UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

FIB

Concepts for Specialised Databases

The Hadoop 2.x Ecosystem: HDFS, HBase, and MapReduce

Objectives

This is a hands-on exercise with HBase and MapReduce. Obviously, HDFS will be in the middle, but we will focus on the first two. The list of objectives for this session is as follows:

- Manage an HBase store (i.e., create / modify / drop tables, insert / drop data),
- Write simple Map / Reduce tasks on top of files.

Lab organization

On session 1, the lecturer will be present at the lab and will help you out. Thus, it is a session to solve your doubts and help you to prepare the final delivery.

On session 2, you must upload the deliverables specified in the corresponding section of this document.

Your lab mate for this practice will be that of the corresponding team creation event.
Important: You are highly advised to attend the first session. Furthermore, if you do not work on this lab during the first week expect a heavy workload for the last one (we estimate 6h of work per person per week). Thus, you are the ultimate responsible for a reasonable scheduling of this session.

Instructions

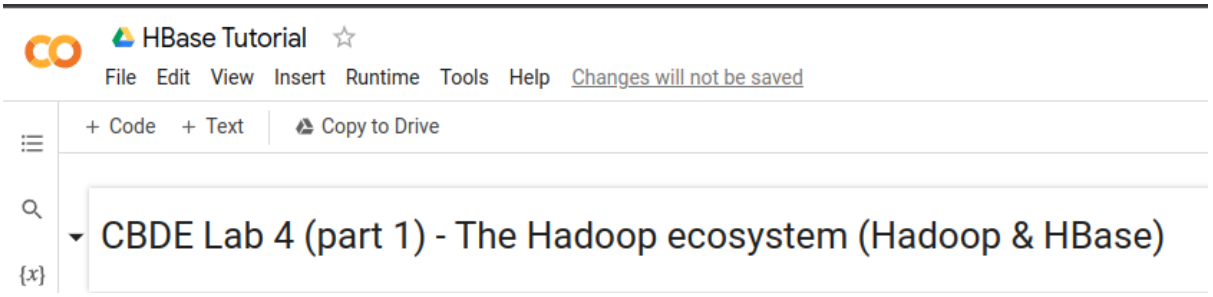
This session consists of two independent parts (i.e., HBase and MapReduce). Both parts will be performed in notebook environments, precisely in Google Colaboratory notebooks. Thus, familiarity with the Python programming language is expected. For each part, a tutorial is provided showcasing in practice the concepts that have been studied in the lecture sessions.

Delivery

This lab will be assessed as follows: HBase (4 pt) + MapReduce (6 pt). Each line (or blocks of code) of your implemented code has to be commented on to let the lecturer understand what you are doing. Upload the corresponding .ipynb files for each notebook, this can be achieved in the contextual menu “File” > “Download” > “Download .ipynb”.

Session assignment

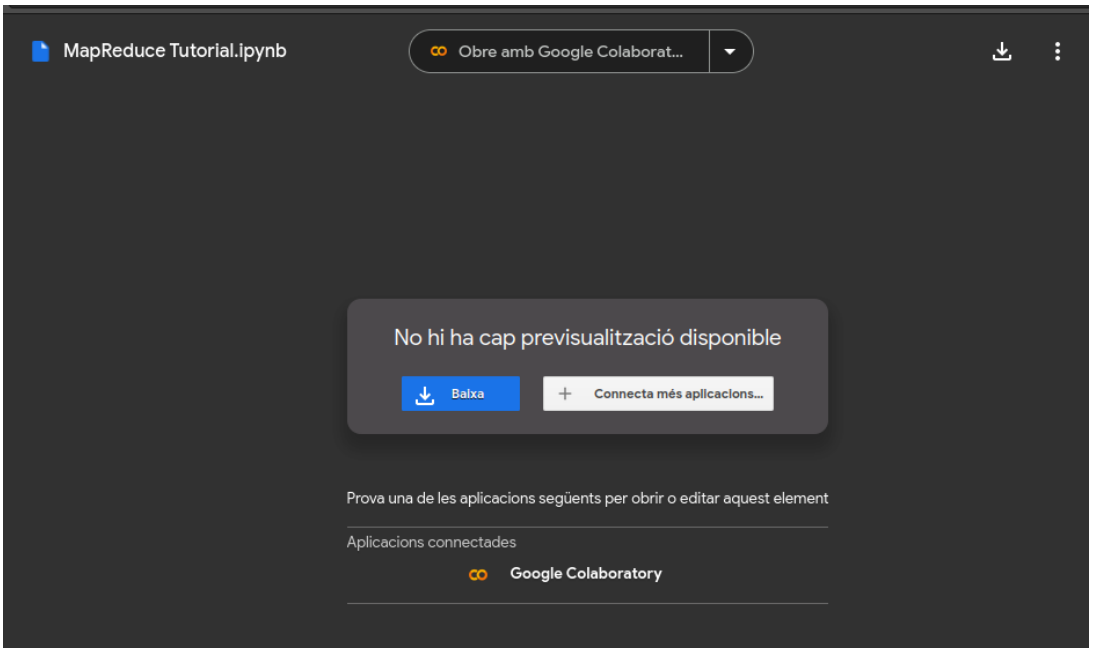
By the beginning of the lecture you will be provided *read-only access* to the notebook template corresponding to the tutorial and assignment to your UPC account. This notebook cannot be modified, thus you should make a copy to your Google Drive using the “Copy to Drive” button shown below.



Each notebook contains the set of exercises to be implemented. We encourage you to, first, understand and get familiar with the provided tutorials.

Troubleshooting

It is possible that when you load a notebook it does not automatically display the Colab interface. For that, click on the “Open with Google Colaboratory” button as shown below.



If the button is not available it means you need to “Connect” your Google Drive with Google Colaboratory. This only needs to be done once.

Remember: all non-obvious implementation decisions must be explained, either as comments in the code, or as text.