José Mariano Portilla Landa 163449 26/03/2022

Executing Map Reduce Programs on Hadoop Environments

Propose a UML component diagram of the Hadoop environment installed on Collab

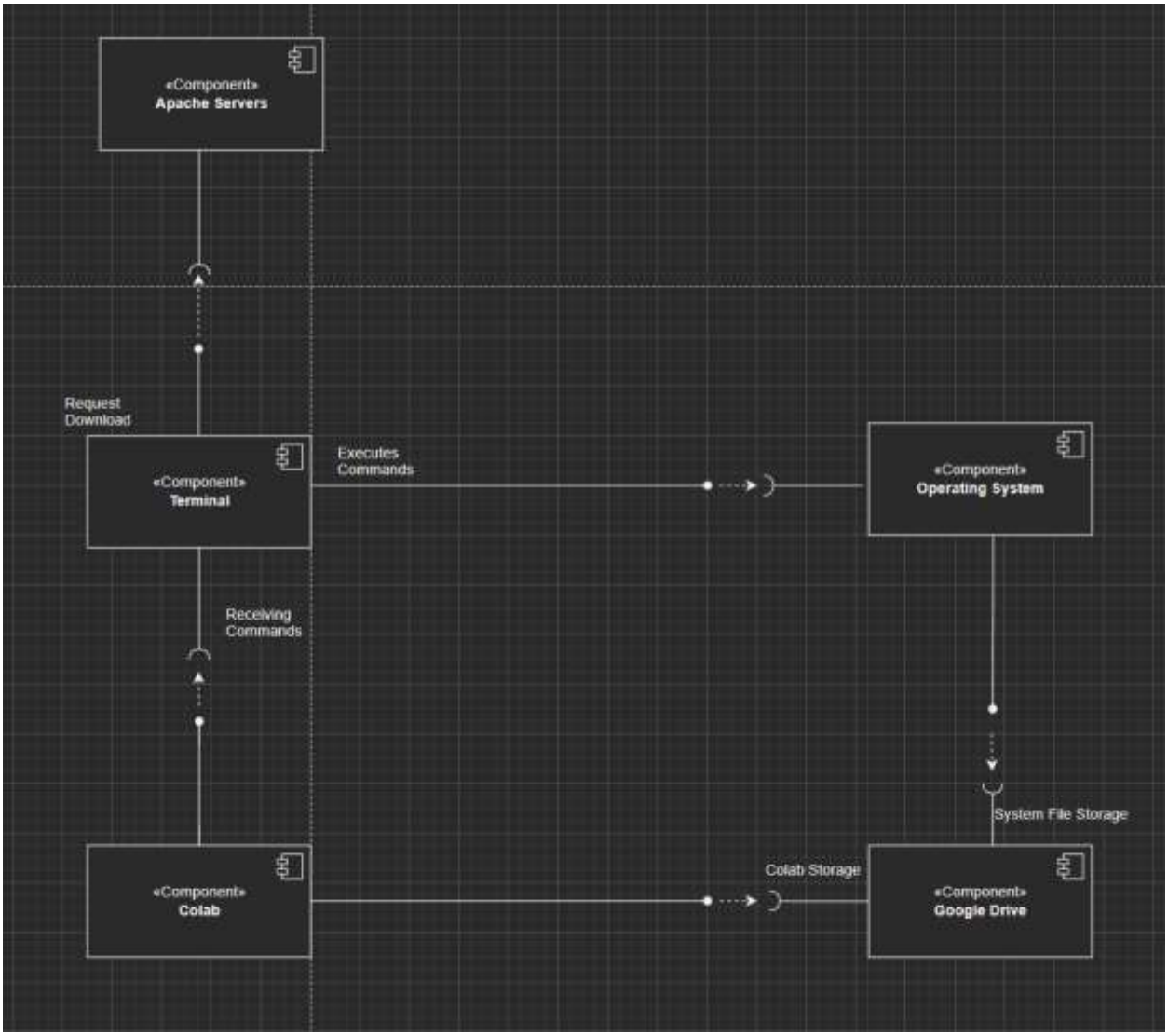

Figure 1: UML component diagram of Hadoop environment installed on colab.

Propose a UML component diagram of the two map-reduce count words programs tested in the lab

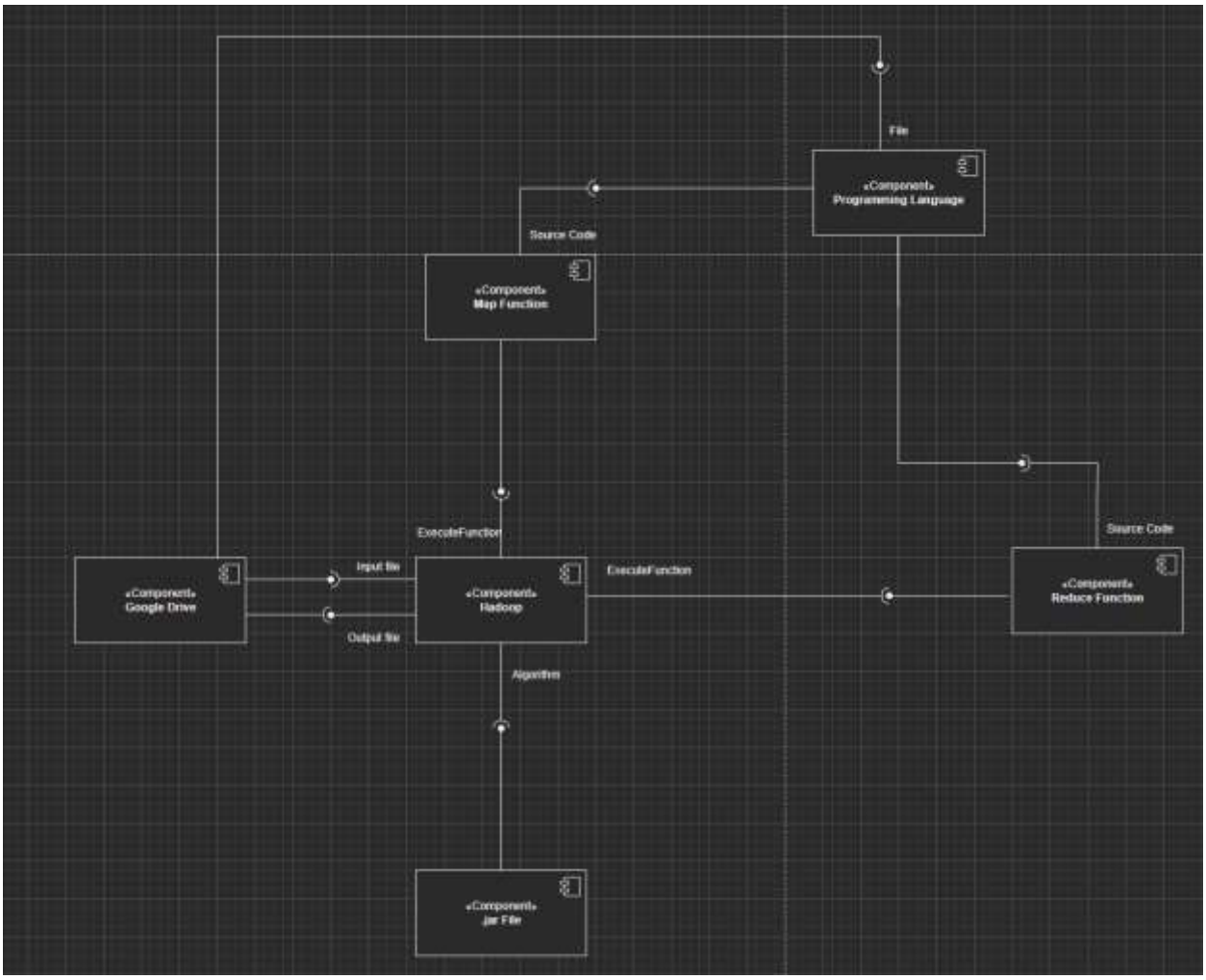

Figure 2: UML component diagram of two map-reduce count words programs.

Explain how the first example implementing a grep operation with a regular expression is executed

First it gathers general information about the process, such as the total input files to process (10) or the number of splits (10). Then it goes file by file making the map process with some defined attributes, such as bufstart being 0 and bufvoid being 104857600. After that it displays some information about the file system counters (such as the number of bytes read and written), the Map-Reduce framework (such as the map input and output records, or the spilled records and failed shuffles), and the file input format counters (bytes read). When the process ends with all of this,

the map task executor is complete and reduce tasks may start by shuffling the data for each file. Then the files are merged and information is displayed for the file system counters, the map-reduce framework, shuffle errors and file output format counters. After the reduce task executor is complete the grep function is called to find the provided pattern and the output is saved in the specified file.

Explain the way the program "count words" is executed in the example

To execute the Wordcount program using python we defined our own map and reduce functions that were passed to the command line arguments to initialize the program. The process is basically the same as above, we are initiating with map tasks and then finishing with reduce tasks and saving the output, the difference is that we are not merging files, we are using a downloaded dataset, and we are programming our own map and reduce functions. In the map function we are only keeping words that are conformed by letters or numbers and we are eliminating punctuations and stop words, while in the reduce function we are just grouping equal words by counting their number of occurrences.

What is the role of google drive in these examples?

Google drive is responsible to store all the files needed by Hadoop, our environment, python files, our dataset, and our output from running the examples.