Assignment 20

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

object Assignment_20 {

  //Create a case class globally to be used inside the main method

  case class Transport(mode: String, amount: Int)

  case class User(id:Int, name:String,age :Int)

  case class Holiday (t_id:Int,Source:String,Destination: String,mode: String, distance:Int,
year:Int)

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

    //Create a spark session object

val spark = SparkSession

      .builder()

      .master("local")

      .appName("Spark SQL basic example")

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

      .getOrCreate()

    println("Spark Session Object created")


    import spark.implicits._                //to convert RDD into DataFrame

    val hl = spark.sparkContext

      .textFile("E:\\Avani\\Acadgild\\Assignment_20\\Data\\Holidays.txt")

      .map(_.split(","))

      .map(attributes => Holiday(attributes(0).trim.toInt,
attributes(1),attributes(2),attributes(3),attributes(4).trim.toInt,attributes(5).trim.toInt))

      .toDF()

    hl.show()

    println("Holiday data")
```

```
 val trans = spark.sparkContext
     .textFile("E:\\Avani\\Acadgild\\Assignment_20\\Data\\Transport.txt")
     .map(_.split(","))
     .map(attributes => Transport(attributes(0), attributes(1).trim.toInt))
     .toDF()
trans.show()
    println("Transport data")

  val user = spark.sparkContext
     .textFile("E:\\Avani\\Acadgild\\Assignment_20\\Data\\User.txt")
     .map(_.split(","))
     .map(attributes => User(attributes(0).trim.toInt,attributes(1),attributes(2).trim.toInt))
     .toDF()
user.show()
  println("User data")
  //Creating temporary tables
   hl.registerTempTable("holiday")
   user.registerTempTable("people")
   trans.registerTempTable("transport")
```

a. <u>What is the distribution of the total number of air-travellers per year?</u>

```
val total = spark.sql("SELECT year,COUNT(*) from holiday group by year")
total.show()
println("Task1 output")
```

b. <u>What is the total air distance covered by each user per year?</u>

```
val dist = spark.sql("SELECT t_id,year,sum(distance) from holiday group by t_id,year
order by t_id,year")
```

```
dist.show()
println("Task2 output")
```

c. Which user has travelled the largest distance till date?

```
val max_d = spark.sql("SELECT t_id,max(d) from (SELECT t_id, sum(distance) d from holiday group by t_id) group by t_id")
  max_d.show(1)
  println("Task3 output")
```

d. What is the most preferred destination for all users?

```
val ct=spark.sql("SELECT Destination from holiday group by 1 having count (Destination)= (select max(c) from (SELECT Destination,count(Destination) c from holiday group by Destination))")

ct.show()

println("Task4 output")
```

e. Which route is generating the most revenue per year?

```
val rev = spark.sql("SELECT h.source, h.destination, sum(h.distance*t.amount) revenue from holiday h join transport t ON h.mode=t.mode group by h.source,h.destination")

   val x= rev.sort($"revenue".desc).first

   println("most revenue generated route-->"+x)
```

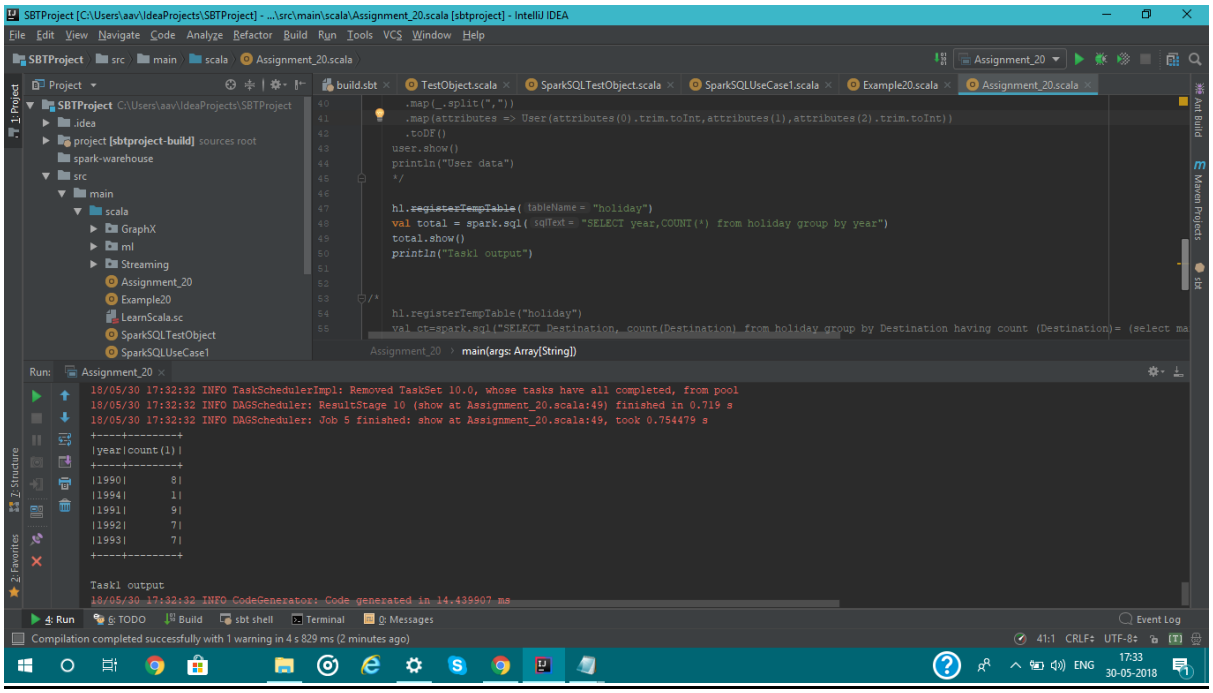
f. What is the total amount spent by every user on air-travel per year?

```
val amt = spark.sql("SELECT h.t_id, h.year,sum(t.amount) from holiday h join transport t ON h.mode=t.mode group by h.t_id,h.year")

   amt.show(32)

   println("Task6")
```

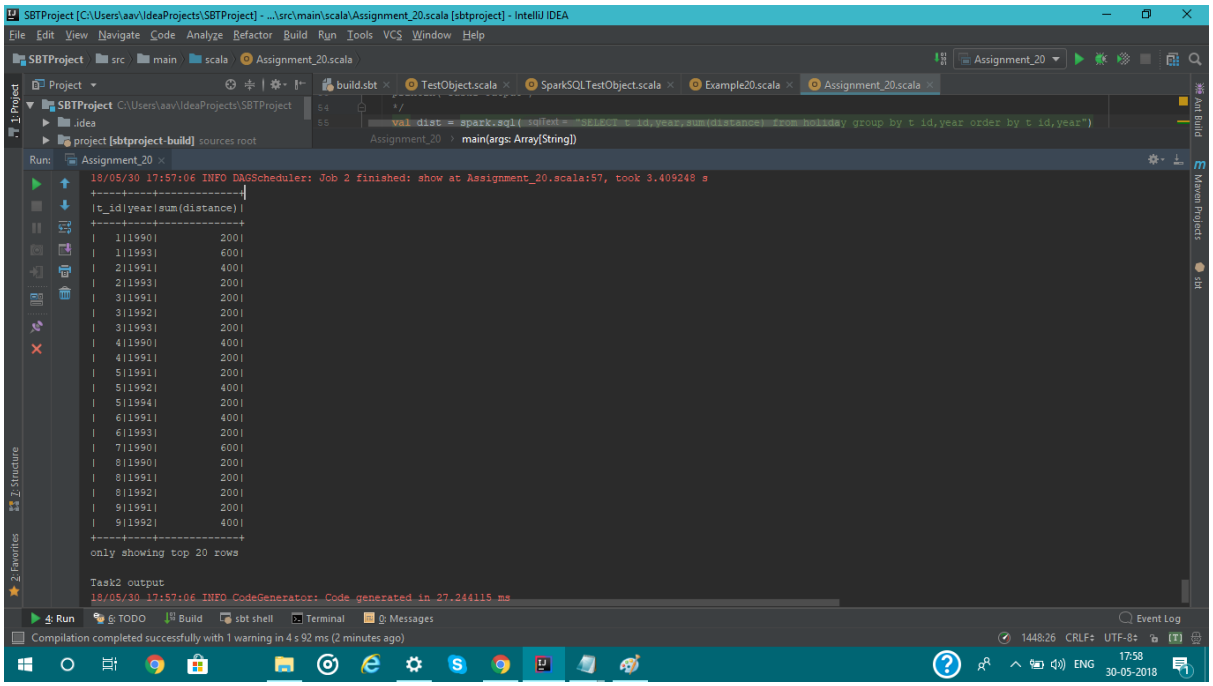
g. Considering age groups of < 20 , 20-35, 35 > ,Which age group is travelling the most every year

```
val new_user = spark.sql("SELECT * from holiday h join people p ON h.t_id = p.id")

   new_user.show()

   new_user.registerTempTable("NewDetails")

   val age = spark.sql("select year, age ,count(age) age_cnt from NewDetails group by
year,age order by age_cnt desc").take(1)

   println("Task7 output--->"+age.foreach(println))
```

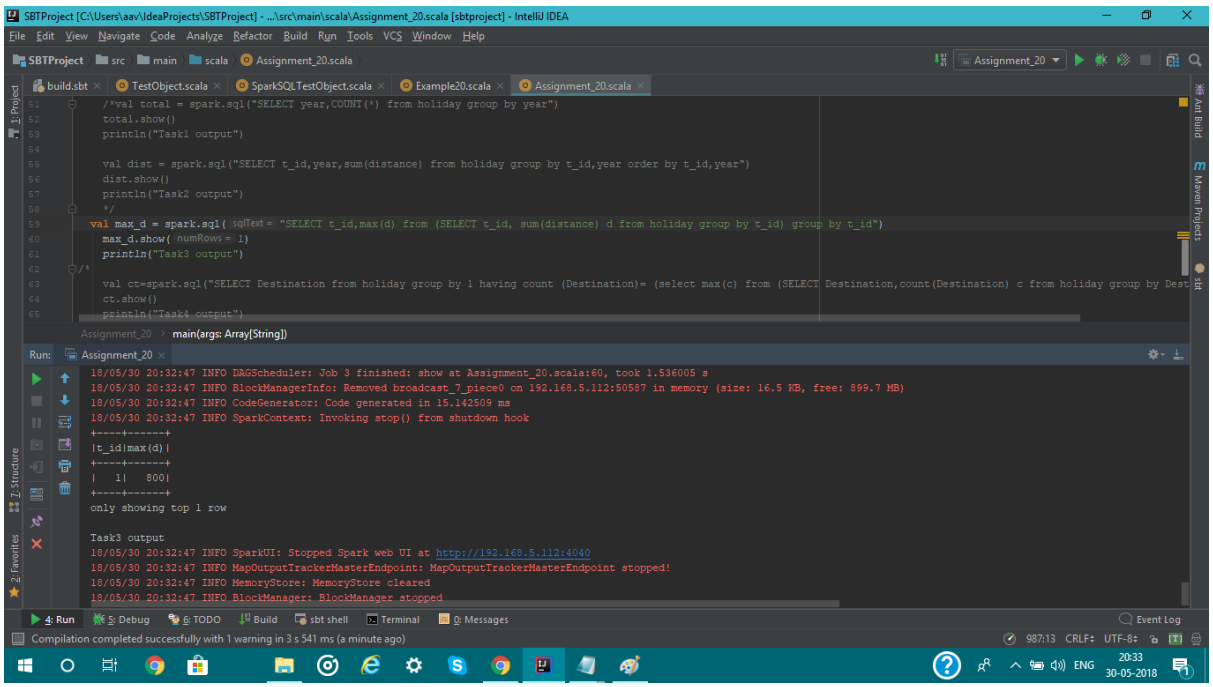
Screenshot



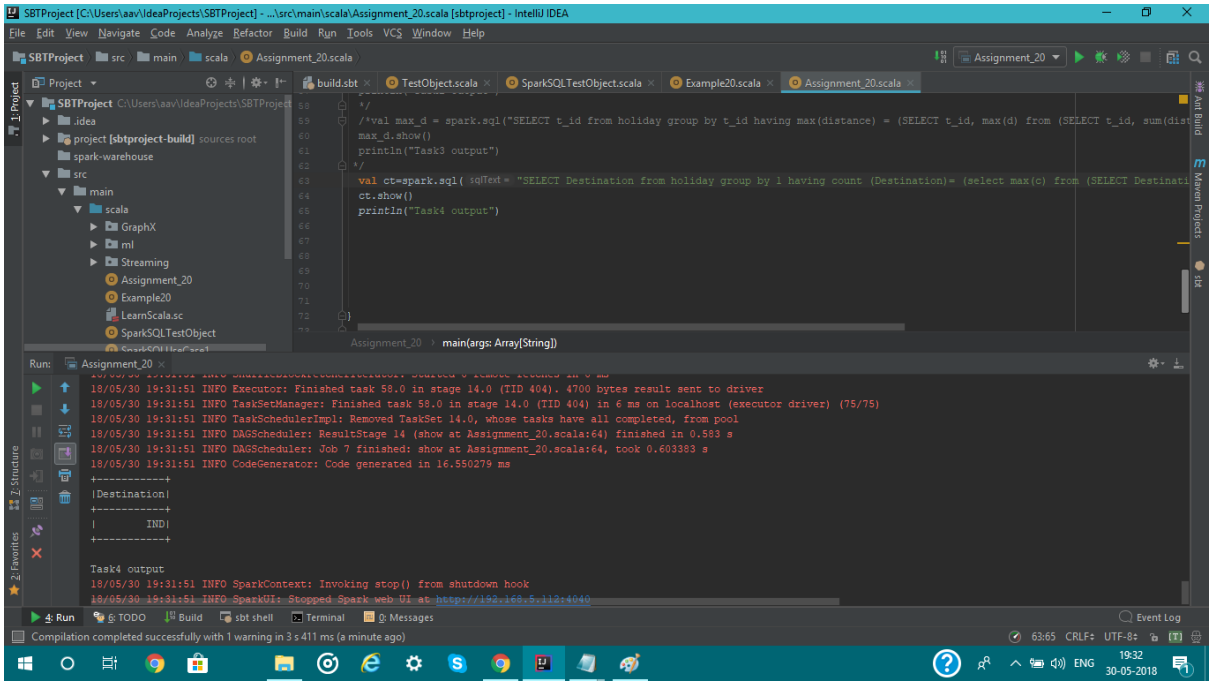
a.



b.


SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\Assignment_20.scala [sbtproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
/*val total = spark.sql("SELECT year,COUNT(*) from holiday group by year")
total.show()
println("Task1 output")
val dist = spark.sql("SELECT t_id,year,sum(distance) from holiday group by t_id,year order by t_id,year")
dist.show()
println("Task2 output")
*/
val max_d = spark.sql(sqlText = "SELECT t_id,max(d) from (SELECT t_id, sum(distance) d from holiday group by t_id) group by t_id")
max_d.show(numRows = 1)
println("Task3 output")
/*
val ct=spark.sql("SELECT Destination from holiday group by 1 having count (Destination)= (select max(c) from (SELECT Destination,count(Destination) c from holiday group by Dest
ct.show()
println("Task4 output")
Run: Assignment_20
18/05/30 20:32:47 INFO DAGScheduler: Job 3 finished: show at Assignment_20.scala:60, took 1.536005 s
18/05/30 20:32:47 INFO BlockManagerInfo: Removed broadcast_7_piece0 on 192.168.5.112:50587 in memory (size: 16.5 KB, free: 899.7 MB)
18/05/30 20:32:47 INFO CodeGenerator: Code generated in 15.142509 ms
18/05/30 20:32:47 INFO SparkContext: Invoking stop() from shutdown hook
+----+------+
|t_id|max(d)|
+----+------+
| 1| 800|
+----+------+
only showing top 1 row
Task3 output
18/05/30 20:32:47 INFO SparkUI: Stopped Spark web UI at http://192.168.5.112:4040
18/05/30 20:32:47 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/30 20:32:47 INFO MemoryStore: MemoryStore cleared
18/05/30 20:32:47 INFO BlockManager: BlockManager stopped
Compilation completed successfully with 1 warning in 3 s 541 ms (a minute ago)

c.


SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\Assignment_20.scala [sbtproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
/*val max_d = spark.sql("SELECT t_id from holiday group by t_id having max(distance) = (SELECT t_id, max(d) from (SELECT t_id, sum(dis
max_d.show()
println("Task3 output")
*/
val ct=spark.sql(sqlText = "SELECT Destination from holiday group by 1 having count (Destination)= (select max(c) from (SELECT Destinat
ct.show()
println("Task4 output")
Run: Assignment_20
18/05/30 19:31:51 INFO Executor: Finished task 58.0 in stage 14.0 (TID 404). 4700 bytes result sent to driver
18/05/30 19:31:51 INFO TaskSetManager: Finished task 58.0 in stage 14.0 (TID 404) in 6 ms on localhost (executor driver) (75/75)
18/05/30 19:31:51 INFO TaskSchedulerImpl: Removed TaskSet 14.0, whose tasks have all completed, from pool
18/05/30 19:31:51 INFO DAGScheduler: ResultStage 14 (show at Assignment_20.scala:64) finished in 0.583 s
18/05/30 19:31:51 INFO DAGScheduler: Job 7 finished: show at Assignment_20.scala:64, took 0.603383 s
18/05/30 19:31:51 INFO CodeGenerator: Code generated in 16.550279 ms
+-----------+
|Destination|
+-----------+
| IND|
+-----------+
Task4 output
18/05/30 19:31:51 INFO SparkContext: Invoking stop() from shutdown hook
18/05/30 19:31:51 INFO SparkUI: Stopped Spark web UI at http://192.168.5.112:4040
Compilation completed successfully with 1 warning in 3 s 411 ms (a minute ago)

d.

```
ct.show()
println("Task4 output")
*/
val rev = spark.sql( sqlText= "SELECT h.source, h.destination, sum(h.distance*t.amount) revenue from holiday h join transport t ON h.mode=t.mode group by h.source,h.destination")
val x= rev.sort(2"revenue".desc).first
println("most revenue generated route-->"+x)
```

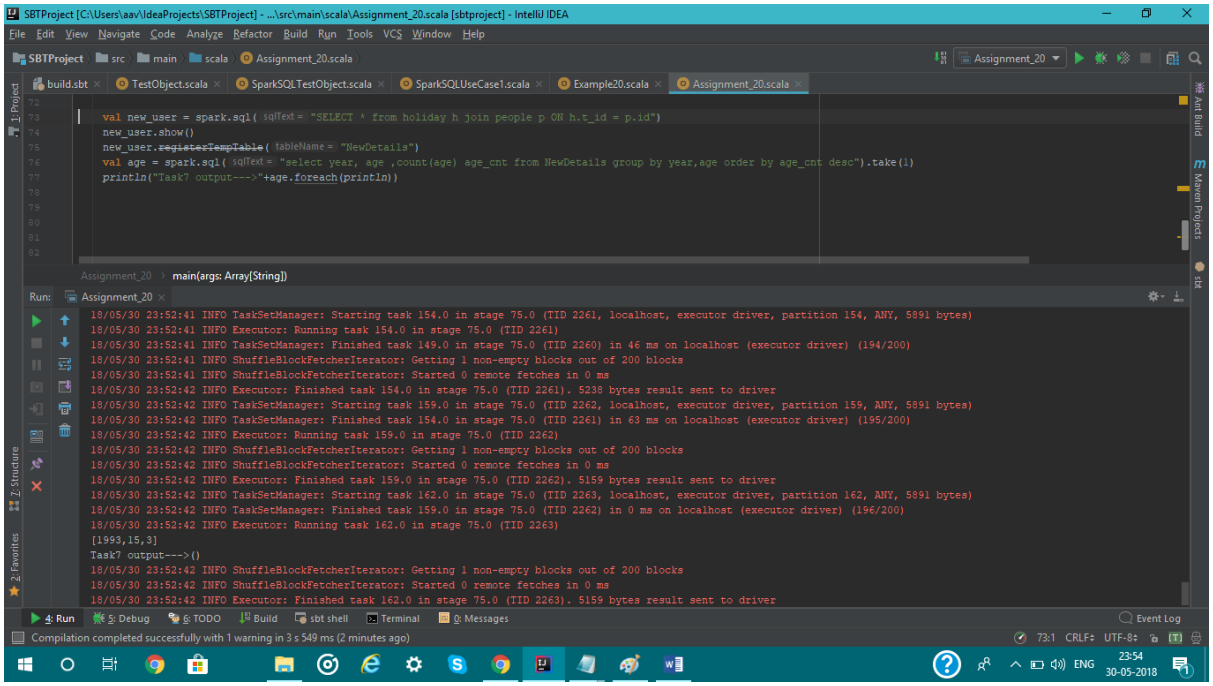
```
18/05/30 22:36:18 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/30 22:36:18 INFO Executor: Finished task 195.0 in stage 3.0 (TID 401). 5333 bytes result sent to driver
18/05/30 22:36:18 INFO TaskSetManager: Finished task 195.0 in stage 3.0 (TID 401) in 15 ms on localhost (executor driver) (200/200)
18/05/30 22:36:18 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
18/05/30 22:36:18 INFO DAGScheduler: ResultStage 3 (first at Assignment_20.scala:69) finished in 1.262 s
18/05/30 22:36:18 INFO DAGScheduler: Job 0 finished: first at Assignment_20.scala:69, took 6.926237 s
18/05/30 22:36:18 INFO CodeGenerator: Code generated in 16.330902 ms
most revenue generated route-->[CHN,IND,136000]
18/05/30 22:36:18 INFO SparkContext: Invoking stop() from shutdown hook
18/05/30 22:36:18 INFO SparkUI: Stopped Spark web UI at http://169.254.29.30:4040
18/05/30 22:36:18 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/30 22:36:18 INFO MemoryStore: MemoryStore cleared
18/05/30 22:36:18 INFO BlockManager: BlockManager stopped
18/05/30 22:36:18 INFO BlockManagerMaster: BlockManagerMaster stopped
18/05/30 22:36:18 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/05/30 22:36:18 INFO SparkContext: Successfully stopped SparkContext
18/05/30 22:36:18 INFO ShutdownHookManager: Shutdown hook called
18/05/30 22:36:18 INFO ShutdownHookManager: Deleting directory C:\Users\aav\AppData\Local\Temp\spark-0568e07d-c4e4-47a2-b152-122e38c999d6

Process finished with exit code 0
```

e.

```
new_user.show()
```

```
+----+----+-----------+
|t_id|year|sum(amount)|
+----+----+-----------+
|   3|1991|        170|
|   6|1993|        170|
|   3|1992|        170|
|   7|1990|        510|
|  10|1993|        170|
|   6|1991|        340|
|   2|1991|        340|
|   4|1991|        170|
|   5|1991|        170|
|   5|1994|        170|
|   8|1991|        170|
|   1|1990|        170|
|   5|1992|        340|
|   4|1990|        340|
|   3|1993|        170|
|  10|1990|        170|
|   2|1993|        170|
|   1|1993|        510|
|   9|1991|        170|
|   9|1992|        340|
|   8|1990|        170|
|  10|1992|        170|
|   8|1992|        170|
+----+----+-----------+

Task6 output
18/05/30 23:57:29 INFO SparkUI: Stopped Spark web UI at http://192.168.5.112:4040
```

f.


SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\Assignment_20.scala [sbtproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SBTProject src main scala Assignment_20.scala
build.sbt TestObject.scala SparkSQLTestObject.scala SparkSQLUseCase1.scala Example20.scala Assignment_20.scala
val new_user = spark.sql(sqlText = "SELECT * from holiday h join people p ON h.t_id = p.id")
new_user.show()
new_user.registerTempTable(tableName = "NewDetails")
val age = spark.sql(sqlText = "select year, age ,count(age) age_cnt from NewDetails group by year,age order by age_cnt desc").take(1)
println("Task7 output--->"+age.foreach(println))
Assignment_20 main(args: Array[String])
Run: Assignment_20
18/05/30 23:52:41 INFO TaskSetManager: Starting task 154.0 in stage 75.0 (TID 2261, localhost, executor driver, partition 154, ANY, 5891 bytes)
18/05/30 23:52:41 INFO Executor: Running task 154.0 in stage 75.0 (TID 2261)
18/05/30 23:52:41 INFO TaskSetManager: Finished task 149.0 in stage 75.0 (TID 2260) in 46 ms on localhost (executor driver) (194/200)
18/05/30 23:52:41 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 200 blocks
18/05/30 23:52:41 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/30 23:52:42 INFO Executor: Finished task 154.0 in stage 75.0 (TID 2261). 5238 bytes result sent to driver
18/05/30 23:52:42 INFO TaskSetManager: Starting task 159.0 in stage 75.0 (TID 2262, localhost, executor driver, partition 159, ANY, 5891 bytes)
18/05/30 23:52:42 INFO TaskSetManager: Finished task 154.0 in stage 75.0 (TID 2261) in 63 ms on localhost (executor driver) (195/200)
18/05/30 23:52:42 INFO Executor: Running task 159.0 in stage 75.0 (TID 2262)
18/05/30 23:52:42 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 200 blocks
18/05/30 23:52:42 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/30 23:52:42 INFO Executor: Finished task 159.0 in stage 75.0 (TID 2262). 5159 bytes result sent to driver
18/05/30 23:52:42 INFO TaskSetManager: Starting task 162.0 in stage 75.0 (TID 2263, localhost, executor driver, partition 162, ANY, 5891 bytes)
18/05/30 23:52:42 INFO TaskSetManager: Finished task 159.0 in stage 75.0 (TID 2262) in 0 ms on localhost (executor driver) (196/200)
18/05/30 23:52:42 INFO Executor: Running task 162.0 in stage 75.0 (TID 2263)
[1993,15,3]
Task7 output--->()
18/05/30 23:52:42 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 200 blocks
18/05/30 23:52:42 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/30 23:52:42 INFO Executor: Finished task 162.0 in stage 75.0 (TID 2263). 5159 bytes result sent to driver
4: Run 5: Debug 6: TODO Build sbt shell Terminal 0: Messages Event Log
Compilation completed successfully with 1 warning in 3 s 549 ms (2 minutes ago)
73:1 CRLF UTF-8
ENG 23:54 30-05-2018

g.