Lab 3.

In this lab, we will continue our investigation on the Amazon fine foods dataset we have been using so far (available in the HFDS shared folder of the BigData@Polito cluster: /data/students/bigdata-01QYD/Lab3/Reviews.csv). Up to now, all the analysis we did were on a “per-row” basis, since we analyzed only the text of the reviews alone. Today we will start looking at the connections between the single reviews, and to do this we will need to transpose the original dataset.The transposed dataset contains one line per reviewer. The first field of each line contains always the id of the reviewer (BXXXXXXXXX), followed by the list of all the products reviewed by her/him (AXXXXXXXXX).

Here’s a sample of such transposed dataset:

B001E4KFG0,A3SGXH7AUHU8GW
B00813GRG4,A1D87F6ZCVE5NK
B000LQOCH0,ABXLMWJIXXAIN
B000UA0QIQ,A395BORC6FGVXV
B006K2ZZ7K,A1UQRSCLF8GW1T,ADT0SRK1MGOEU,A1SP2KVKFXXRU1,A3JRGQVEQN31IQ
B000E7L2R4,A1MZYO9TZK0BBI
B00171APVA,A21BT40VZCCYT4
B0001PB9FE,A3HDKO7OW0QNK4
B0009XLVG0,A2725IB4YY9JEB,A327PCT23YH90
B001GVISJM,A18ECVX2RJ7HUE,A2MUGFV2TDQ47K,A1CZX3CP8IKQIJ,AFKW14U97Z6QO,A2A9X58G2GTBLP,A3IV7CL2C13K2U,A1WO0KGLPR5PV6,AZOF9E17RGZH8,ARYVQL4N737A1,A3KLWF6WQ5BNYO,AJ613OLZZUG7V,A22P2J09NJ9HKE,A3FONPR03H3PJS,A3RXAU2N8KV45G,AAAS38B98HMIK,A1SP2KVKFXXRU1,A3JRGQVEQN31IQ

For the following exercise, you can start from this sample dataset, which you find also on the website of the course at the following link:
http://dbdmg.polito.it/wordpress/wp-content/uploads/2016/04/AmazonTransposedDataset_Sample.txt
You can then reproduce the full transposed dataset (computed on the whole Reviews.csv) on your own (ex. 2) and repeat your investigation on it.

Ex 1. “People also like…”

In this exercise, we try to build a very basic version of a recommending system. Your goal is to find the top 100 pairs of products most often reviewed (and so bought) together.

In this exercise, we consider two products as reviewed (i.e., bought) together if they appear in the same line of the transposed file (AmazonTransposedDataset_Sample.txt). We ignore temporal constraints, so even if a decade or a thousand products have passed between the two reviews, we count the pair, as it represents anyway the tastes of a single user.

In the sample dataset (AmazonTransposedDataset_Sample.txt), you will find only one such pair repeating twice, so you can limit to the top 3 pairs.

Ex 2. Transposing the Reviews.csv dataset

The original review dataset (available in the HFDS shared folder of the BigData@Polito cluster: /data/students/bigdata-01QYD/Lab3/Reviews.csv) lists all the reviews per-row, and is comma-separated. In each line, two of the columns represent the user id and product id.

Write a MapReduce application that transposes the original review dataset (Reviews.csv), listing for each user id all the products he/she has reviewed, as in the format above specified (i.e., the same format of the sample file AmazonTransposedDataset_Sample.txt).

Save the output of this application (i.e. the transposed dataset) to a HDFS folder named ReviewsTransposed in your HDFS home. Then, use it as input for the application you have built in Ex. 1, and save the output in a HDFS folder named MostReviewedTogether. Keep these folders for the future labs.