Start pyspark in databricks:

import pyspark

from pyspark import SparkContext

sc=SparkContext.getOrCreate()

import file from local to databricks:

rdd=sc.textFile('/FileStore/tables/StudentData-1.csv')

To remove the first row which contains column names is :

headers = rdd.first()

rdd = rdd.filter(lambda x: x!=headers)

rdd = rdd.map(lambda x: x.split(','))

Show the number of students in the file:

rdd.count()

Show the total marks achieved by Female and Male students

rdd1=rdd.map(lambda x:(x.split(',')))

rdd2=rdd1.map(lambda x:(x[1],(int(x[5]),1)))

rdd3=rdd2.reduceByKey(lambda x,y:(x[0]+y[0],x[1]+y[1]))

rdd3.collect()

Show the total number of students that have passed and failed. 50+ marks are required to pass the course:

rdd4=rdd2.map(lambda x: 'pass' if x[1][0]>50 else 'fail')

rdd5=rdd4.map(lambda x:(x,1))

rdd6=rdd5.reduceByKey(lambda x,y:x+y)

rdd6.collect()

OR

rdd4=rdd1

passed=rdd4.filter(lambda x: int(x[5]) >50).count()

failed=rdd4.filter(lambda x:int(x[5])<=50).count()

print(passed,failed)

Show the total number of students enrolled per course:

rdd7=rdd1.map(lambda x:(x[3],1))

rdd8=rdd7.reduceByKey(lambda x,y:x+y)

rdd8.collect()

Show the total marks that students have achieved per course:

rdd9=rdd1.map(lambda x:(x[3],int(x[5])))

rdd10=rdd9.reduceByKey(lambda x,y:x+y)

rdd10.collect()

Show the average marks that students have achieved per course:

rdd11=rdd1.map(lambda x:(x[3],(int(x[5]),1)))

rdd12=rdd11.reduceByKey(lambda x,y:(x[0]+y[0],x[1]+y[1]))

rdd13=rdd12.map(lambda x:(x[0],x[1][0]/x[1][1])) or rdd12.mapValues(lambda x: x[0]/x[1])

rdd13.collect()

show maximum and minimum marks:

rdd14=rdd1.map(lambda x: (x[3],int(x[5]) ))

print(rdd14.reduceByKey(lambda x,y : x if x > y else y).collect()

print(rdd14.reduceByKey(lambda x,y : x if x < y else y).collect()

Show the average age of male and female students:

rdd14=rdd1.map(lambda x:(x[1],(int(x[0]),1)))

rdd15=rdd14.reduceByKey(lambda x,y:(x[0]+y[0],x[1]+y[1]))

rdd16=rdd15.map(lambda x:(x[0],x[1][0]/x[1][1])) or rdd15.mapValues(lambda x: x[0]/x[1])

rdd16.collect()