ASSIGNMENT 19.3

Problem Statement:

Create a data frame with 1 to 100 and save as parquet file.

Solution:

Parquet is a column-based storage format for Hadoop that is supported by many other data processing systems. Spark SQL supports both reading and writing Parquet files which preserves the schema of original data automatically.

Here is the Spark code in Scala to create a data frame and save as parquet file:

```
// import required Spark packages
import org.apache.spark.sql.SparkSession

object Assignment19_3 {
  def main(args: Array[String]): Unit = {
// create a SparkSession object that can be used to create various contexts of Spark such as sqlContext
    val spark = SparkSession
      .builder()
      .config("spark.sql.warehouse.dir", "file:///c:/tmp/spark-warehouse")
      .master("local[*]")
      .getOrCreate()

// initialize sqlContext
    val sqlContext = spark.sqlContext

// define a list of values ranging from 1 to 100

    val myList = 1 to 100 toList

// convert the list into a data frame

    val myDF = myList.toDF()
```

// print all the values of the data frame that has been created

println("Reading contents of a data frame:")

myDF.show(100)

// write the contents of the data frame in parquet format and save it into a local file system path

myDF.write.parquet("E:\\Acadgild\\Session 19\\Assignment 19.3\\ sampleParquetFile.parquet")

// read parquet back to data frame

val newDataDF = sqlContext.read.parquet("E:\\Acadgild\\Session 19\\Assignment 19.3\\ sampleParquetFile.parquet")

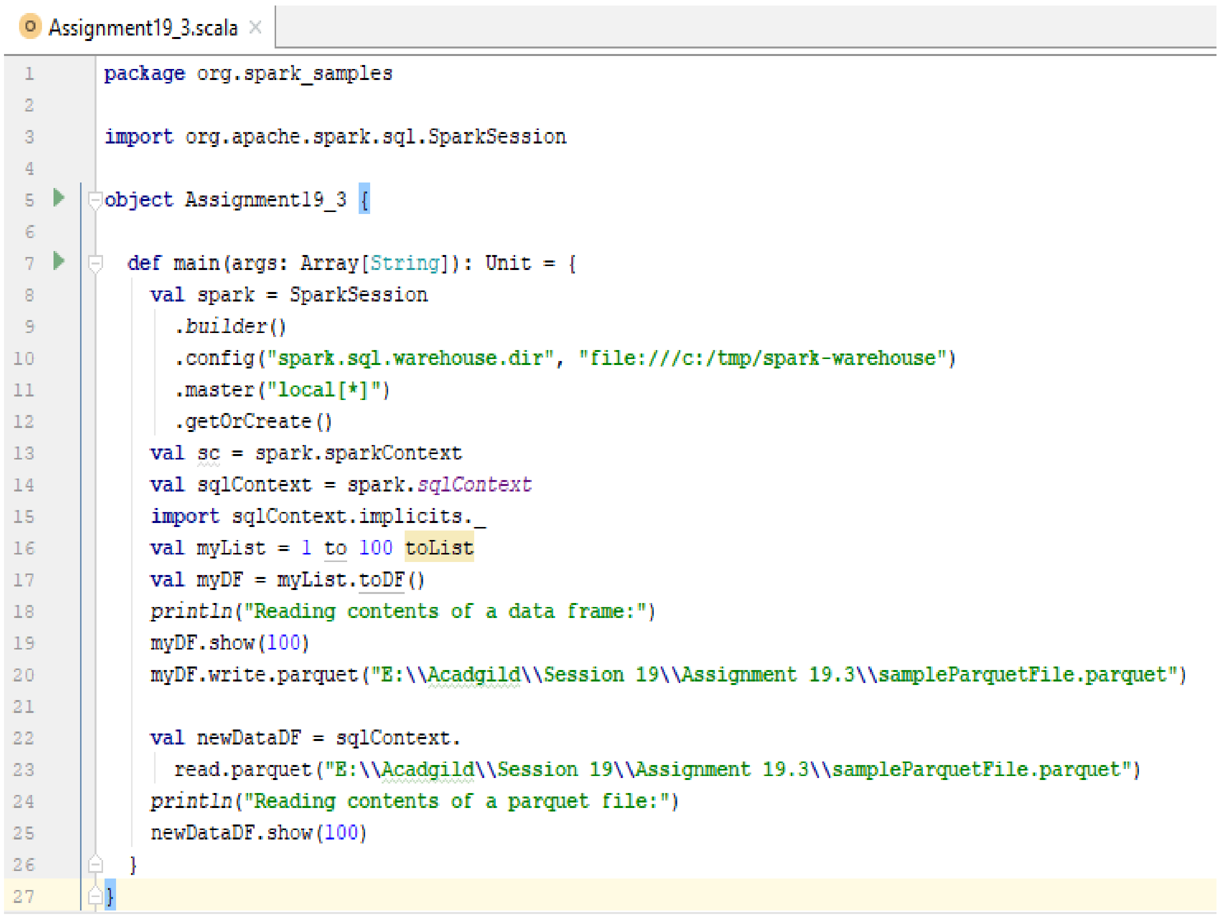
// show contents of new data frame

println("Reading contents of a parquet file:")

newDataDF.show(100)

}

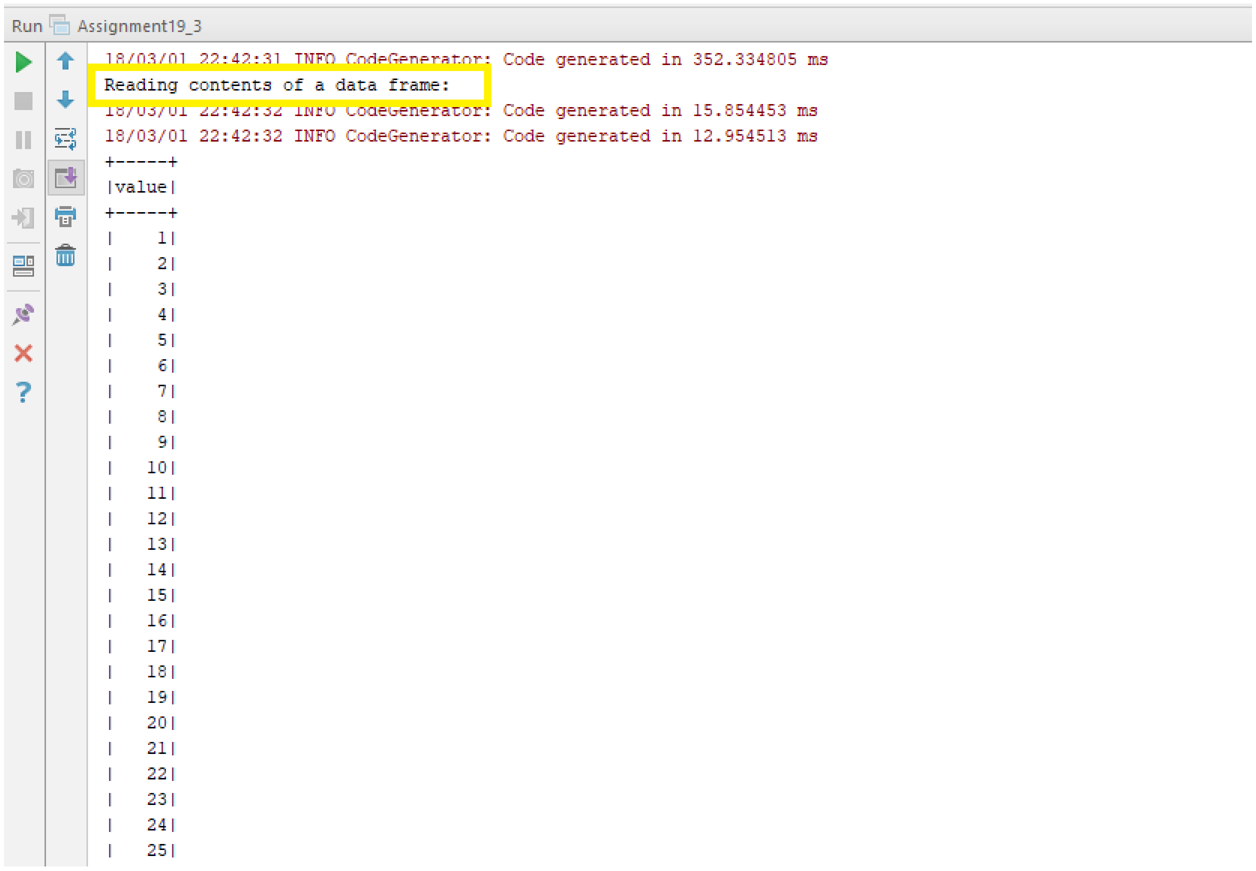
}



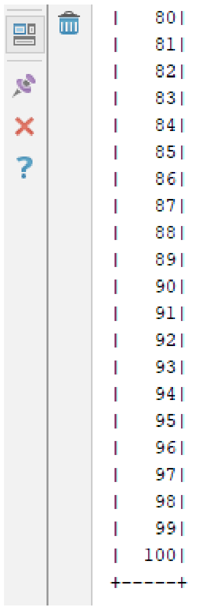
Assignment19_3.scala

```
package org.spark_samples

import org.apache.spark.sql.SparkSession

object Assignment19_3 {

  def main(args: Array[String]): Unit = {
    val spark = SparkSession
      .builder()
      .config("spark.sql.warehouse.dir", "file:///c:/tmp/spark-warehouse")
      .master("local[*]")
      .getOrCreate()
    val sc = spark.sparkContext
    val sqlContext = spark.sqlContext
    import sqlContext.implicits._
    val myList = 1 to 100 toList
    val myDF = myList.toDF()
    println("Reading contents of a data frame:")
    myDF.show(100)
    myDF.write.parquet("E:\\Acadgild\\Session 19\\Assignment 19.3\\sampleParquetFile.parquet")

    val newDataDF = sqlContext.
      read.parquet("E:\\Acadgild\\Session 19\\Assignment 19.3\\sampleParquetFile.parquet")
    println("Reading contents of a parquet file:")
    newDataDF.show(100)
  }
}
```

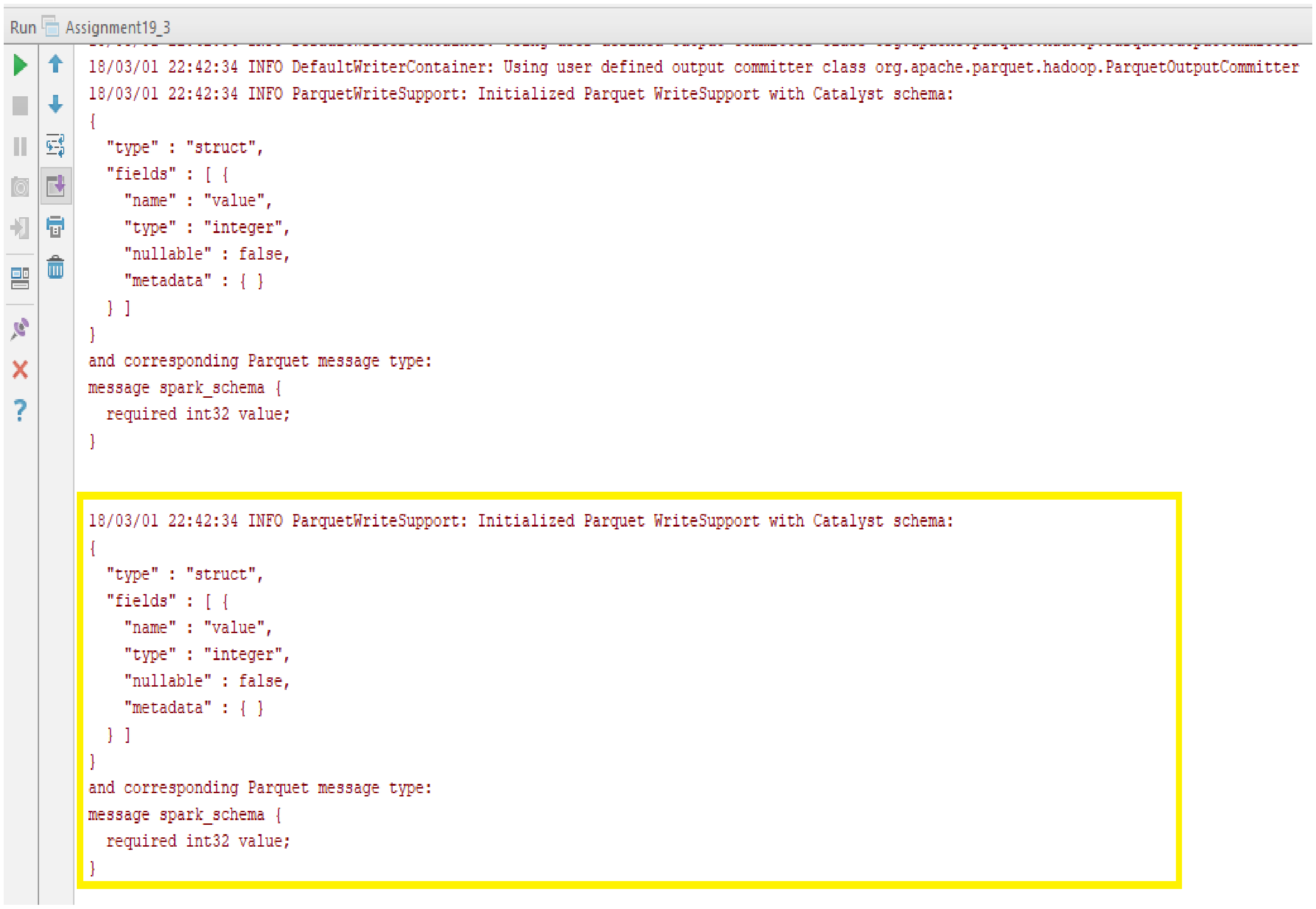
Output:



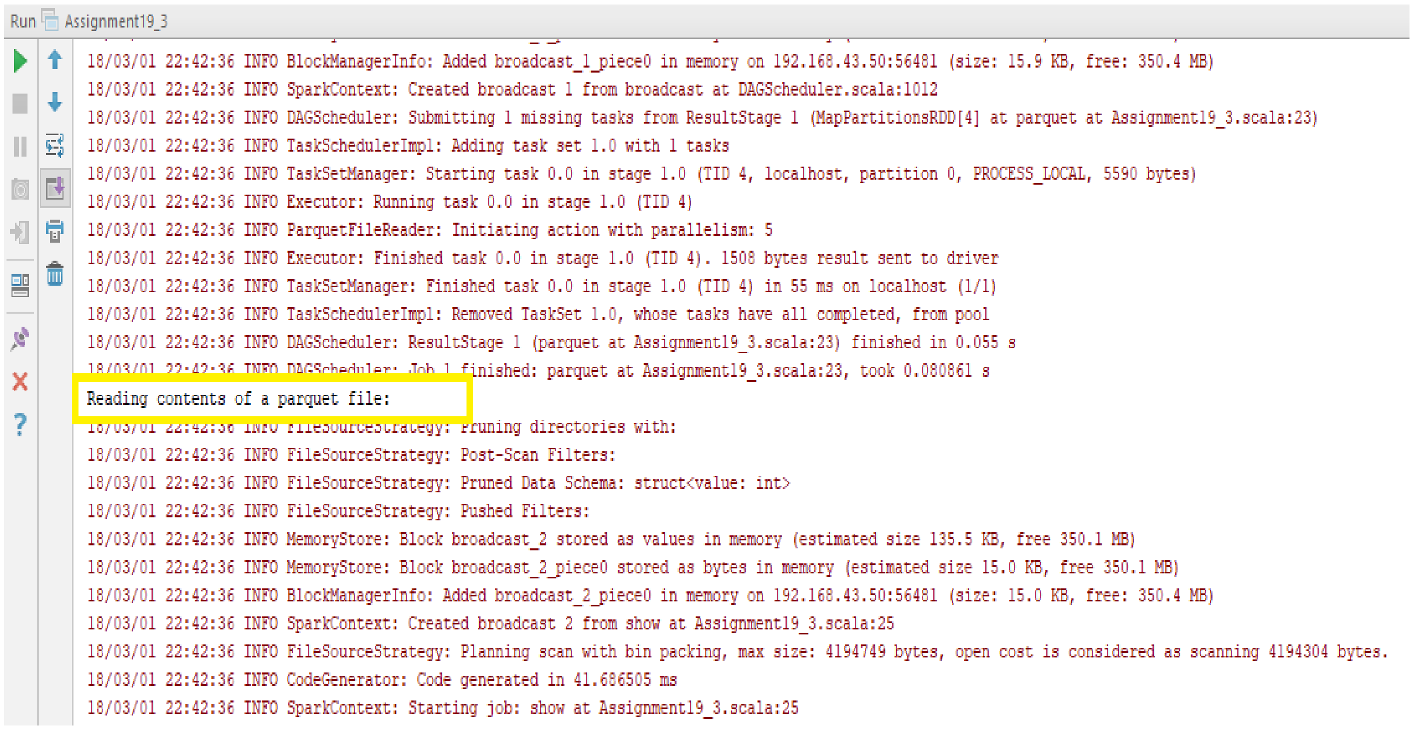
```
18/03/01 22:42:31 INFO CodeGenerator: Code generated in 352.334805 ms
Reading contents of a data frame:
18/03/01 22:42:32 INFO CodeGenerator: Code generated in 15.854453 ms
18/03/01 22:42:32 INFO CodeGenerator: Code generated in 12.954513 ms
+-----+
|value|
+-----+
|    1|
|    2|
|    3|
|    4|
|    5|
|    6|
|    7|
|    8|
|    9|
|   10|
|   11|
|   12|
|   13|
|   14|
|   15|
|   16|
|   17|
|   18|
|   19|
|   20|
|   21|
|   22|
|   23|
|   24|
|   25|
```



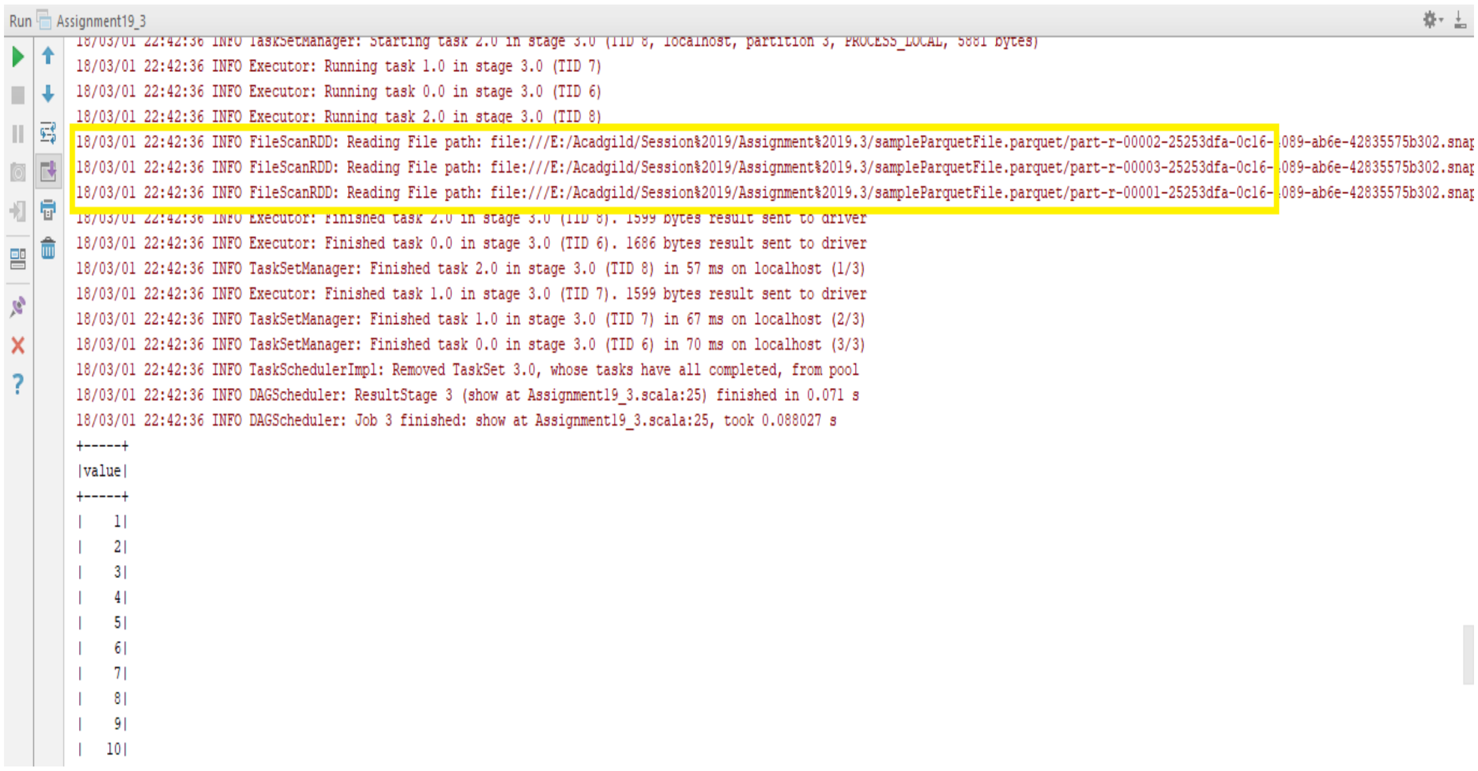
```
|   80|
|   81|
|   82|
|   83|
|   84|
|   85|
|   86|
|   87|
|   88|
|   89|
|   90|
|   91|
|   92|
|   93|
|   94|
|   95|
|   96|
|   97|
|   98|
|   99|
|  100|
+-----+
```



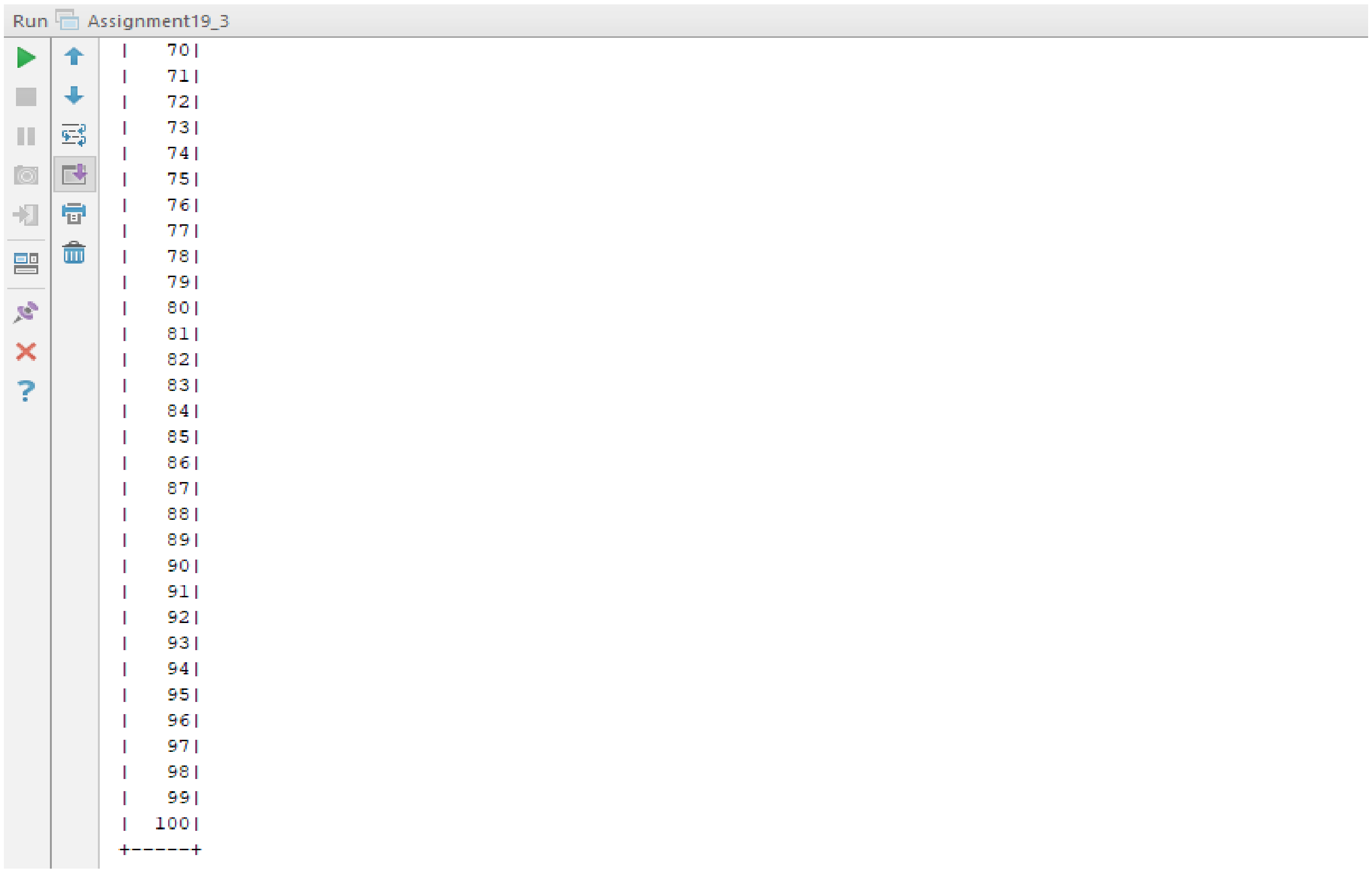
```
18/03/01 22:42:34 INFO DefaultWriterContainer: Using user defined output committer class org.apache.parquet.hadoop.ParquetOutputCommitter
18/03/01 22:42:34 INFO ParquetWriteSupport: Initialized Parquet WriteSupport with Catalyst schema:
{
  "type" : "struct",
  "fields" : [ {
    "name" : "value",
    "type" : "integer",
    "nullable" : false,
    "metadata" : { }
  } ]
}
and corresponding Parquet message type:
message spark_schema {
  required int32 value;
}


18/03/01 22:42:34 INFO ParquetWriteSupport: Initialized Parquet WriteSupport with Catalyst schema:
{
  "type" : "struct",
  "fields" : [ {
    "name" : "value",
    "type" : "integer",
    "nullable" : false,
    "metadata" : { }
  } ]
}
and corresponding Parquet message type:
message spark_schema {
  required int32 value;
}
```



```
18/03/01 22:42:36 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on 192.168.43.50:56481 (size: 15.9 KB, free: 350.4 MB)
18/03/01 22:42:36 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1012
18/03/01 22:42:36 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[4] at parquet at Assignment19_3.scala:23)
18/03/01 22:42:36 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
18/03/01 22:42:36 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 4, localhost, partition 0, PROCESS_LOCAL, 5590 bytes)
18/03/01 22:42:36 INFO Executor: Running task 0.0 in stage 1.0 (TID 4)
18/03/01 22:42:36 INFO ParquetFileReader: Initiating action with parallelism: 5
18/03/01 22:42:36 INFO Executor: Finished task 0.0 in stage 1.0 (TID 4). 1508 bytes result sent to driver
18/03/01 22:42:36 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 4) in 55 ms on localhost (1/1)
18/03/01 22:42:36 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/03/01 22:42:36 INFO DAGScheduler: ResultStage 1 (parquet at Assignment19_3.scala:23) finished in 0.055 s
18/03/01 22:42:36 INFO DAGScheduler: Job 1 finished: parquet at Assignment19_3.scala:23, took 0.080861 s
Reading contents of a parquet file:
18/03/01 22:42:36 INFO FileSourceStrategy: Pruning directories with:
18/03/01 22:42:36 INFO FileSourceStrategy: Post-Scan Filters:
18/03/01 22:42:36 INFO FileSourceStrategy: Pruned Data Schema: struct<value: int>
18/03/01 22:42:36 INFO FileSourceStrategy: Pushed Filters:
18/03/01 22:42:36 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 135.5 KB, free 350.1 MB)
18/03/01 22:42:36 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 15.0 KB, free 350.1 MB)
18/03/01 22:42:36 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 192.168.43.50:56481 (size: 15.0 KB, free: 350.4 MB)
18/03/01 22:42:36 INFO SparkContext: Created broadcast 2 from show at Assignment19_3.scala:25
18/03/01 22:42:36 INFO FileSourceStrategy: Planning scan with bin packing, max size: 4194749 bytes, open cost is considered as scanning 4194304 bytes.
18/03/01 22:42:36 INFO CodeGenerator: Code generated in 41.686505 ms
18/03/01 22:42:36 INFO SparkContext: Starting job: show at Assignment19_3.scala:25
```


Run Assignment19_3

```
18/03/01 22:42:36 INFO TaskSetManager: Starting task 2.0 in stage 3.0 (TID 8, localhost, partition 3, PROCESS_LOCAL, 5881 bytes)
18/03/01 22:42:36 INFO Executor: Running task 1.0 in stage 3.0 (TID 7)
18/03/01 22:42:36 INFO Executor: Running task 0.0 in stage 3.0 (TID 6)
18/03/01 22:42:36 INFO Executor: Running task 2.0 in stage 3.0 (TID 8)
18/03/01 22:42:36 INFO FileScanRDD: Reading File path: file:///E:/Acadgild/Session%2019/Assignment%2019.3/sampleParquetFile.parquet/part-r-00002-25253dfa-0c16-[illegible]089-ab6e-42835575b302.snap
18/03/01 22:42:36 INFO FileScanRDD: Reading File path: file:///E:/Acadgild/Session%2019/Assignment%2019.3/sampleParquetFile.parquet/part-r-00003-25253dfa-0c16-[illegible]089-ab6e-42835575b302.snap
18/03/01 22:42:36 INFO FileScanRDD: Reading File path: file:///E:/Acadgild/Session%2019/Assignment%2019.3/sampleParquetFile.parquet/part-r-00001-25253dfa-0c16-[illegible]089-ab6e-42835575b302.snap
18/03/01 22:42:36 INFO Executor: Finished task 2.0 in stage 3.0 (TID 8). 1599 bytes result sent to driver
18/03/01 22:42:36 INFO Executor: Finished task 0.0 in stage 3.0 (TID 6). 1686 bytes result sent to driver
18/03/01 22:42:36 INFO TaskSetManager: Finished task 2.0 in stage 3.0 (TID 8) in 57 ms on localhost (1/3)
18/03/01 22:42:36 INFO Executor: Finished task 1.0 in stage 3.0 (TID 7). 1599 bytes result sent to driver
18/03/01 22:42:36 INFO TaskSetManager: Finished task 1.0 in stage 3.0 (TID 7) in 67 ms on localhost (2/3)
18/03/01 22:42:36 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 6) in 70 ms on localhost (3/3)
18/03/01 22:42:36 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
18/03/01 22:42:36 INFO DAGScheduler: ResultStage 3 (show at Assignment19_3.scala:25) finished in 0.071 s
18/03/01 22:42:36 INFO DAGScheduler: Job 3 finished: show at Assignment19_3.scala:25, took 0.088027 s
+-----+
|value|
+-----+
|    1|
|    2|
|    3|
|    4|
|    5|
|    6|
|    7|
|    8|
|    9|
|   10|
```


Run Assignment19_3

```
|   70|
|   71|
|   72|
|   73|
|   74|
|   75|
|   76|
|   77|
|   78|
|   79|
|   80|
|   81|
|   82|
|   83|
|   84|
|   85|
|   86|
|   87|
|   88|
|   89|
|   90|
|   91|
|   92|
|   93|
|   94|
|   95|
|   96|
|   97|
|   98|
|   99|
|  100|
+-----+
```