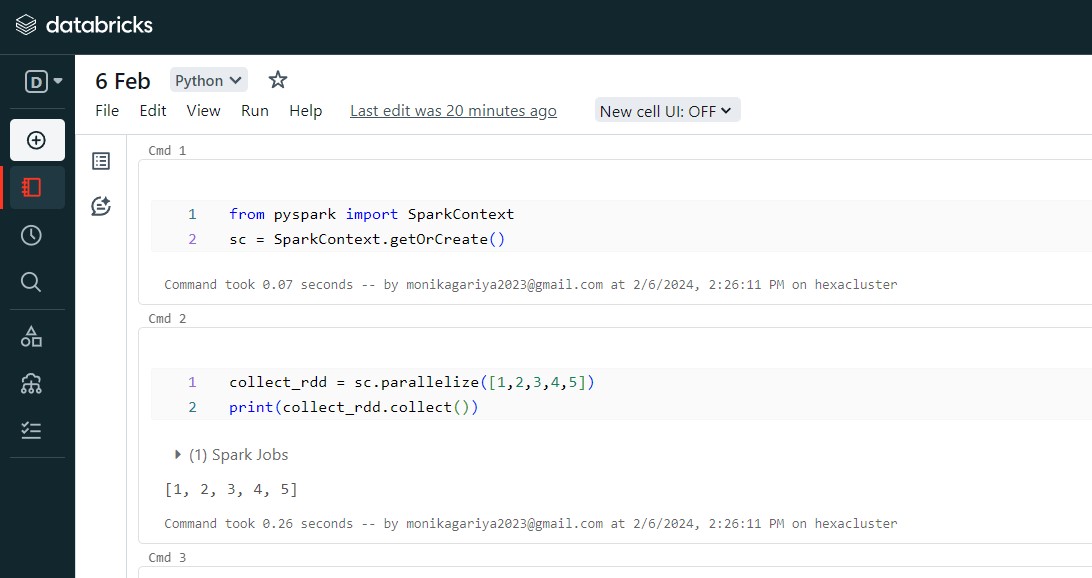
**Actions**are a kind of operation which are applied on an RDD to produce a single value. These methods are applied on a resultant RDD and produces a non-RDD value, thus removing the laziness of the transformation of RDD.

**Actions in PySpark RDDs**

In PySpark RDDs, Actions are a kind of operation that returns a value on being applied to an RDD.

**1)The .collect() Action**

The **.collect()** action on an RDD returns a list of all the elements of the RDD. It’s a great asset for displaying all the contents of our RDD



**2. The .count() Action**

The **.count()** action on an RDD is an operation that returns the number of elements of our RDD.



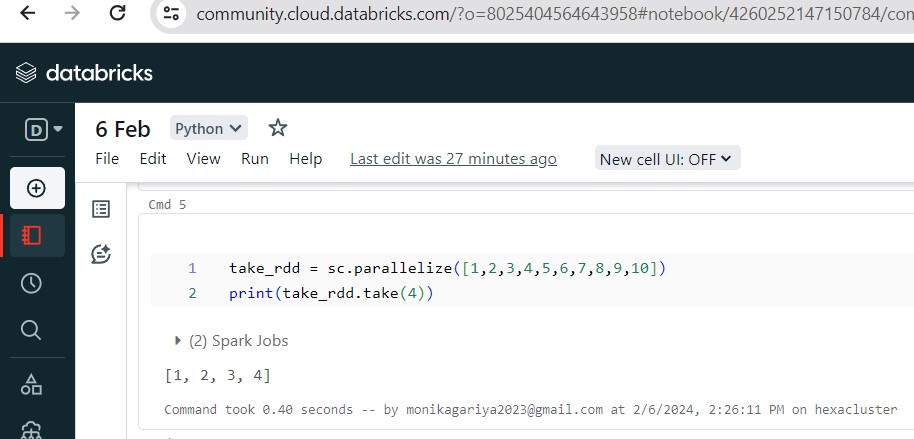
**3. The .first() Action**

The **.first()** action on an RDD returns the first element from our RDD. This can be helpful when we want to verify if the exact kind of data has been loaded in our RDD as per the requirements.



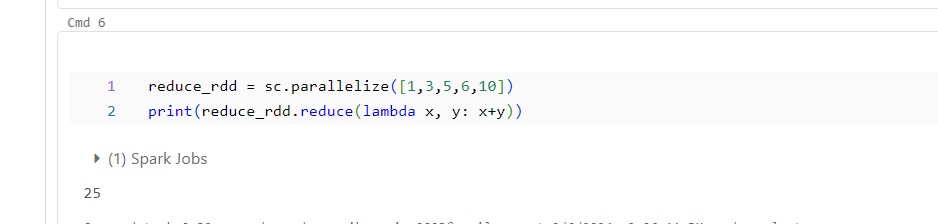
**4. The .take() Action**

The **.take(n)** action on an RDD returns n number of elements from the RDD. The ‘n’ argument takes an integer which refers to the number of elements we want to extract from the RDD



**5. The .reduce() Action**

The **.reduce()** Action takes two elements from the given RDD and operates. This operation is performed using an anonymous function or lambda



**6. The .saveAsTextFile() Action**

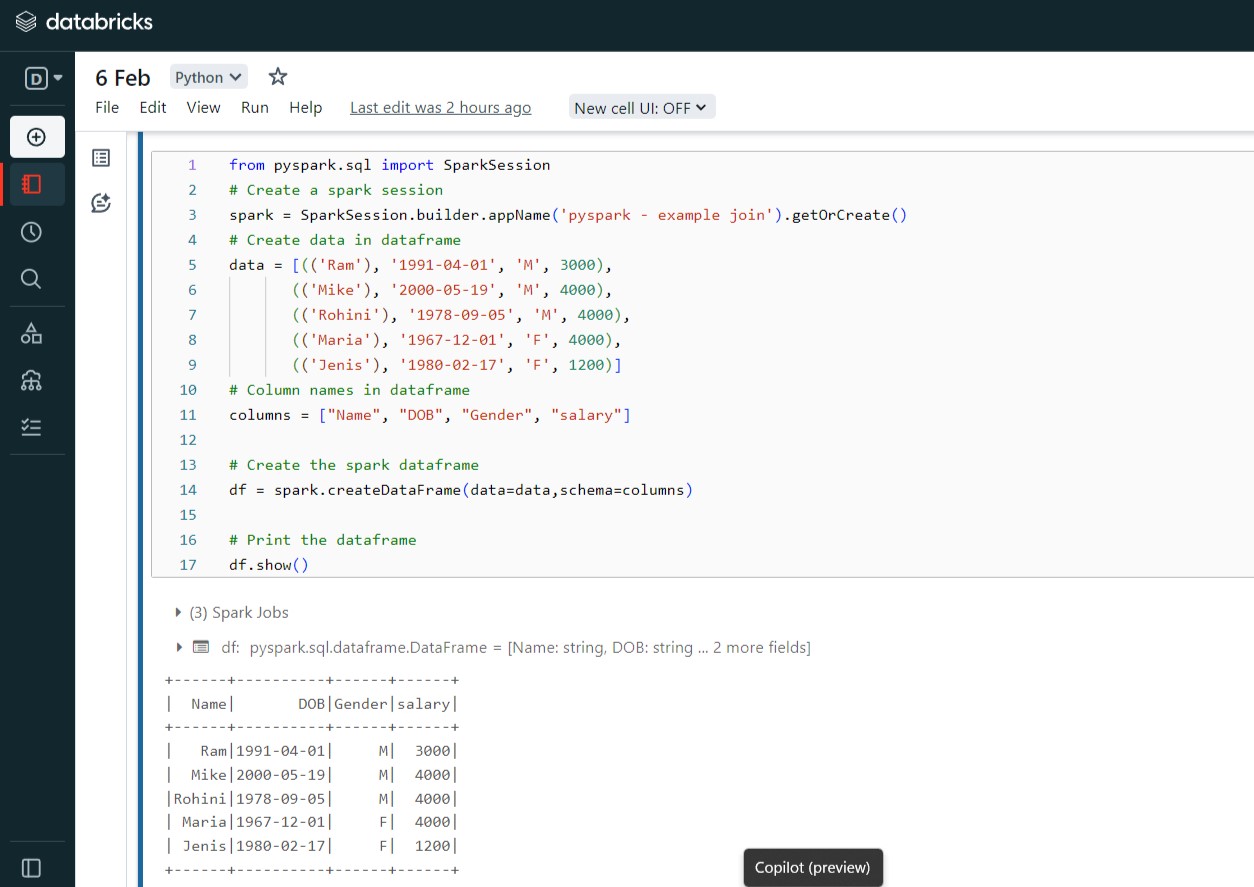
The **.saveAsTextFile()** Action is used to serve the resultant RDD as a text file. We can also specify the path to which file needed to be saved.

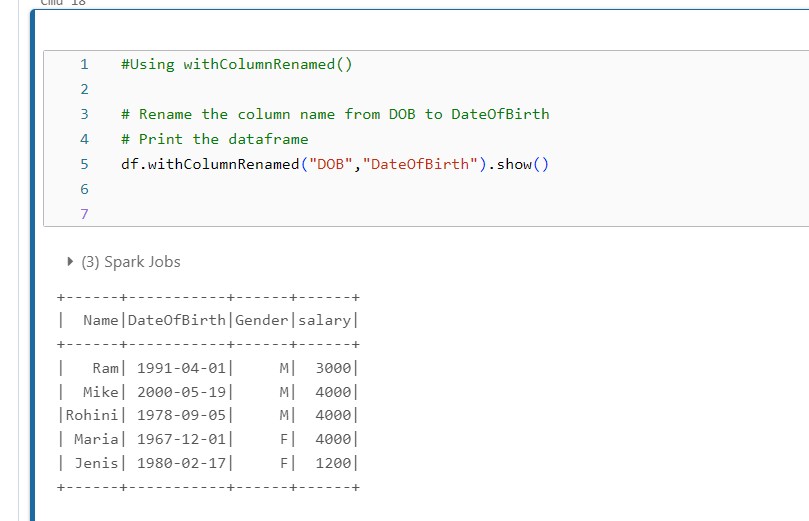


**2.Selecting,renaming and filtering data in panda dataframe**

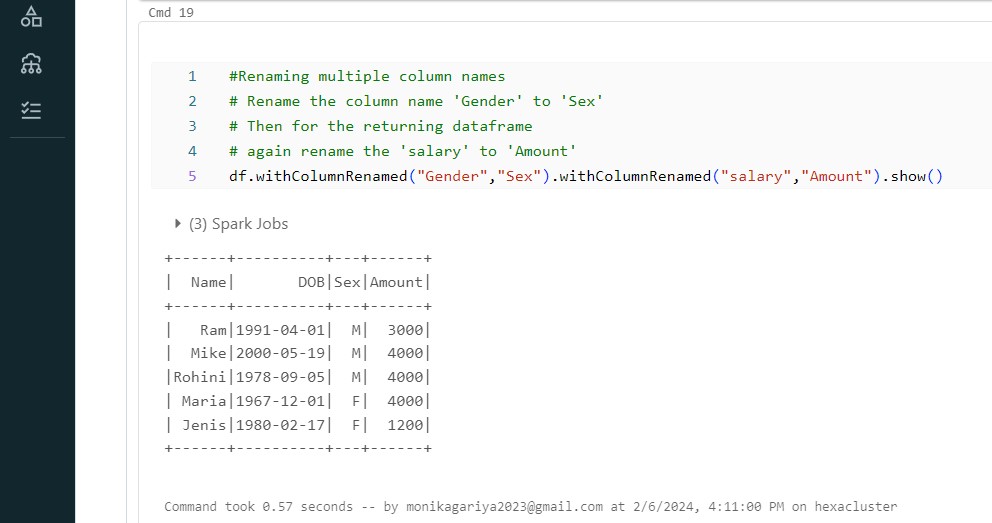
**Method 1: Using with ColumnRenamed()**

We will use of with ColumnRenamed() method to change the column names of pyspark data frame.





**Example 2:** Renaming multiple column names



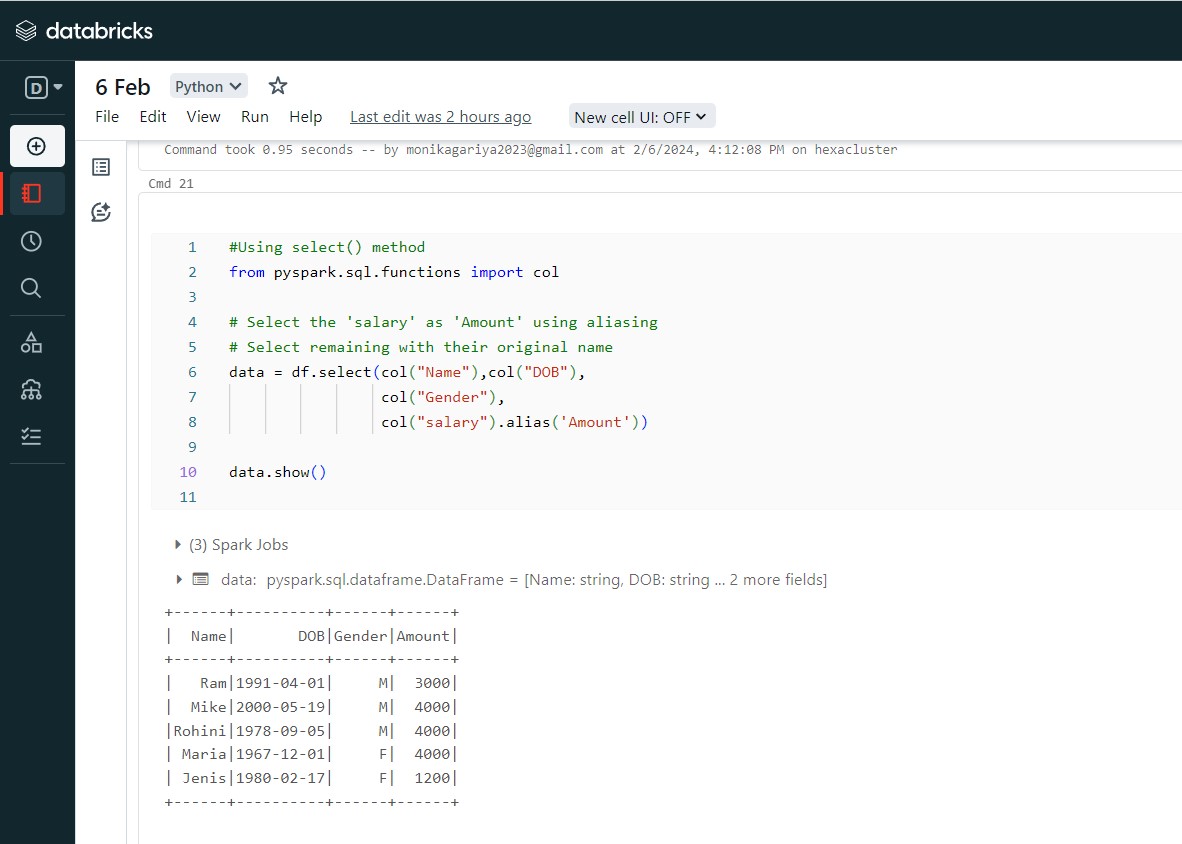
**Method 2: Using selectExpr()**

Renamingthe column names using**selectExpr()**method



**Method 3: Using select() method**

Selects the cols in the dataframe and returns a new DataFrame.



**Method 4: Using toDF()**

This function returns a new DataFrame that with new specified column names.

