Session 19:

RDD DEEP DIVE

Assignment 1

Task 1

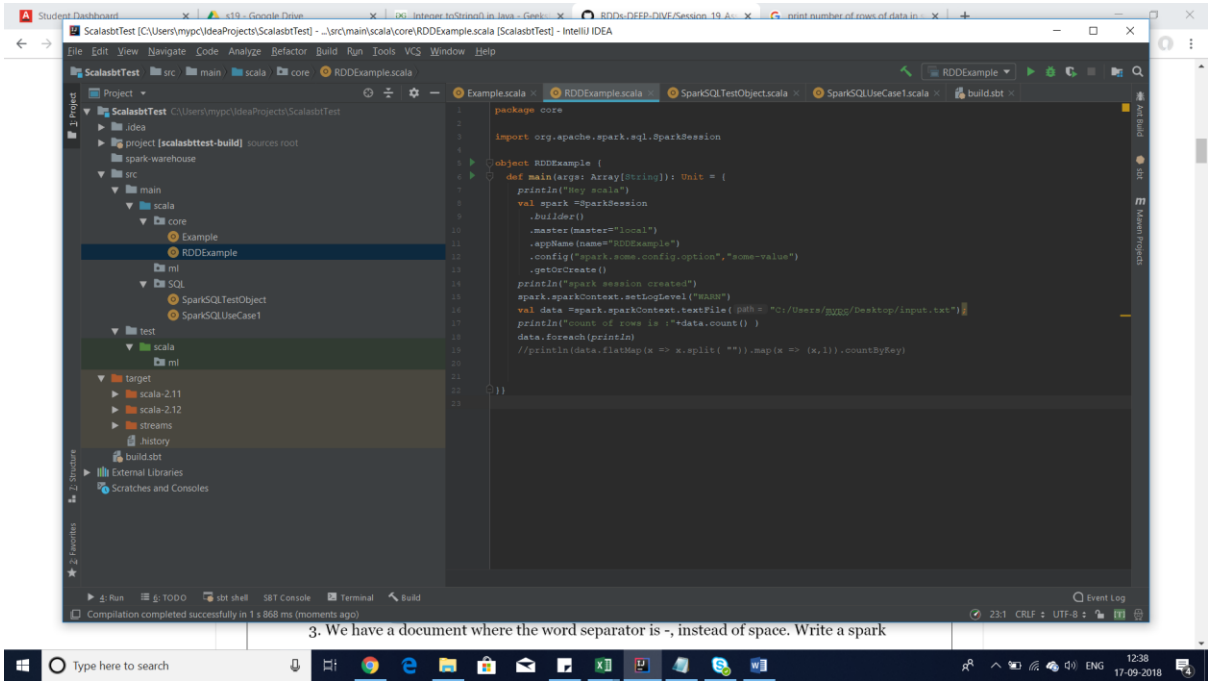
1.Write a program to read a text file and print the number of rows of data in the document.

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

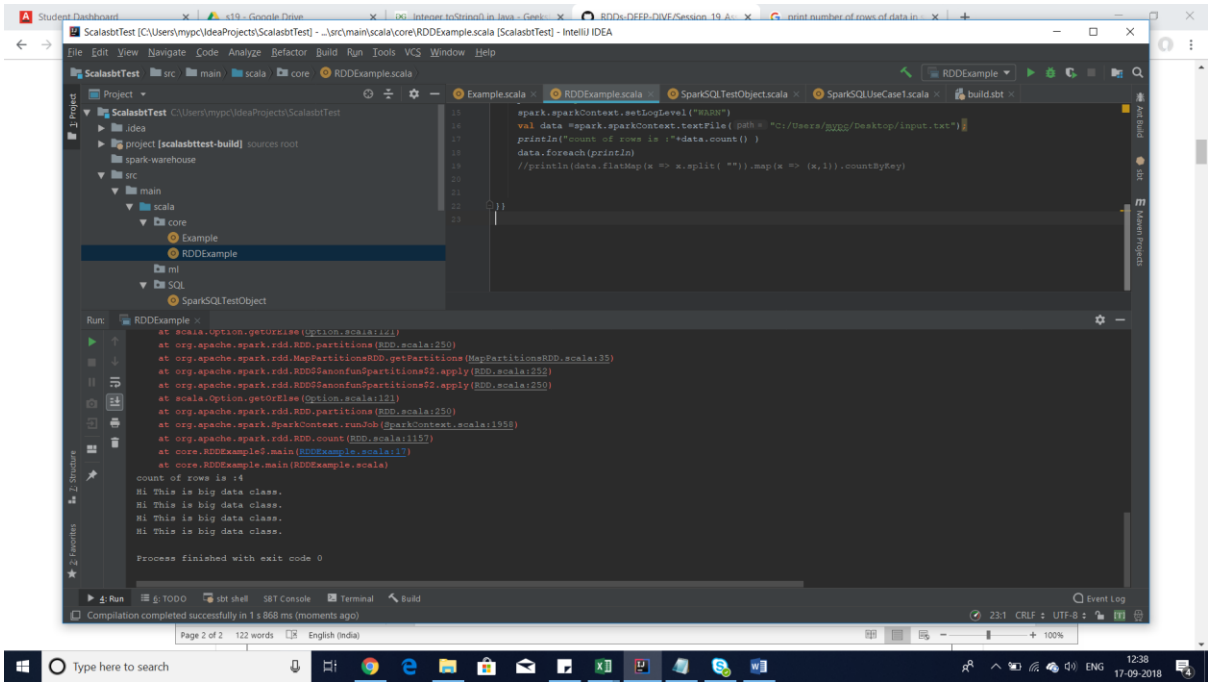
3. We have a document where the word separator is -, instead of space. Write a spark code, to obtain the count of the total number of words present in the document.

```
package core

import org.apache.spark.sql.SparkSession

object RDDExample {
  def main(args: Array[String]): Unit = {
    println("Hey scala")
    val spark =SparkSession
      .builder()
      .master(master="local")
      .appName(name="RDDExample")
      .config("spark.some.config.option","some-value")
      .getOrCreate()
    println("spark session created")
    spark.sparkContext.setLogLevel("WARN")
    val data =spark.sparkContext.textFile("C:/Users/mypc/Desktop/input.txt");
    println("count of rows is :"+data.count() )
    data.foreach(println)
    //println(data.flatMap(x => x.split( "")).map(x => (x,1)).countByKey)
```

}}



3. We have a document where the word separator is -, instead of space. Write a spark

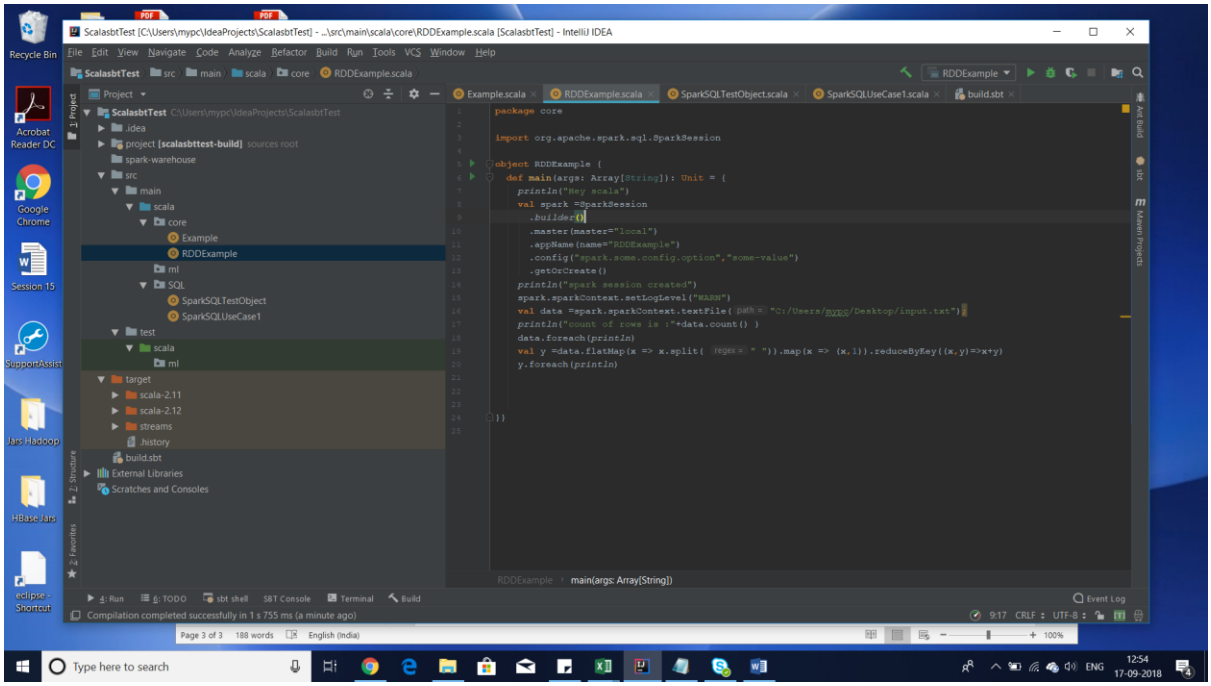


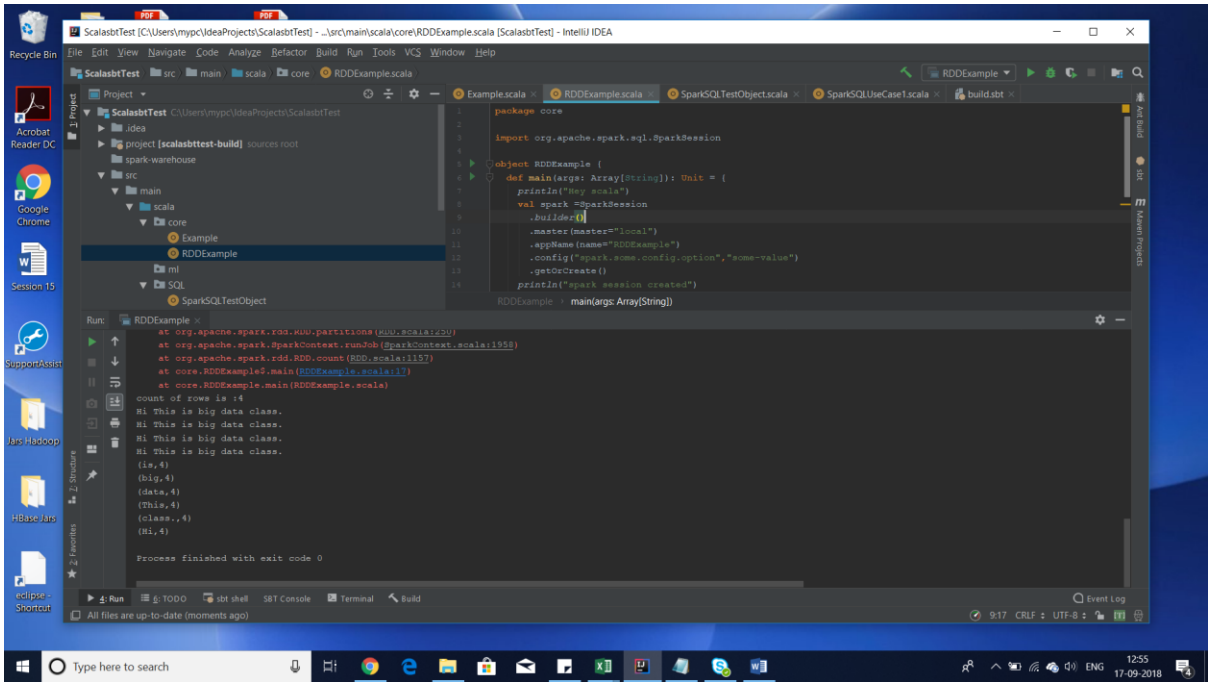
2. Write a program to read a text file and print the number of words in the document.

package core

import org.apache.spark.sql.SparkSession

```
object RDDExample {
  def main(args: Array[String]): Unit = {
    println("Hey scala")
    val spark =SparkSession
      .builder()
      .master(master="local")
      .appName(name="RDDExample")
      .config("spark.some.config.option","some-value")
      .getOrCreate()
    println("spark session created")
    spark.sparkContext.setLogLevel("WARN")
    val data =spark.sparkContext.textFile("C:/Users/mypc/Desktop/input.txt");
    println("count of rows is :"+data.count() )
    data.foreach(println)
    val y =data.flatMap(x => x.split( " ")).map(x => (x,1)).reduceByKey((x,y)=>x+y)
    y.foreach(println)



}}
```



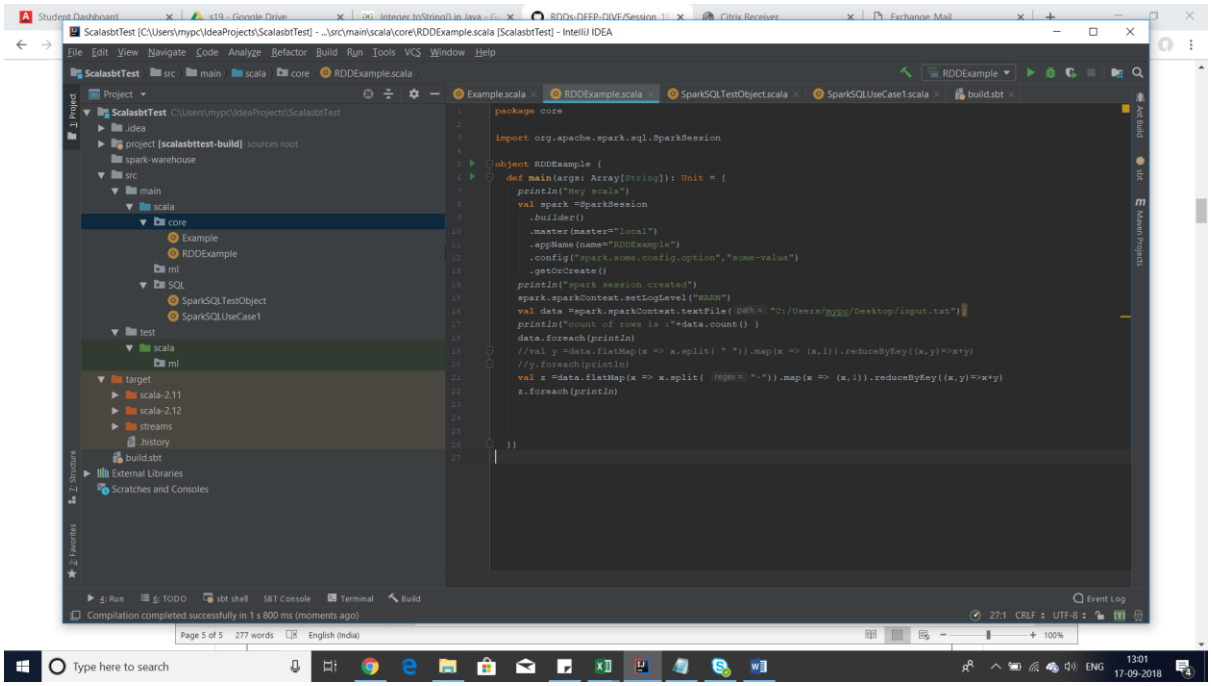


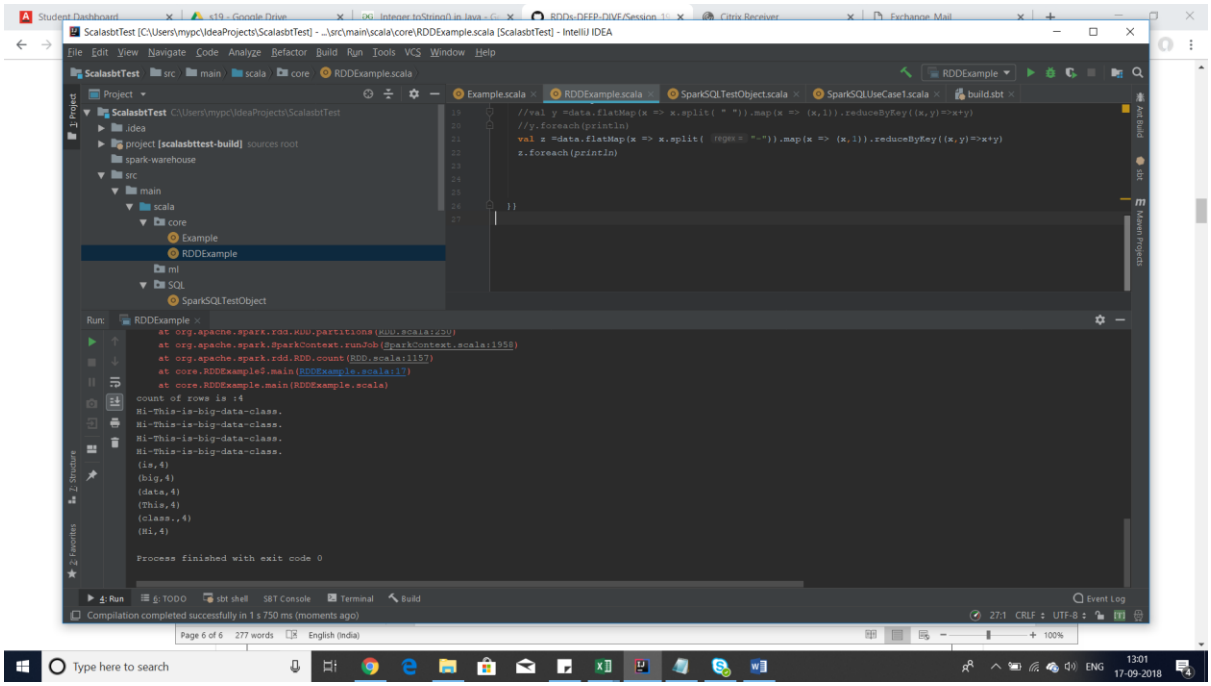
3. We have a document where the word separator is -, instead of space. Write a spark code, to obtain the count of the total number of words present in the document.

package core

import org.apache.spark.sql.SparkSession

```
object RDDExample {
  def main(args: Array[String]): Unit = {
    println("Hey scala")
    val spark =SparkSession
      .builder()
      .master(master="local")
      .appName(name="RDDExample")
      .config("spark.some.config.option","some-value")
      .getOrCreate()
    println("spark session created")
    spark.sparkContext.setLogLevel("WARN")
    val data =spark.sparkContext.textFile("C:/Users/mypc/Desktop/input.txt");
    println("count of rows is :"+data.count() )
    data.foreach(println)
    //val y =data.flatMap(x => x.split( " ")).map(x => (x,1)).reduceByKey((x,y)=>x+y)
    //y.foreach(println)
    val z =data.flatMap(x => x.split( "-")).map(x => (x,1)).reduceByKey((x,y)=>x+y)
    z.foreach(println)



}}
```





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.

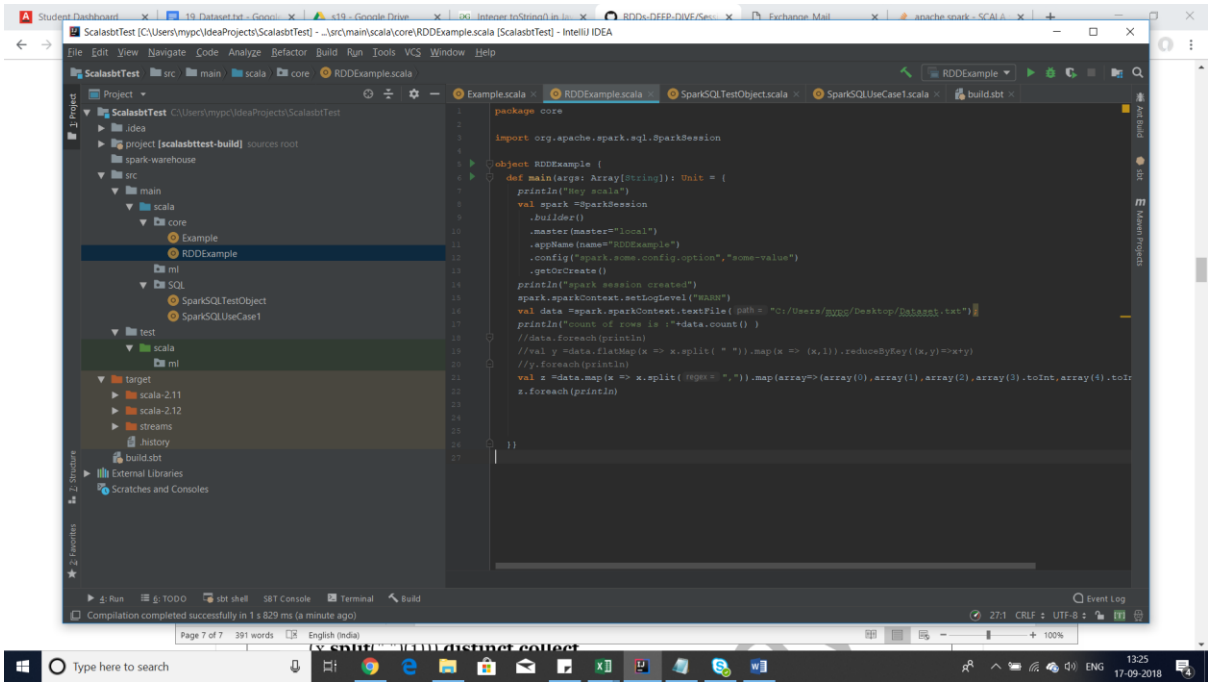
3. What is the distinct number of subjects present in the entire school

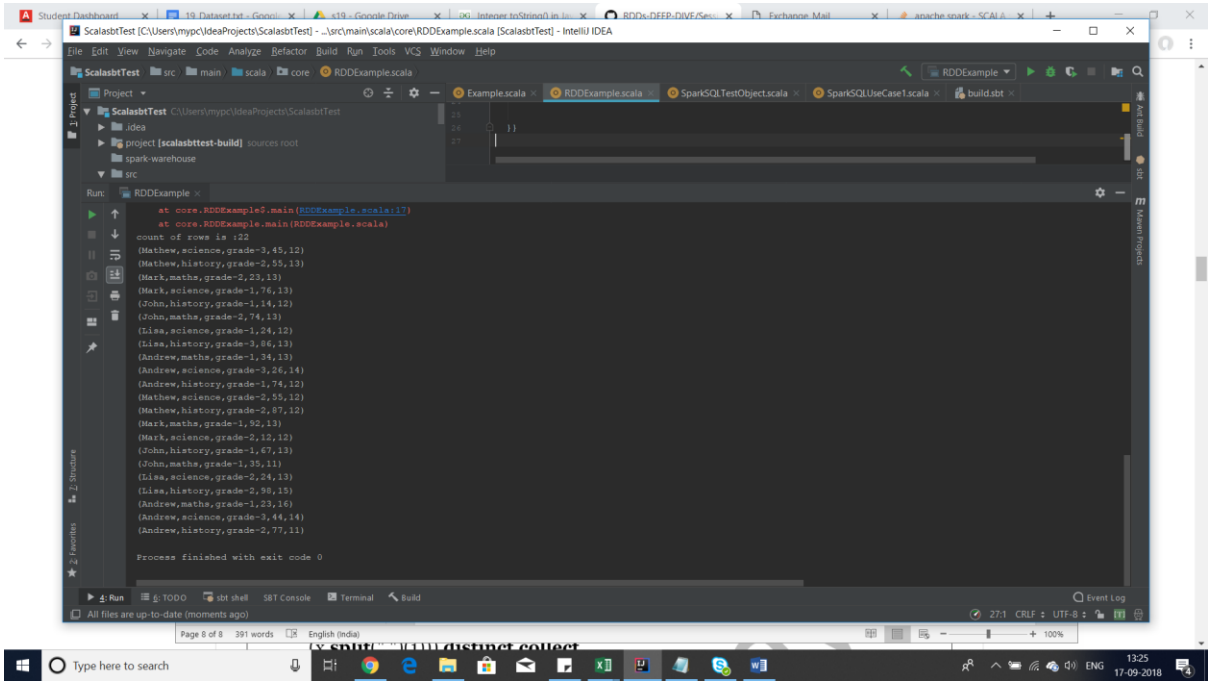
4. What is the count of the number of students in the school, whose name is Mathew and

marks is 55

Solution:

```
package core

import org.apache.spark.sql.SparkSession

object RDDExample {
  def main(args: Array[String]): Unit = {
    println("Hey scala")
    val spark =SparkSession
      .builder()
      .master(master="local")
      .appName(name="RDDExample")
      .config("spark.some.config.option","some-value")
      .getOrCreate()
    println("spark session created")
    spark.sparkContext.setLogLevel("WARN")
    val data =spark.sparkContext.textFile("C:/Users/mypc/Desktop/Dataset.txt");
    println("count of rows is :"+data.count() )
    //data.foreach(println)
    //val y =data.flatMap(x => x.split( " ")).map(x => (x,1)).reduceByKey((x,y)=>x+y)
    //y.foreach(println)
    val z =data.map(x =>
x.split(",")).map(array=>(array(0),array(1),array(2),array(3).toInt,array(4).toInt)).collect
    z.foreach(println)



  }}
```



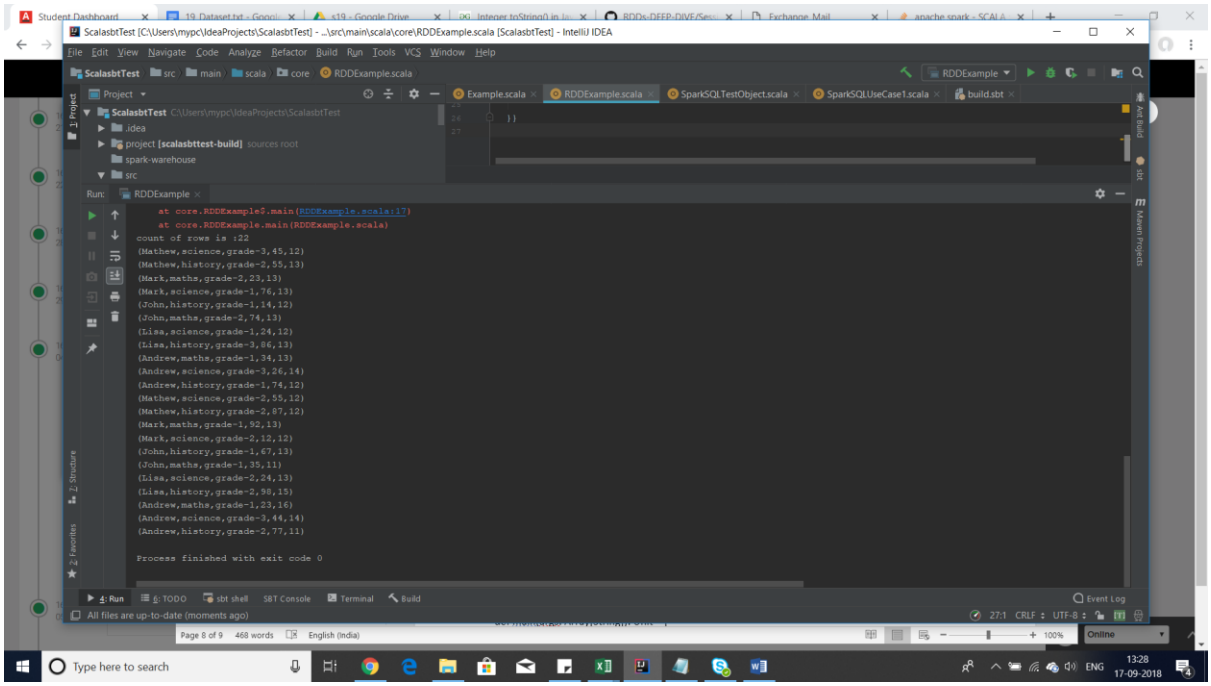


2. Find the count of total number of rows present.

Solution:We can find the count of rows using

“println("count of rows is :"+data.count())” in the program

```
object RDDExample {
  def main(args: Array[String]): Unit = {
    println("Hey scala")
    val spark =SparkSession
      .builder()
      .master(master="local")
      .appName(name="RDDExample")
      .config("spark.some.config.option","some-value")
      .getOrCreate()
    println("spark session created")
    spark.sparkContext.setLogLevel("WARN")
    val data =spark.sparkContext.textFile("C:/Users/mypc/Desktop/Dataset.txt");
    println("count of rows is :"+data.count() )
    val z =data.map(x =>
x.split(",")).map(array=>(array(0),array(1),array(2),array(3).toInt,array(4).toInt)).collect
    z.foreach(println)



  }}
```



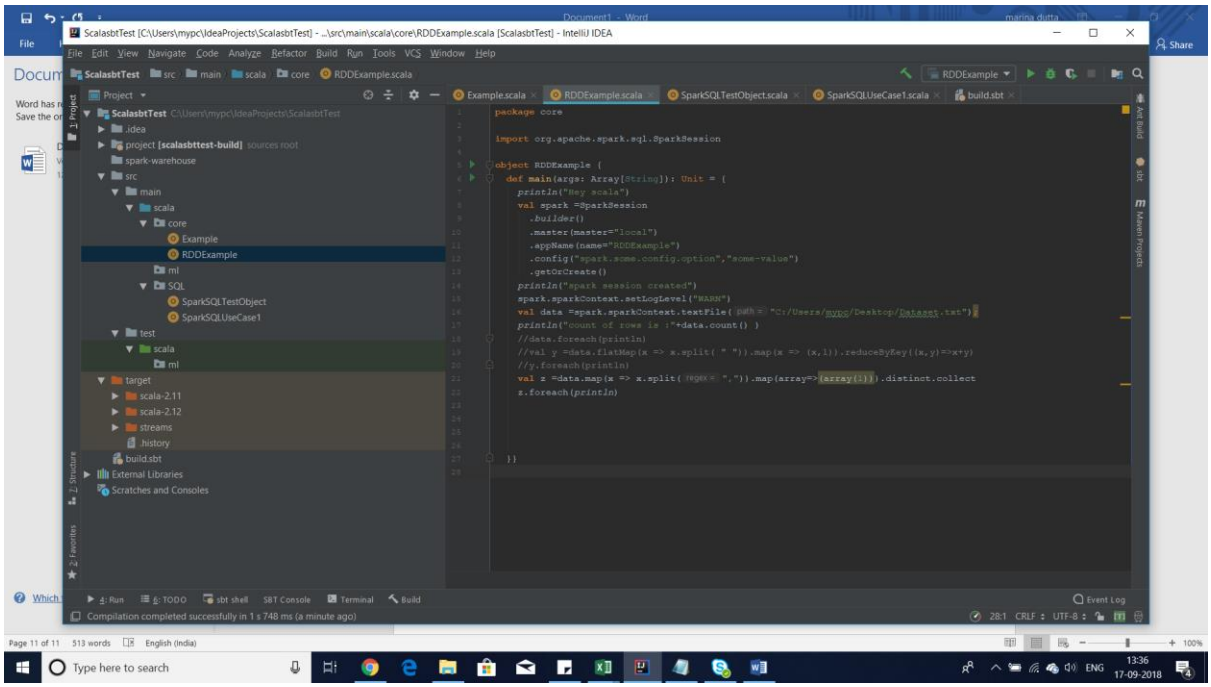
What is the distinct number of subjects present in the entire school

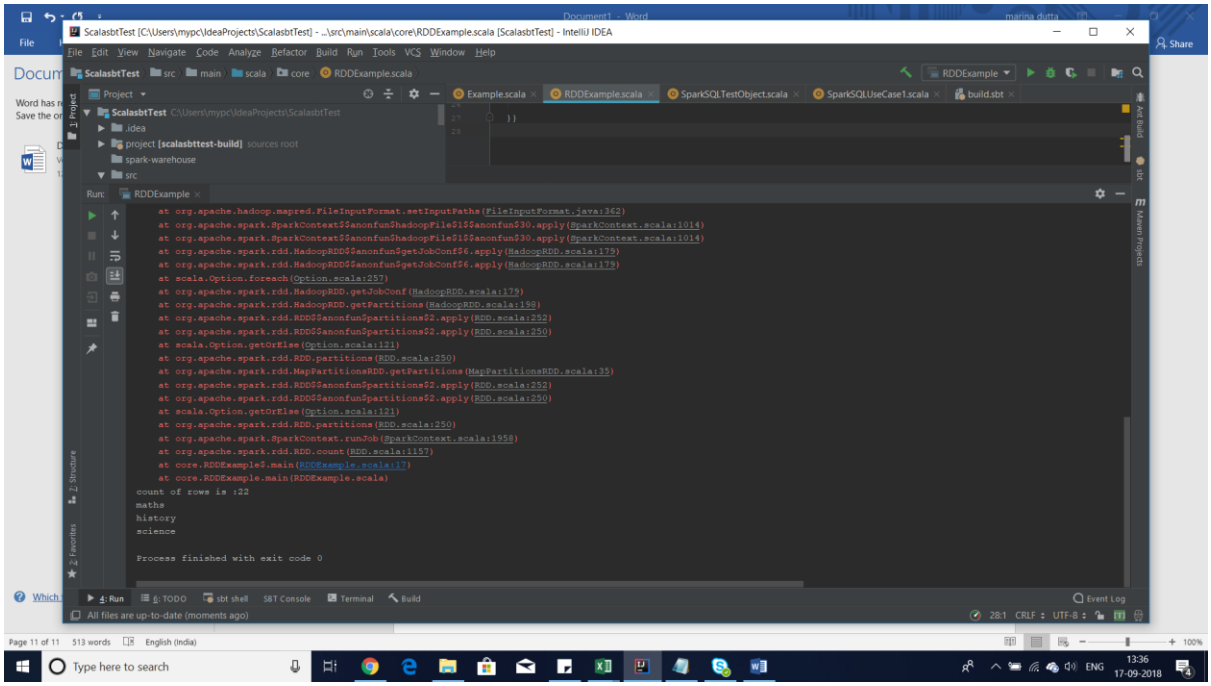
Solution:

```
package core


import org.apache.spark.sql.SparkSession


object RDDExample {
  def main(args: Array[String]): Unit = {
    println("Hey scala")
    val spark =SparkSession
      .builder()
      .master(master="local")
      .appName(name="RDDExample")
      .config("spark.some.config.option","some-value")
      .getOrCreate()
    println("spark session created")
    spark.sparkContext.setLogLevel("WARN")
```

```
val data =spark.sparkContext.textFile("C:/Users/mypc/Desktop/Dataset.txt");
println("count of rows is :"+data.count() )
val z =data.map(x => x.split(",")).map(array=>(array(1))).distinct.collect
z.foreach(println)
```

}}





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

Solution:

package core

import org.apache.spark.sql.SparkSession

object RDDExample {

  def main(args: Array[String]): Unit = {

    println("Hey scala")

    val spark =SparkSession

      .builder()

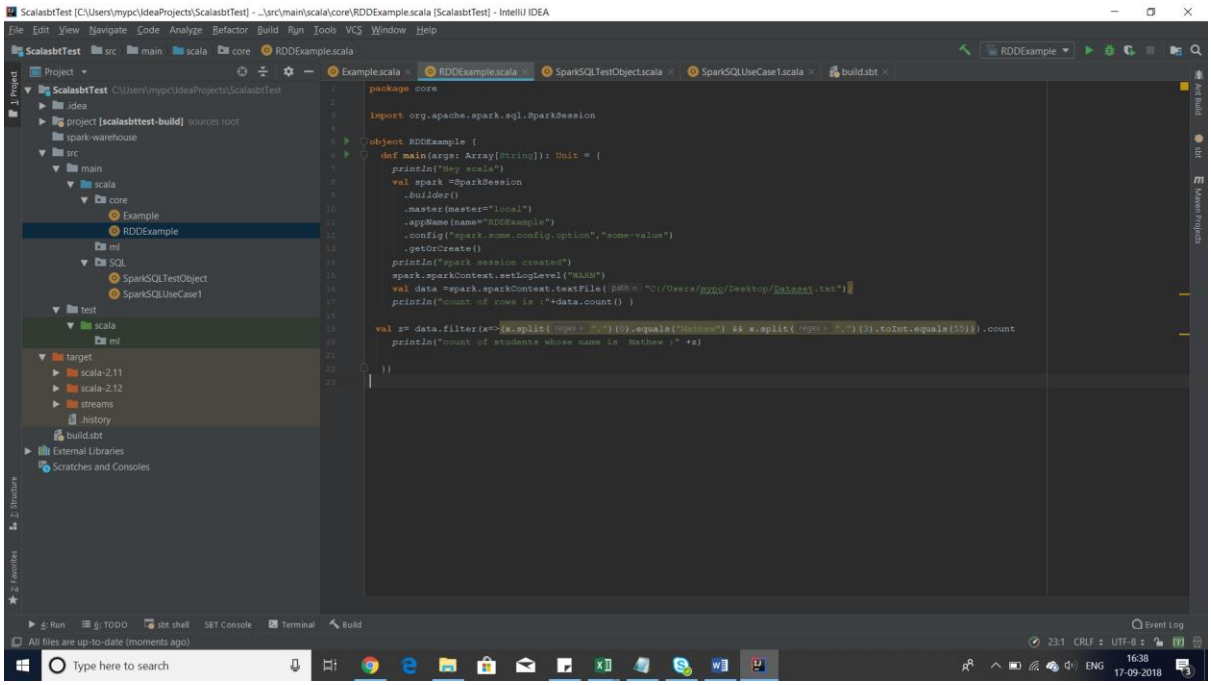
      .master(master="local")

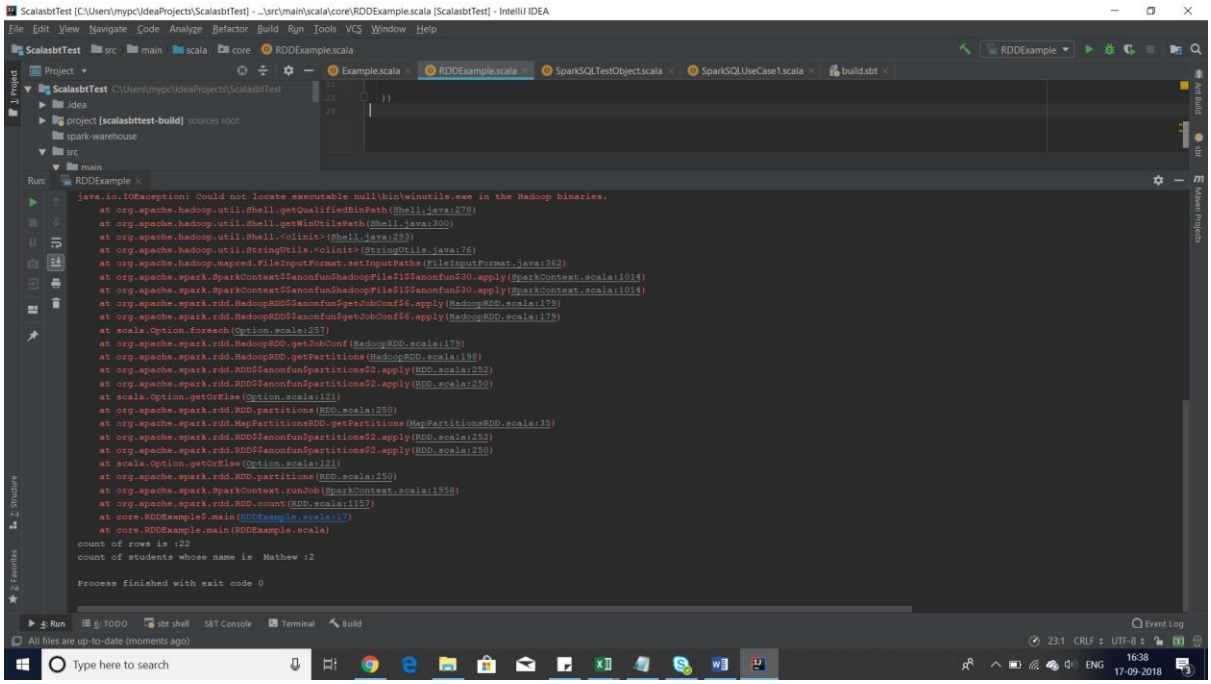
      .appName(name="RDDExample")

      .config("spark.some.config.option","some-value")

      .getOrCreate()

```
  println("spark session created")
  spark.sparkContext.setLogLevel("WARN")
  val data =spark.sparkContext.textFile("C:/Users/mypc/Desktop/Dataset.txt");
  println("count of rows is :"+data.count() )

val z= data.filter(x=>(x.split(",")(0).equals("Mathew") && x.split(",")(3).toInt.equals(55))).count
  println("count of students whose name is  Mathew :" +z)

 }}
```

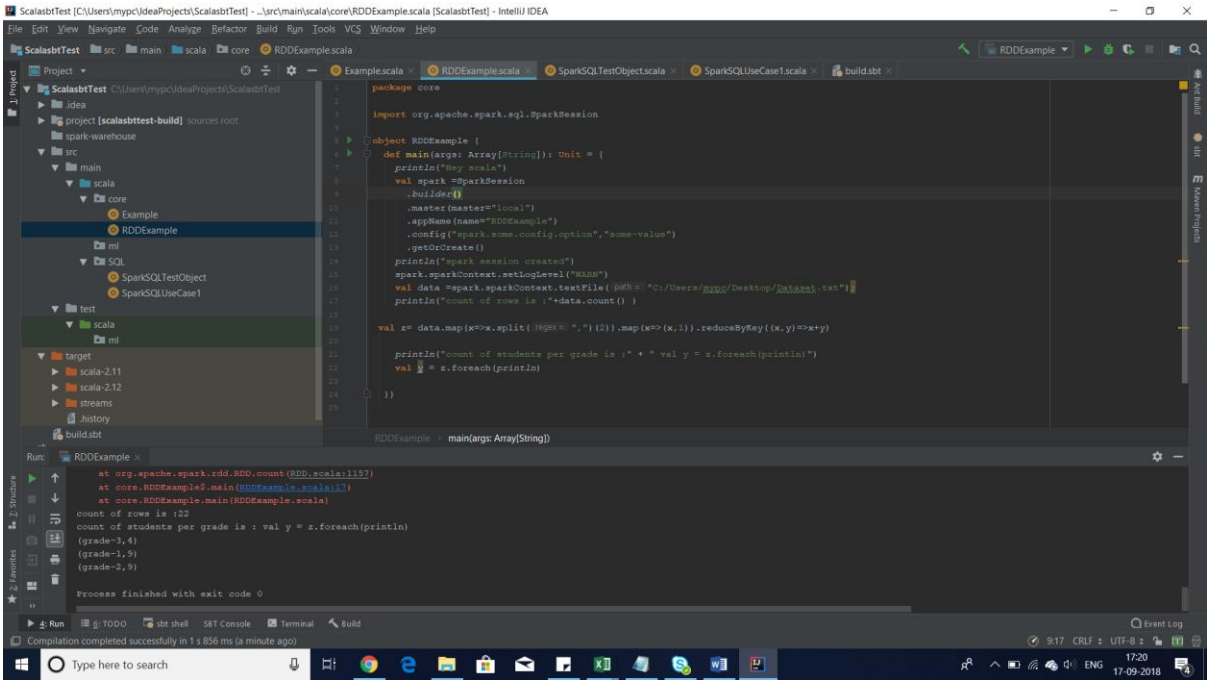


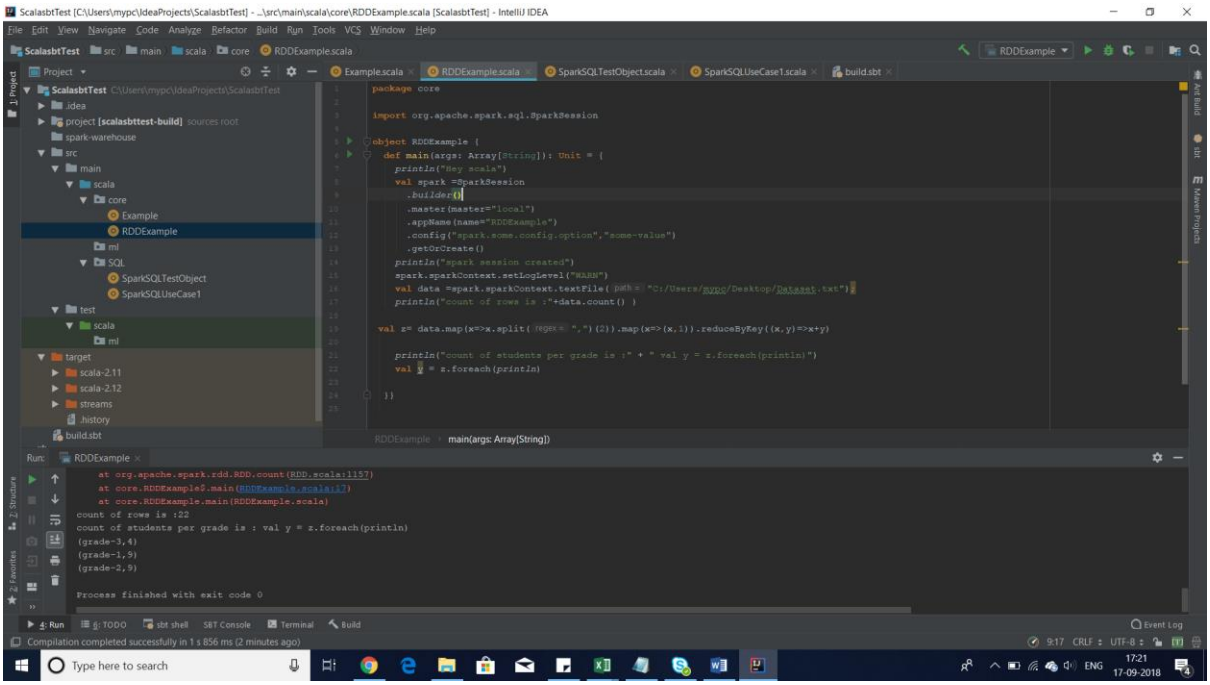


Problem Statement 2:

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

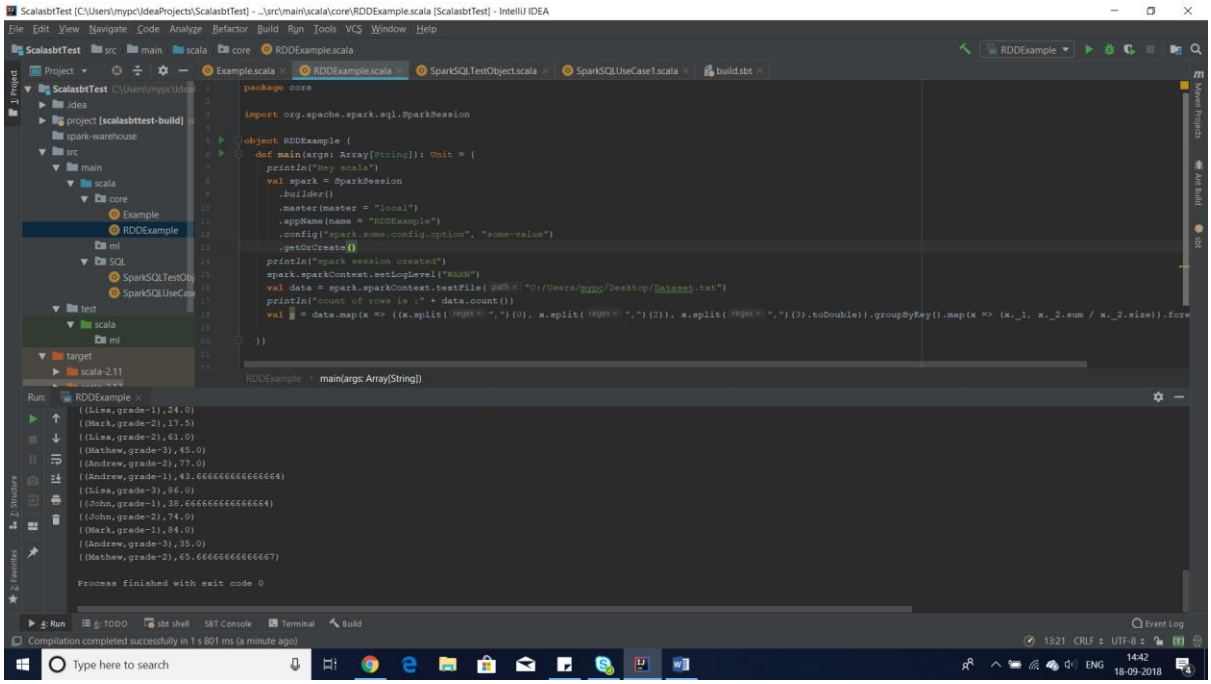
```
import org.apache.spark.sql.SparkSession

object RDDExample {
 def main(args: Array[String]): Unit = {
  println("Hey scala")
  val spark =SparkSession
   .builder()
   .master(master="local")
   .appName(name="RDDExample")
   .config("spark.some.config.option","some-value")
   .getOrCreate()
  println("spark session created")
  spark.sparkContext.setLogLevel("WARN")
  val data =spark.sparkContext.textFile("C:/Users/mypc/Desktop/Dataset.txt");
  println("count of rows is :"+data.count() )

 val z= data.map(x=>x.split(",")(2)).map(x=>(x,1)).reduceByKey((x,y)=>x+y)

  println("count of students per grade is :" + " val y = z.foreach(println)")
  val y = z.foreach(println)

 }}
```

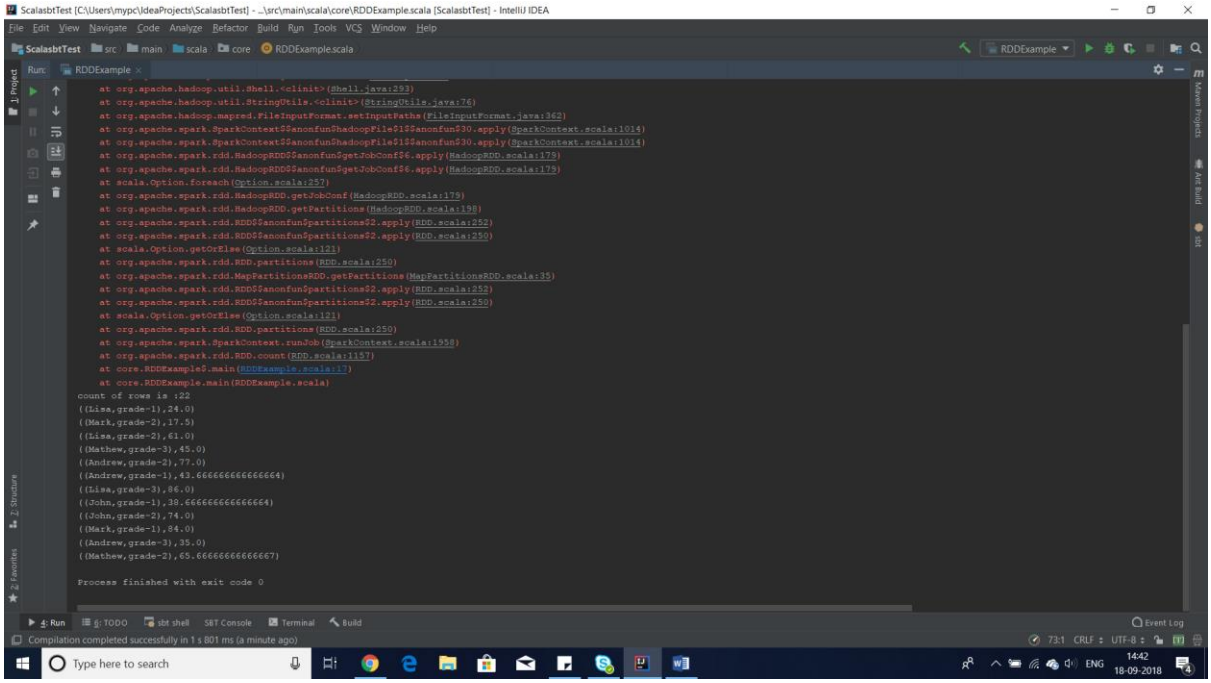




2. Find the average of each student (Note - Mathew is grade-1, is different from Mathew in some other grade!)

```
import org.apache.spark.sql.SparkSession


object RDDExample {
  def main(args: Array[String]): Unit = {
    println("Hey scala")
    val spark = SparkSession
      .builder()
      .master(master = "local")
      .appName(name = "RDDExample")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("spark session created")
    spark.sparkContext.setLogLevel("WARN")
    val data = spark.sparkContext.textFile("C:/Users/mypc/Desktop/Dataset.txt")
    println("count of rows is :" + data.count())
    val z = data.map(x => ((x.split(",")(0), x.split(",")(2)), x.split(",")(3).toDouble)).groupByKey().map(x
=> (x._1, x._2.sum / x._2.size)).foreach(println)


  }}
```

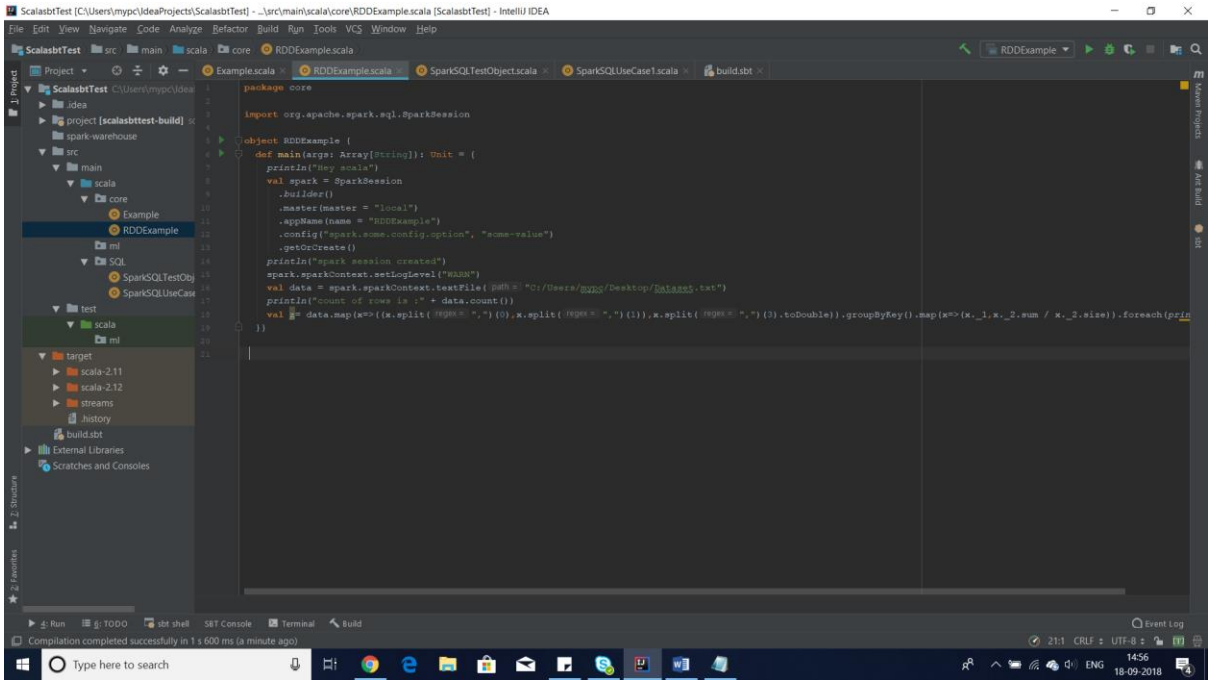


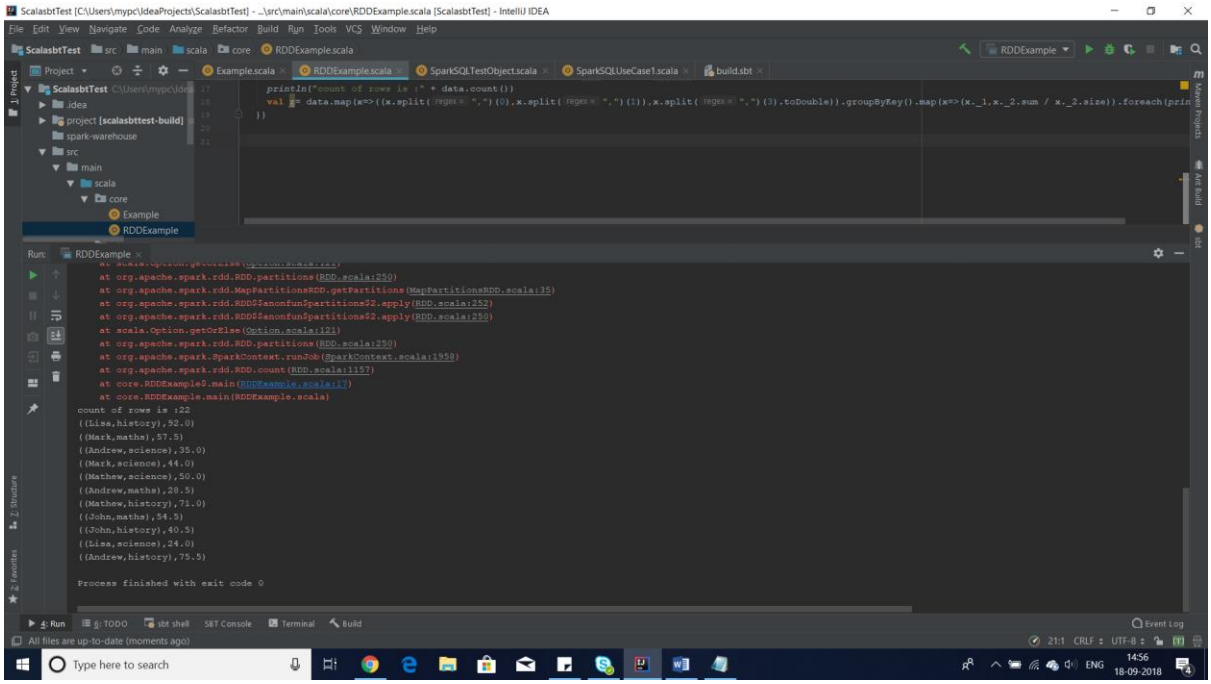


3. What is the average score of students in each subject across all grades?

```
import org.apache.spark.sql.SparkSession

object RDDExample {
```

```
 def main(args: Array[String]): Unit = {
   println("Hey scala")
   val spark = SparkSession
     .builder()
     .master(master = "local")
     .appName(name = "RDDExample")
     .config("spark.some.config.option", "some-value")
     .getOrCreate()
   println("spark session created")
   spark.sparkContext.setLogLevel("WARN")
   val data = spark.sparkContext.textFile("C:/Users/mypc/Desktop/Dataset.txt")
   println("count of rows is :" + data.count())
   val z=
data.map(x=>((x.split(",")(0),x.split(",")(1)),x.split(",")(3).toDouble)).groupByKey().map(x=>(x._1,x._2.
sum / x._2.size)).foreach(println)
 }}
```

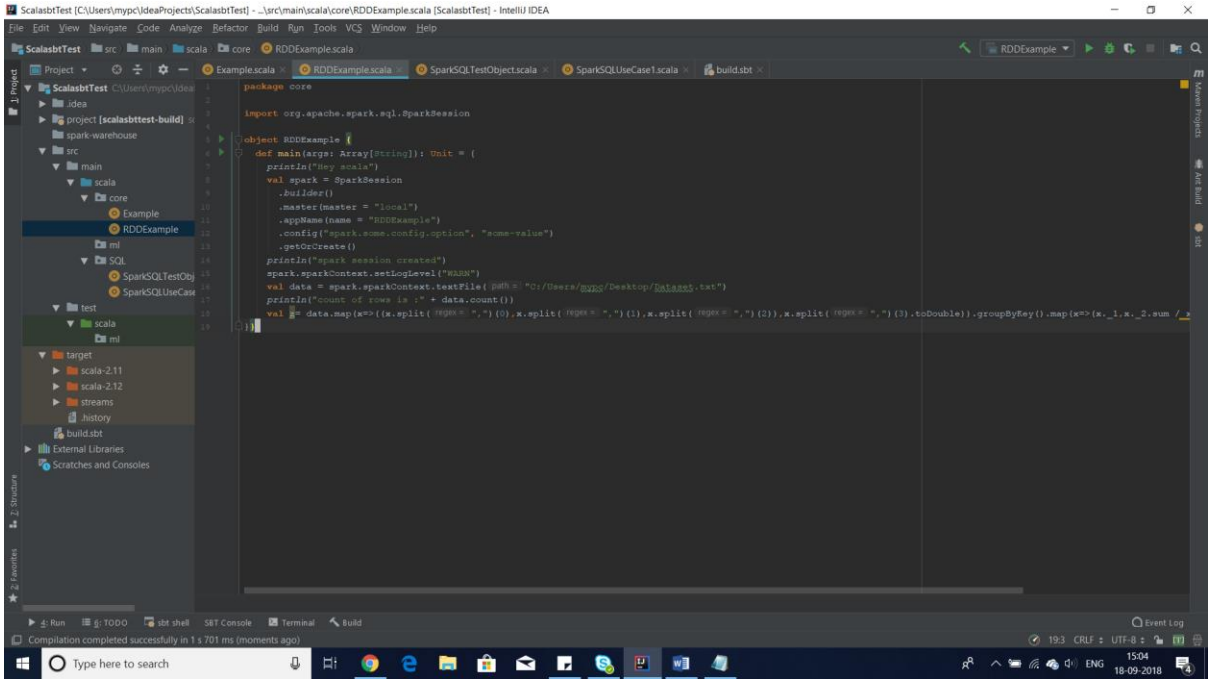


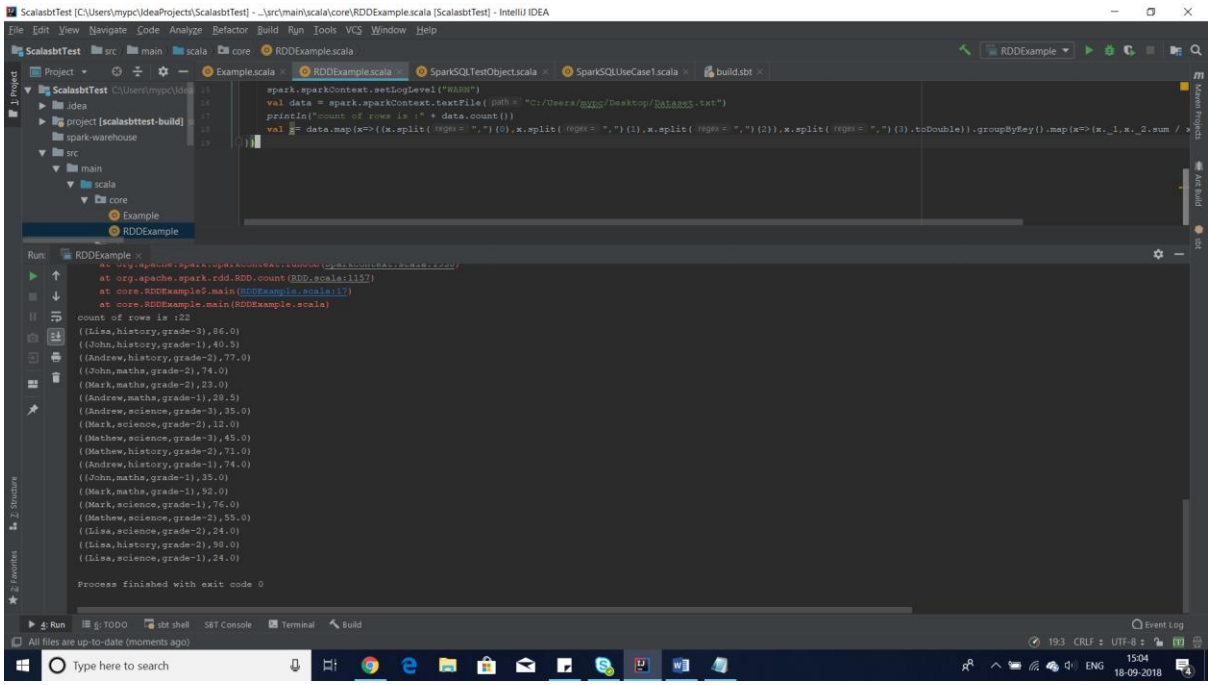


4. What is the average score of students in each subject per grade?

```
package core

import org.apache.spark.sql.SparkSession

object RDDExample {
  def main(args: Array[String]): Unit = {
    println("Hey scala")
    val spark = SparkSession
      .builder()
      .master(master = "local")
      .appName(name = "RDDExample")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("spark session created")
    spark.sparkContext.setLogLevel("WARN")
```

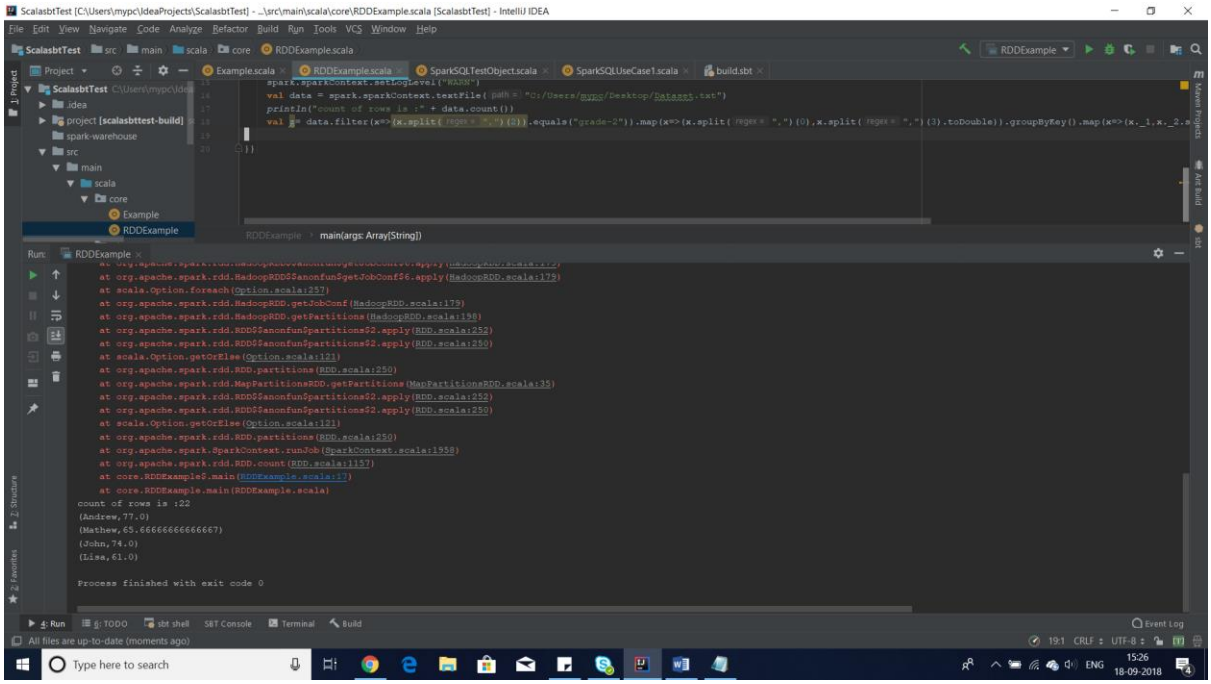
```
val data = spark.sparkContext.textFile("C:/Users/mypc/Desktop/Dataset.txt")

println("count of rows is :" + data.count())

val z=
data.map(x=>((x.split(",")(0),x.split(",")(1),x.split(",")(2)),x.split(",")(3).toDouble)).groupByKey().map(
x=>(x._1,x._2.sum / x._2.size)).foreach(println)

}}
```

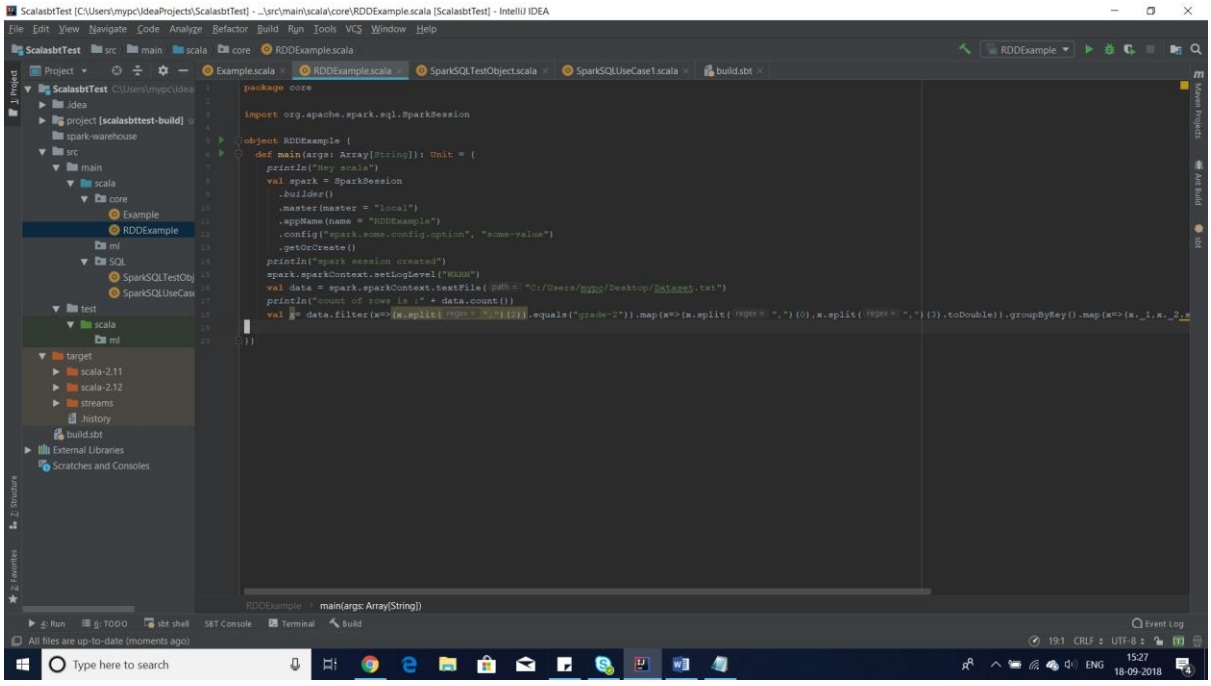




5. For all students in grade-2, how many have average score greater than 50?

```
ackage core

import org.apache.spark.sql.SparkSession

object RDDExample {
  def main(args: Array[String]): Unit = {
    println("Hey scala")
    val spark = SparkSession
      .builder()
      .master(master = "local")
      .appName(name = "RDDExample")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("spark session created")
    spark.sparkContext.setLogLevel("WARN")
    val data = spark.sparkContext.textFile("C:/Users/mypc/Desktop/Dataset.txt")
    println("count of rows is :" + data.count())
    val z= data.filter(x=>(x.split(",")(2)).equals("grade-
2")).map(x=>(x.split(",")(0),x.split(",")(3).toDouble)).groupByKey().map(x=>(x._1,x._2.sum /
x._2.size)).filter(x=>(x._2>50)).foreach(println)

}}
```





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 stud_name per grade