Session 19:
RDD DEEP DIVE
Assignment 1

Task 1

1.Write a program to read a text file and print the number of rows of data in the document.

Code :

```
package Example

import org.apache.spark.sql.SparkSession

object RDD1 {
  def main(args: Array[String]): Unit = {
    println("hey scala") //Let us create a spark session object

    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL Use Case 1 ")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("Spark Session Object created")

      val data =
spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt");
    println("19_Dataset Data->>"+data.count())

  }
}
```

File Edit Format View Help

```
Mathew,science,grade-3,45,12
Mathew,history,grade-2,55,13
Mark,maths,grade-2,23,13
Mark,science,grade-1,76,13
John,history,grade-1,14,12
John,maths,grade-2,74,13
Lisa,science,grade-1,24,12
Lisa,history,grade-3,86,13
Andrew,maths,grade-1,34,13
Andrew,science,grade-3,26,14
Andrew,history,grade-1,74,12
Mathew,science,grade-2,55,12
Mathew,history,grade-2,87,12
Mark,maths,grade-1,92,13
Mark,science,grade-2,12,12
John,history,grade-1,67,13
John,maths,grade-1,35,11
Lisa,science,grade-2,24,13
Lisa,history,grade-2,98,15
Andrew,maths,grade-1,23,16
Andrew,science,grade-3,44,14
Andrew,history,grade-2,77,11
```

Output :

```
18/09/13 23:51:56 INFO DAGScheduler: ResultStage 0 (count at RDD1.scala:18) finished in 0.42
18/09/13 23:51:56 INFO DAGScheduler: Job 0 finished: count at RDD1.scala:18, took 0.780515 s
19_Dataset Data->>22
18/09/13 23:51:56 INFO SparkContext: Invoking stop() from shutdown hook
```

2. Write a program to read a text file and print the number of words in the document.

Code:

```scala
package Example

import org.apache.spark.sql.SparkSession
object RDD2  {
  def main(args: Array[String]): Unit = {
    println("hey scala") //Let us create a spark session object

    val spark = SparkSession
    .builder()
      .master("local")
      .appName("Spark  ")
      .config("spark.some.config.option", "some-value")
```

```
      .getOrCreate()
    println("Spark Session Object created")

    val data = spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Wordcount.txt");

    val words = data.flatMap(word=> word.split(" "))
    println("Word Count->>"+words.count())

  }

}
```

Dataset :

Output:

```
18/09/13 23:45:57 INFO DAGScheduler: Job 0 finished: count at RDD2.scala:20, took 1.065594 s
Word Count->>5
18/09/13 23:45:57 INFO SparkContext: Invoking stop() from shutdown hook
```

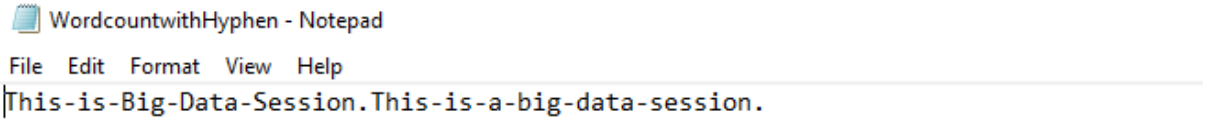
3. We have a document where the word separator is -, instead of space. Write a spark code, to obtain the count of the total number of words present in the document.

Code :

```
package Example

import org.apache.spark.sql.SparkSession

object Task21 {

  def main(args: Array[String]): Unit = {
    println("hey scala") //Let us create a spark session object

    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL Use Case 1 ")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("Spark Session Object created")
```

```
    val rdd = spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/WordcountwithHyphen.txt");
    val words = rdd.flatMap(word=> word.split("-"))

    println("Word Count->>"+words.count())
  }
}
```

Dataset :



Output :

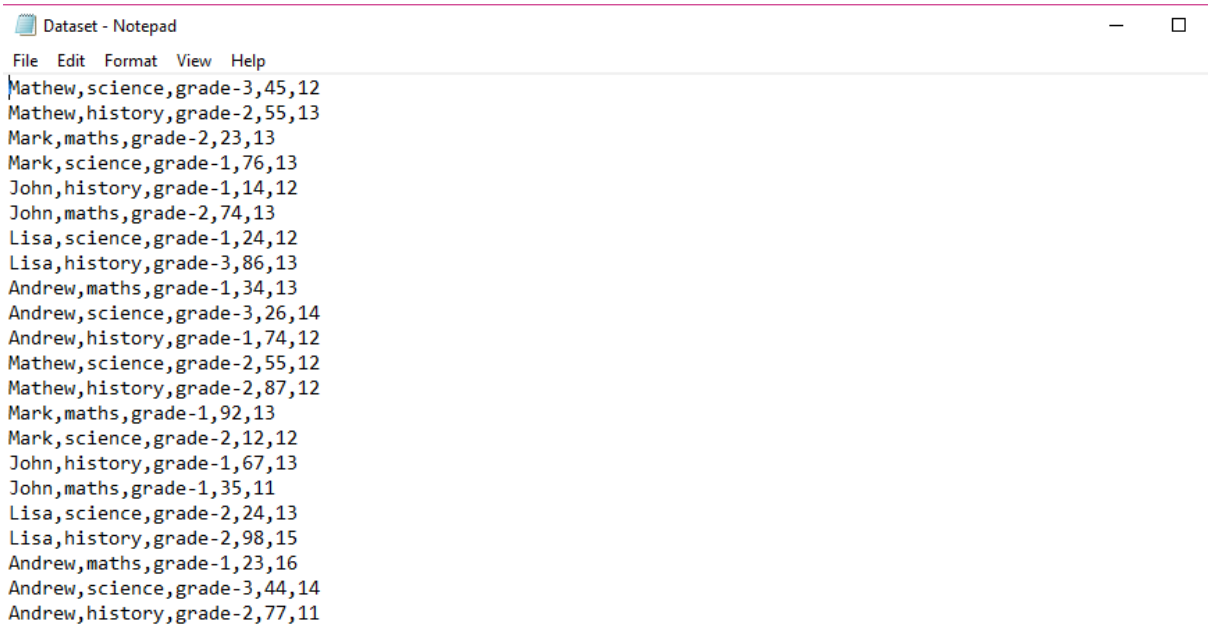
```
18/09/13 23:57:24 INFO DAGScheduler: ResultStage 0 (count at Task21.scala:21) finished in 0
18/09/13 23:57:24 INFO DAGScheduler: Job 0 finished: count at Task21.scala:21, took 0.69606
Word Count->>10
18/09/13 23:57:24 INFO SparkContext: Invoking stop() from shutdown hook
18/09/13 23:57:24 INFO SparkUI: Stopped Spark web UI at http://192.168.100.5:4040
```

Task 2

Problem Statement 1:

1. **Read the text file, and create a tupled rdd.**
2. **Find the count of total number of rows present.**

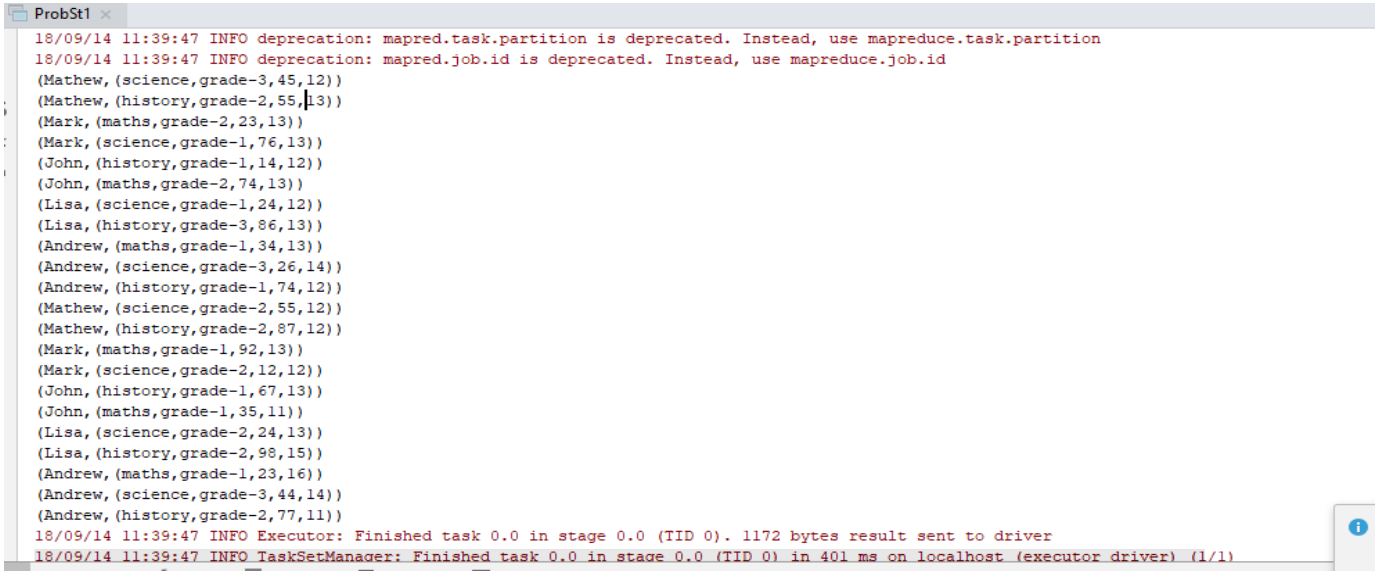
Data set



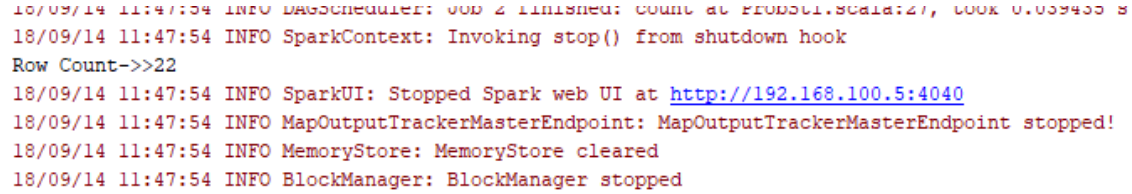
Code :

```
package Example

import org.apache.spark.sql.SparkSession

object ProbSt1 {

  def main(args: Array[String]): Unit = {
    println("hey scala") //Let us create a spark session object

    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL Use Case 1 ")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("Spark Session Object created")



    val baseRDD =
spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt").map(x =>

(x.split(",")(0),(x.split(",")(1),x.split(",")(2),x.split(",")(3).toInt,x.split(","
)(4).toInt)))

   baseRDD.foreach(println)
    baseRDD.count()

   println("Row Count->>"+baseRDD.count())
  }
}
```

Output for Task 1:



Output for task 2:



3. **What is the distinct number of subjects present in the entire school**

Code :

```
package Example

import org.apache.spark.sql.SparkSession

object ProbSt1 {

  def main(args: Array[String]): Unit = {
    println("hey scala") //Let us create a spark session object

    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL Use Case 1 ")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("Spark Session Object created")
```

```
    val baseRDD =
spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt").map(x=> (x.split(",")(1),1))

    val RDDreduce = baseRDD.reduceByKey((x,y)=>(x+y))
    RDDreduce.foreach(println)
  }
}
```

Output:

```
18/09/14 11:52:28 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/14 11:52:28 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 8 ms
(maths,6)
(history,8)
(science,8)
18/09/14 11:52:28 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1809 bytes result sent to driver
18/09/14 11:52:28 INFO DAGScheduler: ResultStage 1 (foreach at ProbSt1.scala:24) finished in 0.153 s
18/09/14 11:52:28 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 152 ms on localhost (executor
```

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

Code:

```
package Example

import org.apache.spark.sql.SparkSession

object ProbSt1 {

  def main(args: Array[String]): Unit = {
    println("hey scala") //Let us create a spark session object

    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL Use Case 1 ")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("Spark Session Object created")


    val baseRDD =
spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt").map(x =>
      ((x.split(",")(0),x.split(",")(3).toInt),1))
    baseRDD.foreach(println)
    val RDDfilter = baseRDD.filter(x=>x._1._1 == "Mathew" && x._1._2 == 55)

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

  }
```

```
}
```

Output:

```
18/09/14 11:58:43 INFO deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
((Mathew,45),1)
((Mathew,55),1)
((Mark,23),1)
((Mark,76),1)
((John,14),1)
((John,74),1)
((Lisa,24),1)
((Lisa,86),1)
((Andrew,34),1)
((Andrew,26),1)
((Andrew,74),1)
((Mathew,55),1)
((Mathew,87),1)
((Mark,92),1)
((Mark,12),1)
((John,67),1)
((John,35),1)
((Lisa,24),1)
((Lisa,98),1)
((Andrew,23),1)
((Andrew,44),1)
((Andrew,77),1)
18/09/14 11:58:43 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1172 bytes result sent to driver
18/09/14 11:58:43 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 331 ms on localhost (executor driver) (1/1)
18/09/14 11:58:43 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
```

```
18/09/14 11:58:43 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/14 11:58:43 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 11 ms
((Mathew,55),2)
18/09/14 11:58:43 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2). 1632 bytes result sent to driver
18/09/14 11:58:43 INFO DAGScheduler: ResultStage 2 (foreach at ProbSt1.scala:24) finished in 0.078 s
```

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?

Code :

```scala
package Example

import org.apache.spark.sql.SparkSession

object ProbSt1 {

  def main(args: Array[String]): Unit = {
    println("hey scala") //Let us create a spark session object
```

```
    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL Use Case 1 ")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("Spark Session Object created")

//1. What is the count of students per grade in the school?
    val baseRDD =
spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt").map(x =>
      (x.split(",")(2),1)).reduceByKey((x,y)=>x+y).foreach(println)


    //Find the average of each student (Note - Mathew is grade-1, is different from
Mathew in
    //some other grade!)

    val baseRDD1 =
spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt").map(x=>((x.split(",")(0),x.split(",")(2)),x.split(
      ",")(3).toInt))
    val RDDmap = baseRDD1.mapValues(x=>(x,1))
    RDDmap.foreach(println)
    val RDDreduce = RDDmap.reduceByKey((x,y) => (x._1 + y._1, x._2 + y._2))
    RDDreduce.foreach(println)
    val StudAvg = RDDreduce.mapValues{case(sum,count)=>(1.0*sum)/count}
    StudAvg.foreach(println)

//What is the average score of students in each subject across all grades?

    val baseRDD2 =
      spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt").map(x=>((x.split(",")(0),x.split(",")(1)),x.split(
        ",")(3).toInt))
    baseRDD2.foreach(println)
    val RDDmap2 = baseRDD2.mapValues(x=>(x,1))
    RDDmap2.foreach(println)
    val RDDreduce2 = RDDmap2.reduceByKey((x,y)=>(x._1+y._1,x._2+y._2))
    RDDreduce2.foreach(println)
    val SubAvg2 = RDDreduce2.mapValues{case(sum,count)=>(1.0*sum)/count}
    SubAvg2.foreach(println)


    //What is the average score of students in each subject per grade?
    val baseRDD3 =
      spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt").map(x=>((x.split(",")(1),x.split(",")(2)),x.split(
        ",")(3).toInt))
    baseRDD3.foreach(println)
    val RDDmapvalue3 = baseRDD3.mapValues(x=>(x,1))
    RDDmapvalue3.foreach(println)
    val RDDreduce3 = RDDmapvalue3.reduceByKey((x,y)=>(x._1+y._1,x._2+y._2))
    RDDreduce3.foreach(println)
    val AvgGrade =
RDDreduce3.mapValues{case(sum,count)=>(1.0*sum)/count}.foreach(println)

  //  5. For all students in grade-2, how many have average score greater than 50?


    val baseRDD4 =
      spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt").map(x=>((x.split(",")(0),x.split(",")(2)),x.split(
        ",")(3).toInt))
    baseRDD4.foreach(println)
    val RDDmap4 = baseRDD4.mapValues(x=>(x,1))
```

```
    RDDmap4.foreach(println)
    val RDDreduce4 = RDDmap4.reduceByKey((x,y)=>(x._1+y._1,x._2+y._2))
    RDDreduce4.foreach(println)
    val RDDavg = RDDreduce.mapValues{case(sum,count)=>(1.0*sum)/count}
    val RDDfiltermap = RDDavg.filter(x=>x._1._2 == "grade-2" && x._2>50).count()

    val RDDfiltermap1 = RDDavg.filter(x=>x._1._2 == "grade-2" &&
x._2>50).foreach(println)
  }
}
```

Output

1. : What is the count of students per grade in the school?)

```
18/09/14 12:14:08 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/14 12:14:08 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 13 ms
(grade-3,4)
(grade-1,9)
(grade-2,9)
18/09/14 12:14:08 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1719 bytes result sent to driver
18/09/14 12:14:08 INFO DAGScheduler: ResultStage 1 (foreach at ProbSt1.scala:20) finished in 0.128 s
18/09/14 12:14:08 INFO DAGScheduler: Job 0 finished: foreach at ProbSt1.scala:20, took 1.458017 s
```

Output:

2 .Find the average of each student (Note - Mathew is grade-1, is different from Mathew in some other grade!)

```
18/09/14 12:14:09 INFO Executor: Running task 0.0 in stage 2.0 (TID 2)
18/09/14 12:14:09 INFO HadoopRDD: Input split: file:/C:/Users/admin/Desktop/Assignment_to_be submitted/s19/Dataset.txt:0+625
((Mathew,grade-3),(45,1))
((Mathew,grade-2),(55,1))
((Mark,grade-2),(23,1))
((Mark,grade-1),(76,1))
((John,grade-1),(14,1))
((John,grade-2),(74,1))
((Lisa,grade-1),(24,1))
((Lisa,grade-3),(86,1))
((Andrew,grade-1),(34,1))
((Andrew,grade-3),(26,1))
((Andrew,grade-1),(74,1))
((Mathew,grade-2),(55,1))
((Mathew,grade-2),(87,1))
((Mark,grade-1),(92,1))
((Mark,grade-2),(12,1))
((John,grade-1),(67,1))
((John,grade-1),(35,1))
((Lisa,grade-2),(24,1))
((Lisa,grade-2),(98,1))
((Andrew,grade-1),(23,1))
((Andrew,grade-3),(44,1))
((Andrew,grade-2),(77,1))
18/09/14 12:14:09 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2). 1082 bytes result sent to driver
18/09/14 12:14:09 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 28 ms on localhost (executor driver) (1/1)
18/09/14 12:14:09 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
```

```
18/09/14 12:28:27 INFO Executor: Finished task 0.0 in stage 4.0 (TID 4). 1632 bytes result sent to driver
((Lisa,grade-1),(24,1))
18/09/14 12:28:27 INFO TaskSetManager: Finished task 0.0 in stage 4.0 (TID 4) in 17 ms on localhost (exec
((Mark,grade-2),(35,2))
18/09/14 12:28:27 INFO TaskSchedulerImpl: Removed TaskSet 4.0, whose tasks have all completed, from pool
((Lisa,grade-2),(122,2))
18/09/14 12:28:27 INFO DAGScheduler: ResultStage 4 (foreach at ProbSt1.scala:31) finished in 0.019 s
((Andrew,grade-2),(77,1))
((Mathew,grade-3),(45,1))
((Andrew,grade-1),(131,3))
((Lisa,grade-3),(86,1))
((John,grade-1),(116,3))
((John,grade-2),(74,1))
((Mark,grade-1),(168,2))
((Andrew,grade-3),(70,2))
((Mathew,grade-2),(197,3))
18/09/14 12:28:27 INFO DAGScheduler: Job 2 finished: foreach at ProbSt1.scala:31, took 0.099768 s
18/09/14 12:28:27 INFO SparkContext: Starting job: foreach at ProbSt1.scala:33
18/09/14 12:28:27 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 1 is 149 bytes
```

```
18/09/14 12:28:27 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/14 12:28:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
((Lisa,grade-1),24.0)
((Mark,grade-2),17.5)
((Lisa,grade-2),61.0)
((Andrew,grade-2),77.0)
((Mathew,grade-3),45.0)
((Andrew,grade-1),43.666666666666664)
((Lisa,grade-3),86.0)
((John,grade-1),38.666666666666664)
((John,grade-2),74.0)
((Mark,grade-1),84.0)
((Andrew,grade-3),35.0)
((Mathew,grade-2),65.66666666666667)
18/09/14 12:28:27 INFO Executor: Finished task 0.0 in stage 6.0 (TID 5). 1553 bytes result sent to driver
18/09/14 12:28:27 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 5) in 21 ms on localhost (executor driver) (1/1)
18/09/14 12:28:27 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool
18/09/14 12:28:27 INFO DAGScheduler: ResultStage 6 (foreach at ProbSt1.scala:33) finished in 0.023 s
```

Output:

3 .What is the average score of students in each subject across all grades?

```
18/09/14 12:28:27 INFO Executor: Running task 0.0 in stage 8.0 (TID 7)
18/09/14 12:28:27 INFO HadoopRDD: Input split: file:/C:/Users/admin/Desktop/Assignment_to_be submitted/s19/Dataset.txt:0+625
((Mathew,science),(45,1))
((Mathew,history),(55,1))
((Mark,maths),(23,1))
((Mark,science),(76,1))
((John,history),(14,1))
((John,maths),(74,1))
((Lisa,science),(24,1))
((Lisa,history),(86,1))
((Andrew,maths),(34,1))
((Andrew,science),(26,1))
((Andrew,history),(74,1))
((Mathew,science),(55,1))
((Mathew,history),(87,1))
((Mark,maths),(92,1))
((Mark,science),(12,1))
((John,history),(67,1))
((John,maths),(35,1))
((Lisa,science),(24,1))
((Lisa,history),(98,1))
((Andrew,maths),(23,1))
((Andrew,science),(44,1))
((Andrew,history),(77,1))
18/09/14 12:28:27 INFO Executor: Finished task 0.0 in stage 8.0 (TID 7). 995 bytes result sent to driver
18/09/14 12:28:27 INFO TaskSetManager: Finished task 0.0 in stage 8.0 (TID 7) in 19 ms on localhost (executor driver) (1/1)
```

```
18/09/14 12:28:27 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/14 12:28:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
((Lisa,history),(184,2))
((Mark,maths),(115,2))
((Andrew,science),(70,2))
((Mark,science),(88,2))
((Mathew,science),(55,1))
((Andrew,maths),(57,2))
((Mathew,science),(45,1))
((Mathew,history),(142,2))
((John,maths),(109,2))
((John,history),(81,2))
((Lisa,science),(48,2))
((Andrew,history),(151,2))
18/09/14 12:28:27 INFO Executor: Finished task 0.0 in stage 10.0 (TID 9). 1553 bytes result sent to driver
18/09/14 12:28:27 INFO TaskSetManager: Finished task 0.0 in stage 10.0 (TID 9) in 17 ms on localhost (executor driver) (1/1)
18/09/14 12:28:27 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks have all completed, from pool
```

```
18/09/14 12:36:13 INFO DAGScheduler: Job 7 finished: foreach at ProbStl.scala:46, took 0.060276 s
((Lisa,history),92.0)
((Mark,maths),57.5)
((Andrew,science),35.0)
((Mark,science),44.0)
((Mathew,science),55.0)
((Andrew,maths),28.5)
((Mathew,science),45.0)
((Mathew,history),71.0)
((John,maths),54.5)
((John,history),40.5)
((Lisa,science),24.0)
((Andrew,history),75.5)
18/09/14 12:36:13 INFO SparkContext: Invoking stop() from shutdown hook
18/09/14 12:36:13 INFO SparkUI: Stopped Spark web UI at http://192.168.100.5:4040
18/09/14 12:36:13 INFO BlockManagerInfo: Removed broadcast_10_piece0 on 192.168.100.5:51708 in memory (size: 2.0 KB, free: 897.
```

Output:

4.What is the average score of students in each subject per grade?

```
18/09/14 12:42:18 INFO HadoopRDD: Input split: file:/C:/Users/admin/Desktop/Assignment_to_be submitted/s19/Dataset.txt:0+625
((science,grade-3),45)
((history,grade-2),55)
((maths,grade-2),23)
((science,grade-1),76)
((history,grade-1),14)
((maths,grade-2),74)
((science,grade-1),24)
((history,grade-3),86)
((maths,grade-1),34)
((science,grade-3),26)
((history,grade-1),74)
((science,grade-2),55)
((history,grade-2),87)
((maths,grade-1),92)
((science,grade-2),12)
((history,grade-1),67)
((maths,grade-1),35)
((science,grade-2),24)
((history,grade-2),98)
((maths,grade-1),23)
((science,grade-3),44)
((history,grade-2),77)
18/09/14 12:42:18 INFO Executor: Finished task 0.0 in stage 13.0 (TID 11). 995 bytes result sent to driver
18/09/14 12:42:18 INFO TaskSetManager: Finished task 0.0 in stage 13.0 (TID 11) in 18 ms on localhost (executor driver) (1/1)
18/09/14 12:42:18 INFO TaskSchedulerImpl: Removed TaskSet 13.0, whose tasks have all completed, from pool
```

```
18/09/14 12:42:18 INFO Executor: Running task 0.0 in stage 14.0 (TID 12)
18/09/14 12:42:18 INFO HadoopRDD: Input split: file:/C:/Users/admin/Desktop/Assignment_to_be submitted/s19/Dataset.txt:0+625
((science,grade-3),(45,1))
((history,grade-2),(55,1))
((maths,grade-2),(23,1))
((science,grade-1),(76,1))
((history,grade-1),(14,1))
((maths,grade-2),(74,1))
((science,grade-1),(24,1))
((history,grade-3),(86,1))
((maths,grade-1),(34,1))
((science,grade-3),(26,1))
((history,grade-1),(74,1))
((science,grade-2),(55,1))
((history,grade-2),(87,1))
((maths,grade-1),(92,1))
((science,grade-2),(12,1))
((history,grade-1),(67,1))
((maths,grade-1),(35,1))
((science,grade-2),(24,1))
((history,grade-2),(98,1))
((maths,grade-1),(23,1))
((science,grade-3),(44,1))
((history,grade-2),(77,1))
18/09/14 12:42:18 INFO Executor: Finished task 0.0 in stage 14.0 (TID 12). 995 bytes result sent to driver
18/09/14 12:42:18 INFO TaskSetManager: Finished task 0.0 in stage 14.0 (TID 12) in 16 ms on localhost (executor driver) (1/1)
```

```
18/09/14 12:42:19 INFO Executor: Running task 0.0 in stage 16.0 (TID 14)
18/09/14 12:42:19 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/14 12:42:19 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
((history,grade-2),(317,4))
((history,grade-3),(86,1))
((maths,grade-1),(184,4))
((science,grade-3),(115,3))
((science,grade-1),(100,2))
((science,grade-2),(91,3))
((history,grade-1),(155,3))
((maths,grade-2),(97,2))
18/09/14 12:42:19 INFO Executor: Finished task 0.0 in stage 16.0 (TID 14). 1632 bytes result sent to driver
18/09/14 12:42:19 INFO TaskSetManager: Finished task 0.0 in stage 16.0 (TID 14) in 12 ms on localhost (executor driver) (1/
18/09/14 12:42:19 INFO TaskSchedulerImpl: Removed TaskSet 16.0, whose tasks have all completed, from pool
18/09/14 12:42:19 INFO DAGScheduler: ResultStage 16 (foreach at ProbSt1.scala:57) finished in 0.017 s
```

```
18/09/14 12:42:19 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
((history,grade-2),79.25)
18/09/14 12:42:19 INFO BlockManagerInfo: Removed broadcast_15_piece0 on 192.168.100.5:51768 in memory (size: 2024.0 B, free: 897.5 MB)
((history,grade-3),86.0)
((maths,grade-1),46.0)
((science,grade-3),38.333333333333336)
((science,grade-1),50.0)
((science,grade-2),30.333333333333332)
((history,grade-1),51.666666666666664)
((maths,grade-2),48.5)
18/09/14 12:42:19 INFO Executor: Finished task 0.0 in stage 18.0 (TID 15). 1640 bytes result sent to driver
18/09/14 12:42:19 INFO TaskSetManager: Finished task 0.0 in stage 18.0 (TID 15) in 21 ms on localhost (executor driver) (1/1)
18/09/14 12:42:19 INFO TaskSchedulerImpl: Removed TaskSet 18.0, whose tasks have all completed, from pool
18/09/14 12:42:19 INFO BlockManagerInfo: Removed broadcast_16_piece0 on 192.168.100.5:51768 in memory (size: 2.1 KB, free: 897.5 MB)
18/09/14 12:42:19 INFO DAGScheduler: ResultStage 18 (foreach at ProbStl.scala:58) finished in 0.023 s
```

Output:

5 .For all students in grade-2, how many have average score greater than 50?

```
18/09/14 12:48:26 INFO Executor: Running task 0.0 in stage 19.0 (TID 16)
18/09/14 12:48:26 INFO HadoopRDD: Input split: file:/C:/Users/admin/Desktop/Assignment_to_be submitted/s19/Dataset.txt:0+625
((Mathew,grade-3),45)
((Mathew,grade-2),55)
((Mark,grade-2),23)
((Mark,grade-1),76)
((John,grade-1),14)
((John,grade-2),74)
((Lisa,grade-1),24)
((Lisa,grade-3),86)
((Andrew,grade-1),34)
((Andrew,grade-3),26)
((Andrew,grade-1),74)
((Mathew,grade-2),55)
((Mathew,grade-2),87)
((Mark,grade-1),92)
((Mark,grade-2),12)
((John,grade-1),67)
((John,grade-1),35)
((Lisa,grade-2),24)
((Lisa,grade-2),98)
((Andrew,grade-1),23)
((Andrew,grade-3),44)
((Andrew,grade-2),77)
18/09/14 12:48:26 INFO Executor: Finished task 0.0 in stage 19.0 (TID 16). 995 bytes result sent to driver
```

```
18/09/14 12:48:26 INFO Executor: Running task 0.0 in stage 20.0 (TID 17)
18/09/14 12:48:26 INFO HadoopRDD: Input split: file:/C:/Users/admin/Desktop/Assignment_to_be submitted/s19/Dataset.txt:0+625
((Mathew,grade-3),(45,1))
((Mathew,grade-2),(55,1))
((Mark,grade-2),(23,1))
((Mark,grade-1),(76,1))
((John,grade-1),(14,1))
((John,grade-2),(74,1))
((Lisa,grade-1),(24,1))
((Lisa,grade-3),(86,1))
((Andrew,grade-1),(34,1))
((Andrew,grade-3),(26,1))
((Andrew,grade-1),(74,1))
((Mathew,grade-2),(55,1))
((Mathew,grade-2),(87,1))
((Mark,grade-1),(92,1))
((Mark,grade-2),(12,1))
((John,grade-1),(67,1))
((John,grade-1),(35,1))
((Lisa,grade-2),(24,1))
((Lisa,grade-2),(98,1))
((Andrew,grade-1),(23,1))
((Andrew,grade-3),(44,1))
((Andrew,grade-2),(77,1))
18/09/14 12:48:26 INFO Executor: Finished task 0.0 in stage 20.0 (TID 17). 995 bytes result sent to driver
```

```
18/09/14 12:48:26 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/14 12:48:26 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
((Lisa,grade-1),(24,1))
((Mark,grade-2),(35,2))
((Lisa,grade-2),(122,2))
((Andrew,grade-2),(77,1))
((Mathew,grade-3),(45,1))
((Andrew,grade-1),(131,3))
((Lisa,grade-3),(86,1))
((John,grade-1),(116,3))
((John,grade-2),(74,1))
((Mark,grade-1),(168,2))
((Andrew,grade-3),(70,2))
((Mathew,grade-2),(197,3))
18/09/14 12:48:26 INFO Executor: Finished task 0.0 in stage 22.0 (TID 19). 1632 bytes result sent to driver
18/09/14 12:48:26 INFO TaskSetManager: Finished task 0.0 in stage 22.0 (TID 19) in 18 ms on localhost (executor driver) (1/1)
18/09/14 12:48:26 INFO TaskSchedulerImpl: Removed TaskSet 22.0, whose tasks have all completed, from pool
18/09/14 12:48:26 INFO DAGScheduler: ResultStage 22 (foreach at ProbSt1.scala:70) finished in 0.019 s
18/09/14 12:48:26 INFO DAGScheduler: Job 14 finished: foreach at ProbSt1.scala:70, took 0.096555 s
```

```
18/09/14 12:48:26 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
((Lisa,grade-2),61.0)
((Andrew,grade-2),77.0)
((John,grade-2),74.0)
((Mathew,grade-2),65.66666666666667)
18/09/14 12:48:26 INFO Executor: Finished task 0.0 in stage 26.0 (TID 21). 1553 bytes result sent to driver
```

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

Code :

```scala
package Example

import org.apache.spark.sql.SparkSession

object Intersection {

  def main(args: Array[String]): Unit = {
    println("hey scala") //Let us create a spark session object

    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL Use Case 1 ")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("Spark Session Object created")

  //Prob Statement 3: Are there any students in the college that satisfy the below
```

```
criteria:
  val baseRDD5 =
    spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt").map(x=>(x.split(",")(0),x.split(",")(3).toInt))
  baseRDD5.foreach(println)
  val studAvg = baseRDD5.mapValues(x=>(x,1))
  studAvg.foreach(println)
  val studReduce = studAvg.reduceByKey((x,y)=> (x._1+y._1,x._2+y._2))
  studReduce.foreach(println)
  val Avg_Stud = studReduce.mapValues{case (sum,count) => (1.0 * sum)/count}
  Avg_Stud.foreach(println)

  val baseRDD6 =
    spark.sparkContext.textFile("C:/Users/admin/Desktop/Assignment_to_be
submitted/s19/Dataset.txt").map(x=>((x.split(",")(0),x.split(",")(2)),x.split(
      ",")(3).toInt))
  baseRDD6.foreach(println)
  val grade = baseRDD6.mapValues(x=>(x,1))
  grade.foreach(println)
  val gradeReduce = grade.reduceByKey((x,y)=> (x._1+y._1,x._2+y._2))
  gradeReduce.foreach(println)
  val gradeAvg = gradeReduce.mapValues{case(sum,count) => (1.0*sum)/count}
  gradeAvg.foreach(println)

  val flatgradeAvg = gradeAvg.map(x=> x._1._1 + "," + x._2.toDouble)
  flatgradeAvg.foreach(println)
  val flatAvg_Stud = Avg_Stud.map(x=> x._1 + "," + x._2)
  flatAvg_Stud.foreach(println)
  val commanval = flatgradeAvg.intersection(flatAvg_Stud)
  commanval.foreach(println)

}

}
```

Output:

```
18/09/14 13:19:59 INFO deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
(Mathew,45)
(Mathew,55)
(Mark,23)
(Mark,76)
(John,14)
(John,74)
(Lisa,24)
(Lisa,86)
(Andrew,34)
(Andrew,26)
(Andrew,74)
(Mathew,55)
(Mathew,87)
(Mark,92)
(Mark,12)
(John,67)
(John,35)
(Lisa,24)
(Lisa,98)
(Andrew,23)
(Andrew,44)
(Andrew,77)
18/09/14 13:19:59 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1085 bytes result sent to driver
18/09/14 13:19:59 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 332 ms on localhost (executor driver) (1/1)
18/09/14 13:19:59 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
```

```
18/09/14 13:19:59 INFO HadoopRDD: Input split: file:/C:/Users/admin/Desktop/Assignment_to_be submitted/s19/Dataset.txt:0+625
(Mathew,(45,1))
(Mathew,(55,1))
(Mark,(23,1))
(Mark,(76,1))
(John,(14,1))
(John,(74,1))
(Lisa,(24,1))
(Lisa,(86,1))
(Andrew,(34,1))
(Andrew,(26,1))
(Andrew,(74,1))
(Mathew,(55,1))
(Mathew,(87,1))
(Mark,(92,1))
(Mark,(12,1))
(John,(67,1))
(John,(35,1))
(Lisa,(24,1))
(Lisa,(98,1))
(Andrew,(23,1))
(Andrew,(44,1))
(Andrew,(77,1))
18/09/14 13:19:59 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 995 bytes result sent to driver
18/09/14 13:19:59 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 30 ms on localhost (executor driver) (1/1)
18/09/14 13:19:59 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/09/14 13:19:59 INFO DAGScheduler: ResultStage 1 (foreach at Intersection.scala:23) finished in 0.030 s
```

```
18/09/14 13:19:59 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 20 ms
(Mark,(203,4))
(Andrew,(278,6))
(Mathew,(197,3))
(Mathew,(45,1))
(John,(190,4))
(Lisa,(232,4))
18/09/14 13:19:59 INFO Executor: Finished task 0.0 in stage 3.0 (TID 3). 1722 bytes result sent to driver
18/09/14 13:19:59 INFO DAGScheduler: ResultStage 3 (foreach at Intersection.scala:25) finished in 0.100 s
```

```
18/09/14 13:19:59 INFO Executor: Running task 0.0 in stage 5.0 (TID 4)
(Mark,50.75)
(Andrew,46.333333333333336)
(Mathew,65.66666666666667)
(Mathew,45.0)
(John,47.5)
(Lisa,58.0)
18/09/14 13:19:59 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
```

```
18/09/14 13:19:59 INFO HadoopRDD: Input split: file:/C:/Users/admin/Desktop/Assignment_to_be submitted/s19/Dataset.txt:0+625
((Mathew,grade-3),45)
((Mathew,grade-2),55)
((Mark,grade-2),23)
((Mark,grade-1),76)
((John,grade-1),14)
((John,grade-2),74)
((Lisa,grade-1),24)
((Lisa,grade-3),86)
((Andrew,grade-1),34)
18/09/14 13:19:59 INFO Executor: Finished task 0.0 in stage 6.0 (TID 5). 995 bytes result sent to driver
18/09/14 13:19:59 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 5) in 25 ms on localhost (executor driver) (1/1)
18/09/14 13:19:59 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool
18/09/14 13:19:59 INFO DAGScheduler: ResultStage 6 (foreach at Intersection.scala:32) finished in 0.026 s
18/09/14 13:19:59 INFO DAGScheduler: Job 4 finished: foreach at Intersection.scala:32, took 0.052391 s
((Andrew,grade-3),26)
((Andrew,grade-1),74)
((Mathew,grade-2),55)
((Mathew,grade-2),87)
((Mark,grade-1),92)
((Mark,grade-2),12)
((John,grade-1),67)
((John,grade-1),35)
((Lisa,grade-2),24)
((Lisa,grade-2),98)
((Andrew,grade-1),23)
((Andrew,grade-3),44)
((Andrew,grade-2),77)
18/09/14 13:19:59 INFO SparkContext: Starting job: foreach at Intersection.scala:34
18/09/14 13:19:59 INFO DAGScheduler: Got job 5 (foreach at Intersection.scala:34) with 1 output partitions
18/09/14 13:19:59 INFO DAGScheduler: Final stage: ResultStage 7 (foreach at Intersection.scala:34)
```

```
18/09/14 13:19:59 INFO HadoopRDD: Input split: file:/C:/Users/admin/Desktop/Assignment_to_be submitted/s19/Dataset.txt:0+625
((Mathew,grade-3),(45,1))
18/09/14 13:19:59 INFO Executor: Finished task 0.0 in stage 7.0 (TID 6). 908 bytes result sent to driver
((Mathew,grade-2),(55,1))
((Mark,grade-2),(23,1))
((Mark,grade-1),(76,1))
((John,grade-1),(14,1))
((John,grade-2),(74,1))
((Lisa,grade-1),(24,1))
((Lisa,grade-3),(86,1))
((Andrew,grade-1),(34,1))
((Andrew,grade-3),(26,1))
((Andrew,grade-1),(74,1))
((Mathew,grade-2),(55,1))
((Mathew,grade-2),(87,1))
((Mark,grade-1),(92,1))
((Mark,grade-2),(12,1))
((John,grade-1),(67,1))
((John,grade-1),(35,1))
((Lisa,grade-2),(24,1))
((Lisa,grade-2),(98,1))
((Andrew,grade-1),(23,1))
((Andrew,grade-3),(44,1))
((Andrew,grade-2),(77,1))
18/09/14 13:20:00 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 6) in 30 ms on localhost (executor driver) (1/1)
18/09/14 13:20:00 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed, from pool
18/09/14 13:20:00 INFO DAGScheduler: ResultStage 7 (foreach at Intersection.scala:34) finished in 0.030 s
```

```
18/09/14 13:20:00 INFO Executor: Finished task 0.0 in stage 9.0 (TID 8). 1474 bytes result sent to driver
((Lisa,grade-1),(24,1))
((Mark,grade-2),(35,2))
((Lisa,grade-2),(122,2))
((Andrew,grade-2),(77,1))
((Mathew,grade-3),(45,1))
((Andrew,grade-1),(131,3))
((Lisa,grade-3),(86,1))
((John,grade-1),(116,3))
((John,grade-2),(74,1))
((Mark,grade-1),(168,2))
((Andrew,grade-3),(70,2))
((Mathew,grade-2),(197,3))
18/09/14 13:20:00 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 8) in 30 ms on localhost (executor driver) (1/1)
```

```
18/09/14 13:20:00 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
((Lisa,grade-1),24.0)
((Mark,grade-2),17.5)
((Lisa,grade-2),61.0)
((Andrew,grade-2),77.0)
((Mathew,grade-3),45.0)
((Andrew,grade-1),43.666666666666664)
((Lisa,grade-3),86.0)
((John,grade-1),38.666666666666664)
((John,grade-2),74.0)
((Mark,grade-1),84.0)
((Andrew,grade-3),35.0)
((Mathew,grade-2),65.66666666666667)
18/09/14 13:20:00 INFO Executor: Finished task 0.0 in stage 11.0 (TID 9). 1545 bytes result sent to driver
18/09/14 13:20:00 INFO TaskSetManager: Finished task 0.0 in stage 11.0 (TID 9) in 10 ms on localhost (executor drive
```

```
18/09/14 13:20:00 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
Lisa,24.0
18/09/14 13:20:00 INFO Executor: Finished task 0.0 in stage 13.0 (TID 10). 1545 bytes result sent to driver
Mark,17.5
18/09/14 13:20:00 INFO TaskSetManager: Finished task 0.0 in stage 13.0 (TID 10) in 10 ms on localhost (execu
Lisa,61.0
Andrew,77.0
18/09/14 13:20:00 INFO TaskSchedulerImpl: Removed TaskSet 13.0, whose tasks have all completed, from pool
Mathew,45.0
Andrew,43.666666666666664
18/09/14 13:20:00 INFO DAGScheduler: ResultStage 13 (foreach at Intersection.scala:41) finished in 0.010 s
Lisa,86.0
18/09/14 13:20:00 INFO DAGScheduler: Job 8 finished: foreach at Intersection.scala:41, took 0.027511 s
John,38.666666666666664
John,74.0
Mark,84.0
Andrew,35.0
Mathew,65.66666666666667
18/09/14 13:20:00 INFO SparkContext: Starting job: foreach at Intersection.scala:43
```

```
18/09/14 13:20:00 INFO Executor: Running task 0.0 in stage 15.0 (TID 11)
Mark,50.75
Andrew,46.333333333333336
Mathew,65.66666666666667
Mathew,45.0
John,47.5
Lisa,58.0
18/09/14 13:20:00 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/14 13:20:00 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
```

```
18/09/14 13:20:00 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
Mathew,45.0
Mathew,65.66666666666667
18/09/14 13:20:00 INFO Executor: Finished task 0.0 in stage 20.0 (TID 14). 1545 bytes result sent to driver
18/09/14 13:20:00 INFO TaskSetManager: Finished task 0.0 in stage 20.0 (TID 14) in 60 ms on localhost (executor driver) (1,
```