1. Create RDD for above text file with 3 partitions.

readTxtFile=sc.textFile("gs://cksproject/rdd\_Sample\_file.txt",3)

1. display the contents of file using RDD.

For i in readTxtFile.collect(): print i

1. create individual RDD for first 3 lines.
2. splitWithLine=readTxtFile.map(lambda x:x.split(“,”))

rddOfLine1=splitWithLine.map(lambda x:x[0])

rddOfLine2=splitWithLine.map(lambda x:x[1])

rddOfLine3=splitWithLine.map(lambda x:x[2])

for i in rddOfLine1.collect() : print i

for i in rddOfLine2.collect(): print i

for i in rddOfLine3.collect(): print i

1. Give a wordcount for each token in the file and save it to another file.

word=readTxtFile.flatMap(lambda x: x.split(" "))

wordLower=word.map(lambda x: (x.split(",")[0].lower(),1))

for i in wordLower.collect() : print i

wordLower.countByKey().saveAsTextFile("gs://cksproject/output1.txt")

1. Differentiate between functionality of map and flatmap.

Map function does one to one operation on each element in a rdd , the input number of element and output element in RDDare same . it will be an array of n number of array .

FlatMap flatten the output , when we apply it on a rdd of size N, it will output a rdd of size 1. It will be array of only one array .

6. Display data in each partition individually and load them into separate files.

readTxtFile.glom().collect()