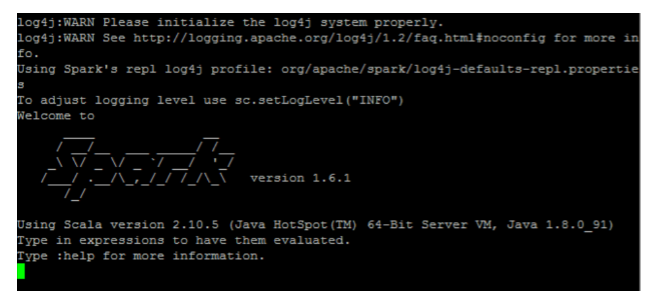
**Name : Hemalatha Karumuri**

**Date : 23-12-23**

Pyspark Commands

1)Opening spark shell:

To start the shell command prompt we need to open spark-shell from our installed spark folder. When it opened it looks like this.



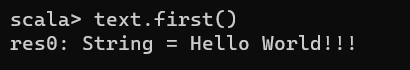
2)Read a file from the local system:

We need to use this syntax:



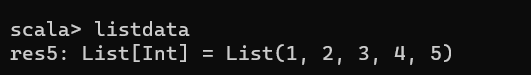
Here “sc” is the spark context. To read “data.txt” located in the home directory, you don’t need to specify the full path. However, you must provide the full path if it’s located elsewhere.

3)To display the contents of the file:

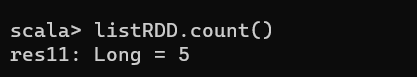


4)Create RDD through parallelizing





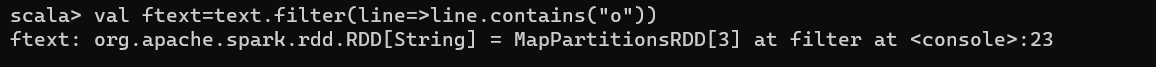
5)To count the number of elements in the list



6) Filter on RDD

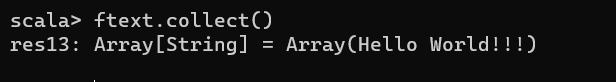
Let’s create a new RDD for items that contain “o”

To apply the transformation filter to an existing RDD and generate a new list of items, you must filter for the word “o”



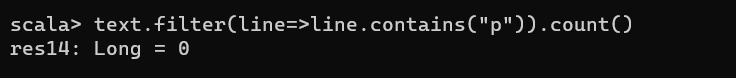
7)Collect():

This function returns all RDD’s content to the driver program. This helps debug at various steps of the writing program.



8)Chain operation:

Here filter transformation and count action acted together. This is called chain operation.



**9)Cache a File**

Caching is an optimization technique. Caching RDD means RDD will reside in memory, and all future computations will be done on those RDD in memory. It saves disk read time and improves performance. In short, it reduces the time to access the data.



10)To show total content of the csv file

