**Session 6**

**Assignment 6.4**

Student Name: Karthik K

Course: Big Data Hadoop & Spark Training

Problem Statement

Problem Statement 1:

1. Read the text file, and create a tupled rdd.
2. Find the count of total number of rows present.
3. What is the distinct number of subjects present in the entire school
4. What is the count of the number of students in the school, whose name is Mathew and Marks is 55

Problem Statement 2:

1. What is the count of students per grade in the school?
2. Find the average of each student (Note - Mathew is grade-1, is different from Mathew in some other grade!)
3. What is the average score of students in each subject across all grades?
4. What is the average score of students in each subject per grade?
5. For all students in grade-2, how many have average score greater than 50?

Problem Statement 3:

Are there any students in the college that satisfy the below criteria:

1. Average score per student\_name across all grades is same as average score per student\_name per grade

Hint - Use Intersection Property.

Spark Operations – Problem Statement 1

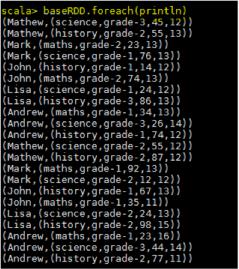
Task 1.1 Read the text file, and create a tupled rdd.

Below is the code we use to read the text file using spark context and creating a tuple RDD,

***scala> val baseRDD = sc.textFile("/home/acadgild/hadoop/17.2\_Dataset.txt").map(x => (x.split(",")(0),(x.split(",")(1),x.split(",")(2),x.split(",")(3).toInt,x.split(",")(4).toInt)))***

***scala> baseRDD.foreach(println)***

We have create a tuple RDD with name as Key and the subject, grades and the marks as values.

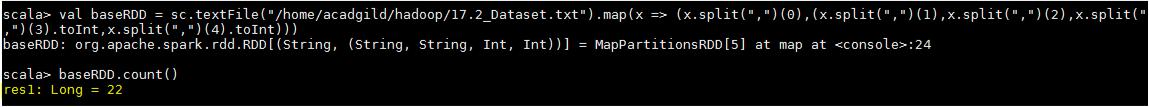


*Task1.2 -* Find the count of total number of rows present

***scala> baseRDD.count()***

***res1: Long = 22***

By using count() function we can see the number of lines present in the text file.



Task1.3 - What is the distinct number of subjects present in the entire school?

Please see the below codes used for this task,

***val baseRDD = sc.textFile("/home/acadgild/hadoop/17.2\_Dataset.txt").map(x=> (x.split(",")(1),1))val RDDreduce = baseRDD.reduceByKey((x,y)=>(x+y))***

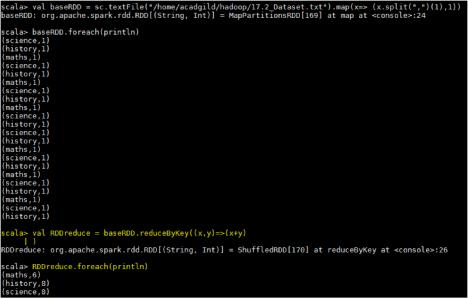
***scala> RDDreduce.foreach(println)***

***(maths,6)***

***(history,8)***

***(science,8)***

First we are creating a RDD to read the file and selecting only subject name and mapping them with value 1 and counting the values of occurrences using **reduceByKey** to get distinct number of subjects.



Task1.4 - What is the count of the number of students in the school, whose name is Mathew and marks is 55

Below are the codes,

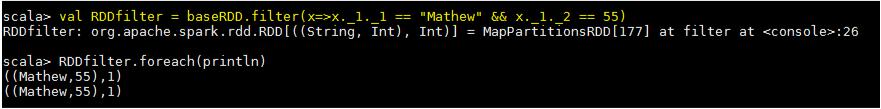
***val baseRDD = sc.textFile("/home/acadgild/hadoop/17.2\_Dataset.txt").map(x => ((x.split(",")(0),x.split(",")(3).toInt),1))***

***val RDDfilter = baseRDD.filter(x=>x.\_1.\_1 == "Mathew" && x.\_1.\_2 == 55)***

***val RDDreduce = RDDfilter.reduceByKey((x,y)=> x+y).foreach(println)***



Filter the tuple RDD by providing the condition Mather and mark as 55,



Now we are counting each occurrences using the **reduceByKey** and the required output is below,

