Lab 6

In this lab, we continue our work on the Amazon dataset, using Apache Spark. Your task is the same as in Lab 3: given the original Amazon food dataset (that you can find at /data/students/bigdata-01QYD/Lab3/Reviews.csv), find all the pairs of items frequently bought together, that is, reviewed by the same user.
In Ex1 you find the steps that you are required to perform on the dataset, that are similar to the one you already implemented in Hadoop.

Ex. 1

Write a single Spark application that:

- · Transposes the original dataset, obtaining an RDD ot the type:
<user_id> → < list of the product_ids reviewed>;
- · Counts the frequencies of all the pairs of products reviewed together;
- · Writes on the output file all the pairs that appear more than once and their frequency.

Inspect the output to search for interesting (or funny!) facts, and analyse the job execution as usual (performances, executors, etc…).

Compare your results with the ones you obtained in Lab 3.

Bonus task

In the same application of Ex1, add a final print statement to write to the standard output the top 10, most frequent, pairs and their frequency.