**Name:** V Venkata Sri Prasad

**Batch:** Data Engineering

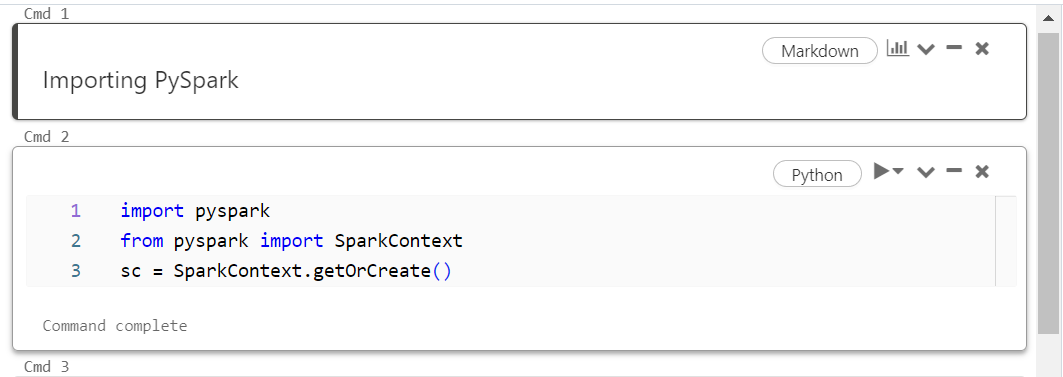
**Date:** 6/02/2024 – (Day 15)

**Topics**

1. Transforming data with PySpark RDDs & its hands on
2. Selecting, Renaming, Filtering Data in a Pandas DataFrame
3. Dataframes in pyspark

**PySpark:**

PySpark is the Python API for Apache Spark, an open-source distributed computing system. It provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.



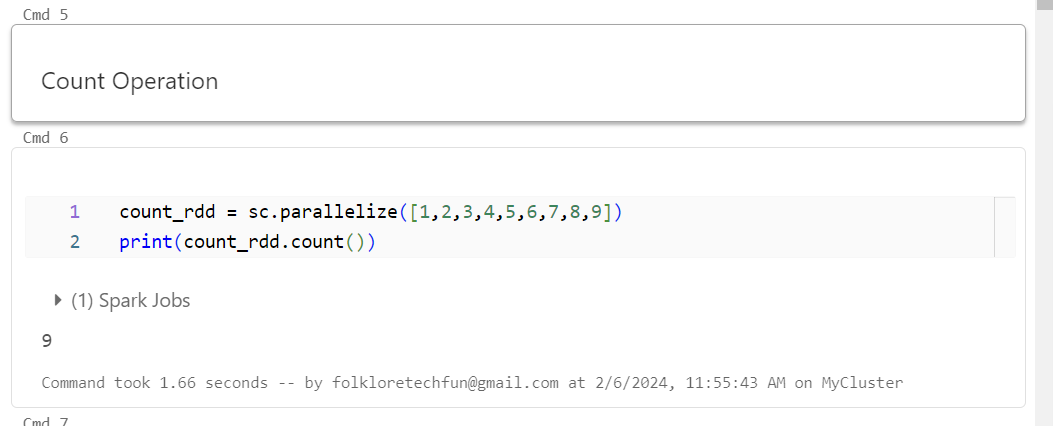
**collect():**

In PySpark, collect() is an action that retrieves all elements of the RDD (Resilient Distributed Dataset) or DataFrame and returns them as a list or array to the driver program. It should be used with caution as it brings all the data to the driver, which can cause out-of-memory errors if the dataset is too large.



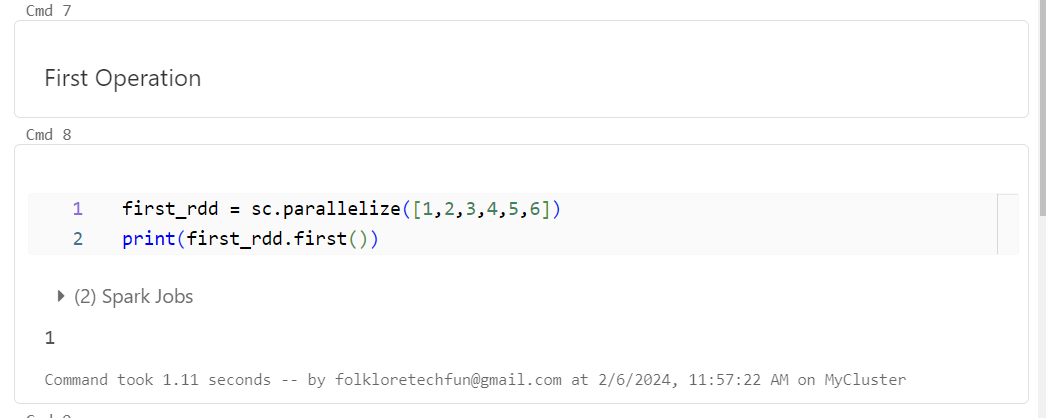
**count():**

The count() action is used to count the number of elements in an RDD or DataFrame. It returns the total count of elements in the dataset.



**first():**

The first() action returns the first element in an RDD or DataFrame. It is often used to quickly inspect the data or extract the first element for further processing.



**reduce():**

reduce() is a higher-order function used in PySpark to aggregate the elements of an RDD or DataFrame using a specified function. It takes a function that operates on two elements of the dataset and repeatedly applies it to pairs of elements until only one element remains.



**take():** The take() action is used to retrieve a specified number of elements from an RDD or DataFrame and returns them as a list or array. It is often used for quickly inspecting a small subset of the data.



**saveAsFile:**

This seems to be a typo; it should be saveAsFile(). In PySpark, saveAsFile() is a method used to save the contents of an RDD to a file or storage system. It is used to write the RDD data to an external location such as HDFS, Amazon S3, or local filesystem.



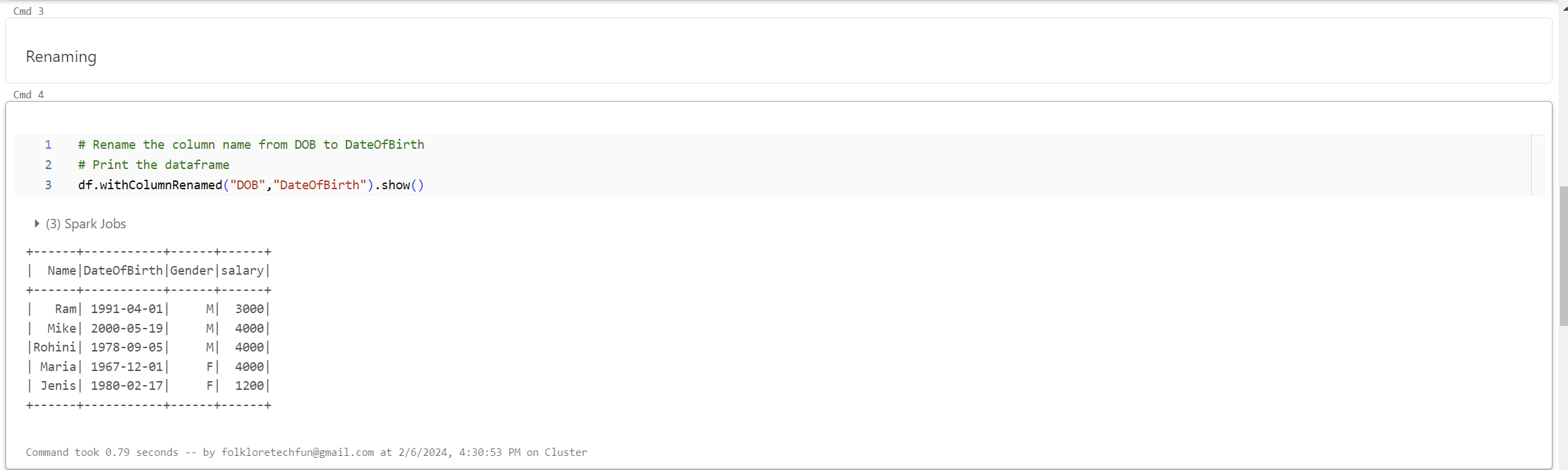
**Method 1: Using withColumnRenamed()**

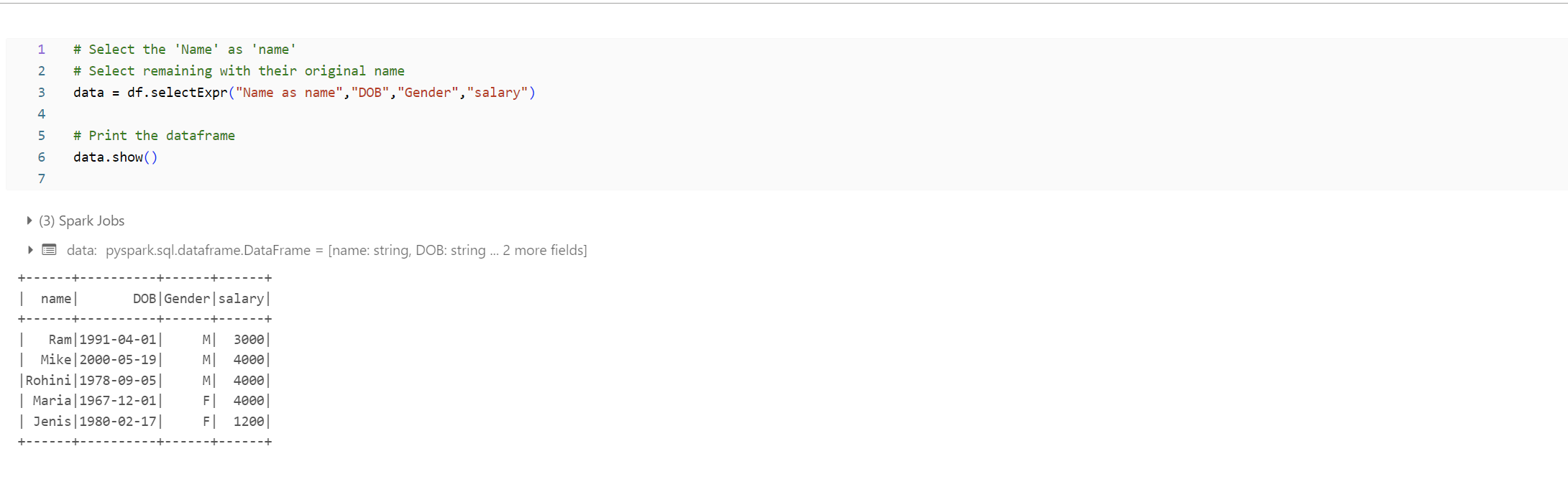
**Purpose: This method is specifically for renaming columns in a DataFrame.**



**Method 2: Using selectExpr()**

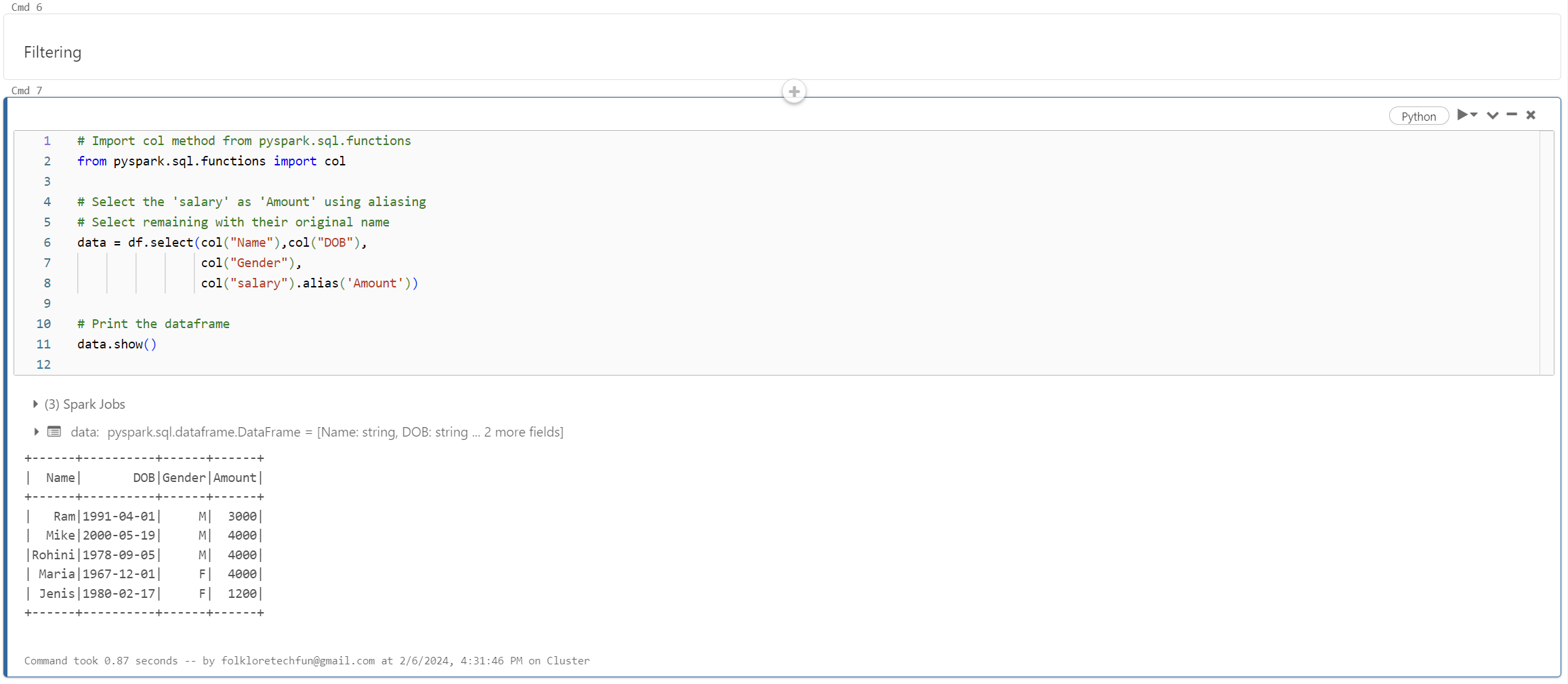
**Purpose: This method allows you to select columns and apply expressions to them simultaneously.**





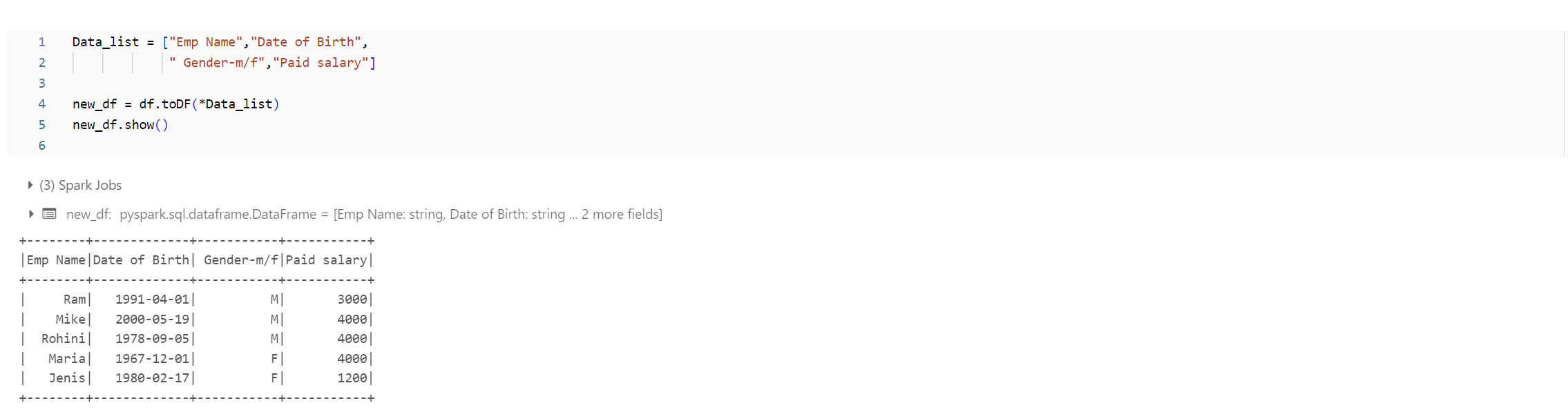
**Method 3: Using select() method**

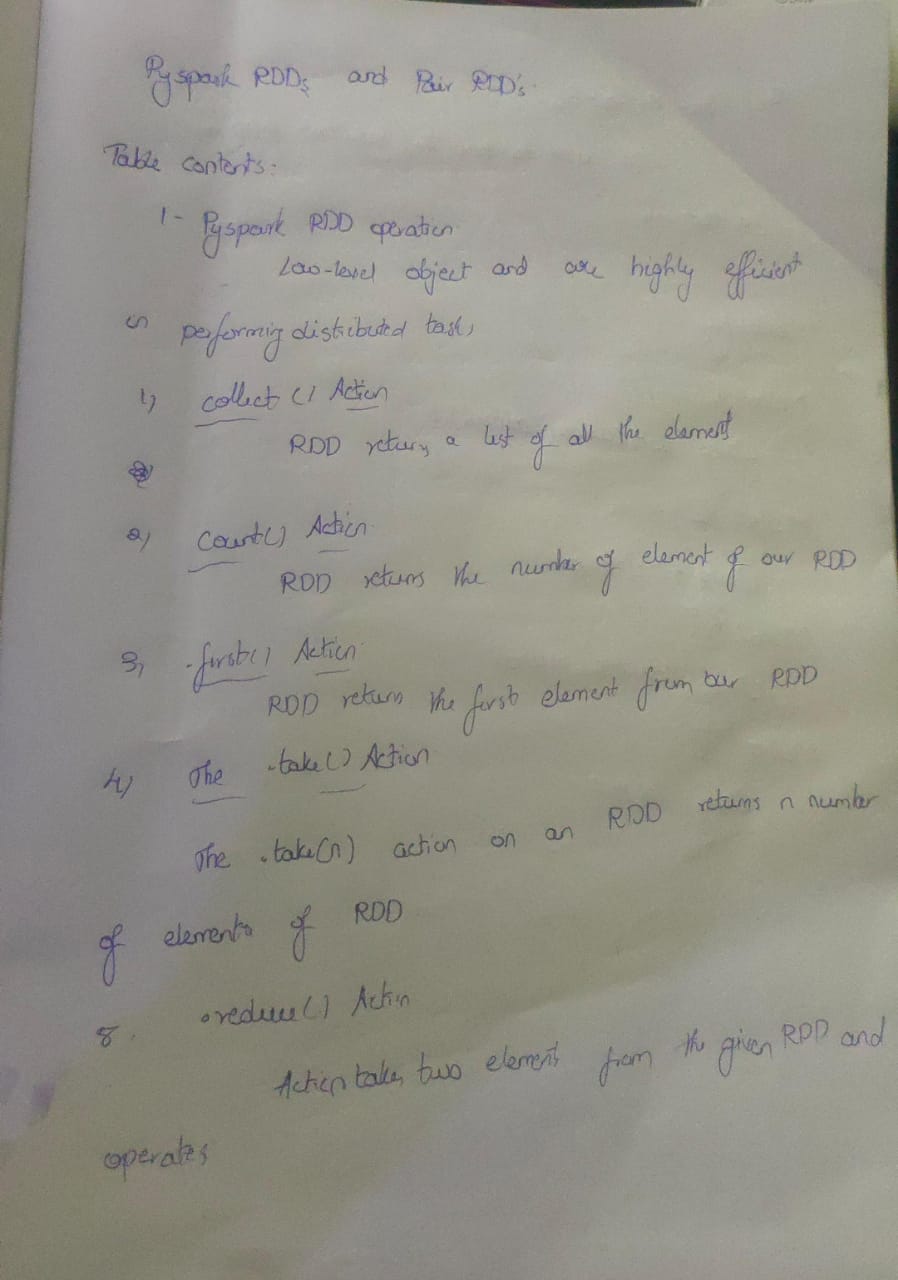
**Purpose: This method is used for selecting one or more columns from a DataFrame.**



**Method 4: Using toDF()**

**Purpose: This method is used to change the column names of a DataFrame altogether.**



**Daily Notes**

