Part 2

user_artists.dat:

- userID artistID weight
- 2 51 13883
- 2 52 11690
- 2 53 11351
- 2 54 10300
- 2 55 8983

Main Steps:

Load the dataset as RDD:

```
> val lines = sc.textFile("YOUR_SPARK_HOME/user_artists.dat")
```

Construct custome data schema:

```
val schema = StructType(colNames.map(fieldName => StructField(fieldName, IntegerType)))
```

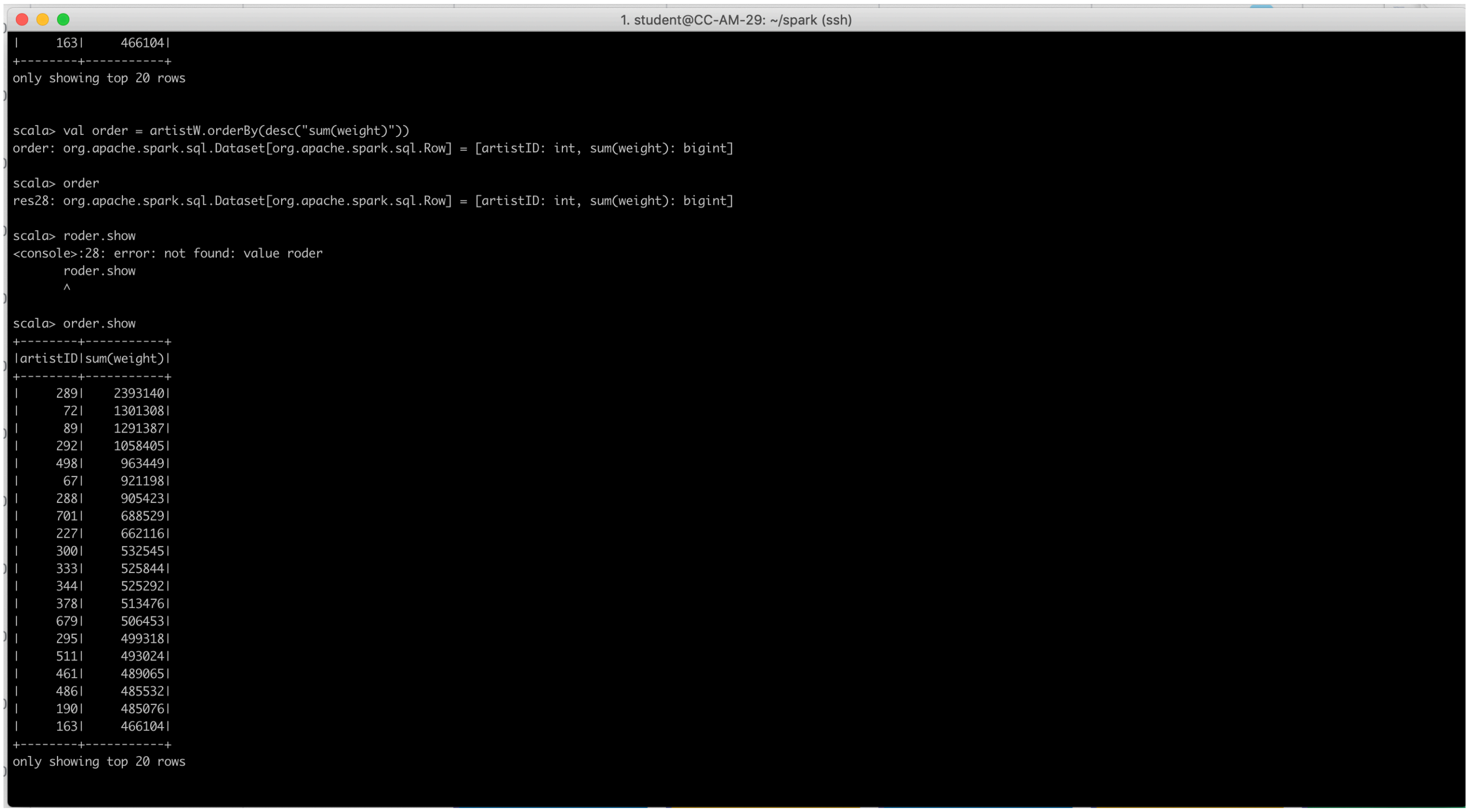
Transform the RDD into spark DateFrame:

```
val data = spark.createDataFrame(rowRDD,schema)
```

Get Result by impelmenting spark SQL query on dataframe. Group by artist ID and sum the weight. Order in descending order:

```
val artistW = data.groupBy("artistID").agg(sum("weight"))
val order = artistW.orderBy(desc("sum(weight)"))
```

Show the result dataframe:



```
|     163|     466104|
+--------+-----------+
only showing top 20 rows


scala> val order = artistW.orderBy(desc("sum(weight)"))
order: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [artistID: int, sum(weight): bigint]

scala> order
res28: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [artistID: int, sum(weight): bigint]

scala> roder.show
<console>:28: error: not found: value roder
       roder.show
       ^

scala> order.show
+--------+-----------+
|artistID|sum(weight)|
+--------+-----------+
|     289|    2393140|
|      72|    1301308|
|      89|    1291387|
|     292|    1058405|
|     498|     963449|
|      67|     921198|
|     288|     905423|
|     701|     688529|
|     227|     662116|
|     300|     532545|
|     333|     525844|
|     344|     525292|
|     378|     513476|
|     679|     506453|
|     295|     499318|
|     511|     493024|
|     461|     489065|
|     486|     485532|
|     190|     485076|
|     163|     466104|
+--------+-----------+
only showing top 20 rows
```