MANDAPALLI HEMA

ASSESSMENT-18

BASIC NEW COMMANDS:

1. To start the Spark shell:

$bin\spark-shell is used to start the spark shell.

2. Read File from local filesystem and create an RDD.:

[php]scala> val data = sc.textFile(“data.txt”)[/php]

3. Create RDD through parallelizing:

PySpark shell provides SparkContext variable “sc”, use sc.parallelize() to create an RDD.

example:rdd = sc.parallelize([1,2,3,4,5,6,7,8,9,10])

1. Count Items in RDD:

Scala>Newdata.count()

1. Collect():

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

1. Save output/processed data into the text file:

Scala> counts.saveAsTextFile(“outputs”)

1. Filter on RDD:

Scala>val DFData = datafilter(line => line.contains(“Yes”)

1. Chain operation:

Scala>Datafilter(line => line.contains(“DataFair”)).count()

9.10)To show total content of the csv file:



