Write an application using Python, Scala or Java that will use Spark to do the following:

*   Read the data file ‘data.csv’.
*   Create an optimised parquet file with the same data
*   Load the parquet file into Spark
*   Aggregate the values by country
*   Write the results to a parquet file

Input:

Expected output:

|  |  |
| --- | --- |
| **Country** | **Values** |
| Canada | 47;97;33;94;6 |
| Canada | 59;98;24;83;3 |
| Canada | 77;63;93;86;62 |
| China | 86;71;72;23;27 |
| China | 74;69;72;93;7 |
| China | 58;99;90;93;41 |
| England | 40;13;85;75;90 |
| England | 39;13;33;29;14 |
| England | 99;88;57;69;49 |
| Germany | 67;93;90;57;3 |
| Germany | 0;9;15;20;19 |
| Germany | 77;64;46;95;48 |
| India | 90;49;91;14;70 |
| India | 70;83;38;27;16 |
| India | 86;21;19;59;4 |

**Solution:**

1. val csv = spark.read.option(“header”, “true”).csv(“data.csv”)
2. val parquet = csv.parquet(“csv\_to\_parquet”)
3. val load\_parquet = spark.read.option(“header”, “true”).parquet(“csv\_to\_parquet”)
4. val agg = parquet.groupBy(“country”).sum(“value”).save(parquet)