

Assignment 19.3:

Problem Statement:

Create a dataframe with 1 to 100 and store it to a parquet file

Step1: Create a dataframe 1 to 100

Use parallelize for 1 to 100 and convert to dataframe using toDF method with field num and put to numbersDF

Source code is as below:

```
val numbersDF = sc.parallelize(1 to 100).toDF("num")
```

Step2: Store dataframe numbersDF as a parquet file

Store the dataframe numbersDF as parquet file numbers.parquet by using write.parquet on numbersDF

Source Code is as below:

```
numbersDF.write.parquet("numbers.parquet")
```

Screenshots of Step1 and Step2 is as below:

```
scala> val numbersDF = sc.parallelize(1 to 100).toDF("num")
numbersDF: org.apache.spark.sql.DataFrame = [num: int]

scala> numbersDF.write.parquet("numbers.parquet")
```

Step3: Read from parquet file

Read from parquet file numbers.parquet using method read.parquet on sqlContext and put this to newNumbersDF. Next, display all the 100 numbers by using show method with argument 100 on newNumbersDF

Code is as below

```
val newNumbersDF = sqlContext.read.parquet("numbers.parquet")
```

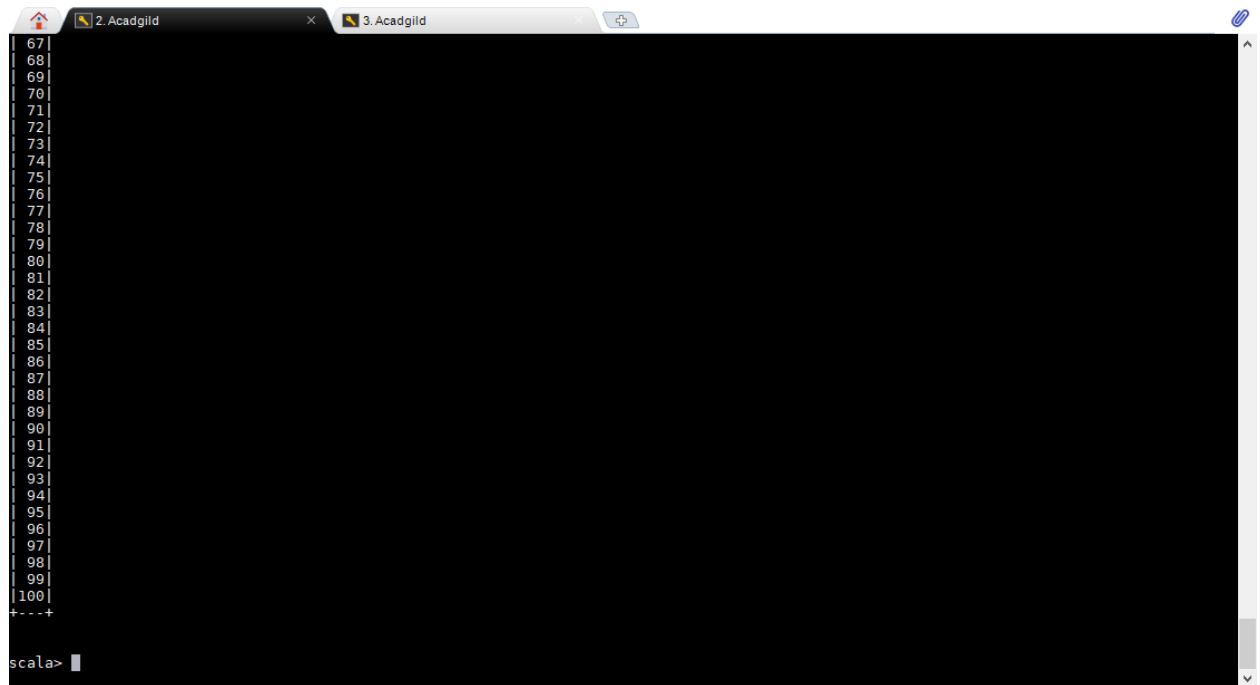
```
newNumbersDF.show(100)
```

Screenshots are as below:

```
scala> val newNumbersDF = spark.sqlContext.read.parquet("numbers.parquet")
newNumbersDF: org.apache.spark.sql.DataFrame = [num: int]

scala> newNumbersDF.show(100)
+---+
|num|
+---+
| 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|
```

```
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|
```



```
| 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> |
```