Spark Assignment 19.3

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

This assignment is done in the spark shell within the acadgildVM.

Steps Followed:

1) Created a range of Int using the below command and converted it to Dataframe

   val r = 1 to 100

   val primitiveDS = r.toDF

2) Using dataframe.write.parquet() function, saving it as parquet file

   primitiveDF.write.parquet("assignment_19_3.parquet")

Spark-shell output

```
scala> val r = 1 to 100
r: scala.collection.immutable.Range.Inclusive = Range(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, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55,
 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 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)

scala> val primitiveDF = r.toDF
primitiveDF: org.apache.spark.sql.DataFrame = [value: int]

scala> primitiveDF.write.parquet("assignment_19_3.parquet")
```

3) We can read the parquet file and check if the contents are stored. We used read.parquet function of SparkSession.

   val parquetFileDF = spark.read.parquet("assignment_19_3.parquet")

4) Parquet files can also be used to create a temporary view and then used in SQL statements

   parquetFileDF.createOrReplaceTempView("parquetFile")

5) Displaying the contents of parquetFile using the below command.

   val printing = spark.sql("SELECT * FROM parquetFile").show()

Spark-shell output

```
scala> val printing = spark.sql("SELECT * FROM parquetFile").show()
+-----+
|value|
+-----+
|    1|
|    2|
|    3|
|    4|
|    5|
|    6|
|    7|
|    8|
|    9|
|   10|
|   11|
|   12|
|   13|
|   14|
|   15|
|   16|
|   17|
|   18|
|   19|
|   20|
+-----+
only showing top 20 rows

printing: Unit = ()
```