SPARK SQL ASSIGNMENT 19.3

PROBLEM STATEMENT:

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

Input Commands:

```
//importing the package to use ListBuffer
import scala.collection.mutable.ListBuffer;
//defining the function to make a Listbuffer of integers from 1 to 100
def populatelist(n:ListBuffer[Integer])=
{
var i = 1
while(i<=100){
n+=i
i+=1
}
}
//defining the Listbuffer object of integers
var mynewlistbuffer = new ListBuffer[Integer]()
//Invoking the User defined function and passing the ListBuffer object as an argument
populatelist(mynewlistbuffer)
//Conversion of ListBuffer Object to List
val mynewlist = mynewlistbuffer.toList
//Importing the SPARK & sqlContext packages to use Dataframe
import org.apache.spark.sql._
```

import sqlContext.implicits._

//Creation of RDD with List of integers

val rdd1= sc.parallelize(mynewlist)

//Creation of Dataframe with the help of RDD defined above

val dfs = sqlContext.createDataFrame(rdd1.map(Tuple1.apply)).toDF("Column")

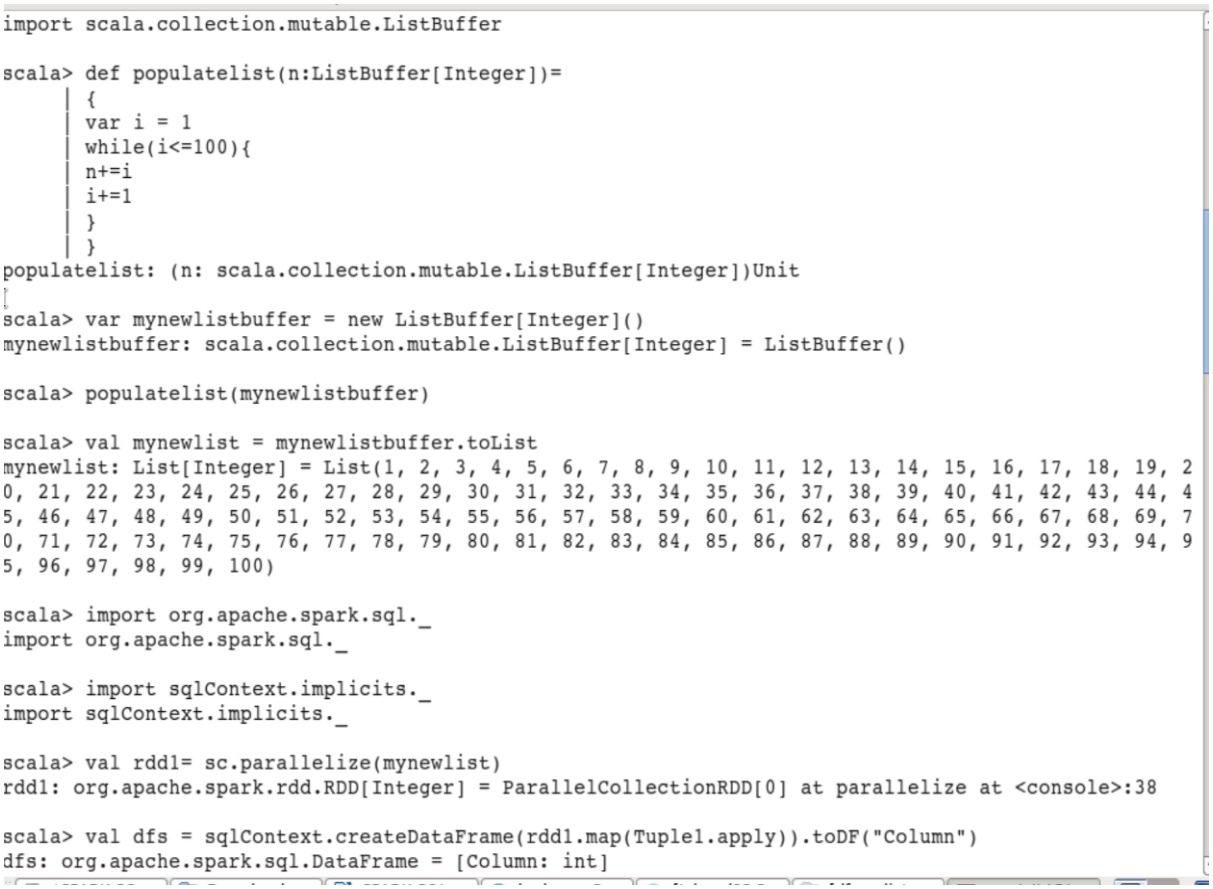
//Saving the Dataframe as Parquet File

dfs.saveAsParquetFile("file:/home/acadgild/Downloads/dfsmynewlist.parquet")

//Saving the Dataframe as CSV File

dfs.map(_.mkString("_")).saveAsTextFile("file:/home/acadgild/Downloads/dfsmynewlist.csv")

Screenshots:



```
import scala.collection.mutable.ListBuffer

scala> def populatelist(n:ListBuffer[Integer])=
     | {
     | var i = 1
     | while(i<=100){
     | n+=i
     | i+=1
     | }
     | }
populatelist: (n: scala.collection.mutable.ListBuffer[Integer])Unit

scala> var mynewlistbuffer = new ListBuffer[Integer]()
mynewlistbuffer: scala.collection.mutable.ListBuffer[Integer] = ListBuffer()

scala> populatelist(mynewlistbuffer)

scala> val mynewlist = mynewlistbuffer.toList
mynewlist: List[Integer] = List(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 2
0, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 4
5, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 7
0, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 9
5, 96, 97, 98, 99, 100)

scala> import org.apache.spark.sql._
import org.apache.spark.sql._

scala> import sqlContext.implicits._
import sqlContext.implicits._

scala> val rdd1= sc.parallelize(mynewlist)
rdd1: org.apache.spark.rdd.RDD[Integer] = ParallelCollectionRDD[0] at parallelize at <console>:38

scala> val dfs = sqlContext.createDataFrame(rdd1.map(Tuple1.apply)).toDF("Column")
dfs: org.apache.spark.sql.DataFrame = [Column: int]
```

```
scala> dfs.map(_.mkString("_")).saveAsTextFile("file:/home/acadgild/Downloads/dfsmynewlist.csv")
[
scala> dfs.saveAsParquetFile("file:/home/acadgild/Downloads/dfsmynewlist.parquet")
warning: there were 1 deprecation warning(s); re-run with -deprecation for details
```

Output PARQUET file which has been generated as output , has been attached to GITHUB directory.

Similarly CSV file has been generated as output and it has been attached to GITHUB Directory.