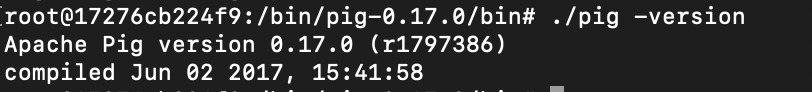
Practical 6

Dawid Skraba 19433692

***Exercise 2.***

Task 0:



I installed pig in the container and went inside bin to check its version, showing that its installed correctly.

Task 1:



Graphical user interface, text

Description automatically generated

Text

Description automatically generated

In the above command I loaded in my “web\_access” file from the /test/input folder in HDFS. I used the CSVExcelStorage command with certain parameters to omit the first row of the column of the file. I did this as this row is used for the column headers. The ‘AS’ command specifies the types of the values in the columns. Using the LIMIT command I showed the correct output. So now the value ‘A’ holds the correct values which I will use later to produce the desired output.

Then for each line in the loaded data A, I multiply the given time by a 1000 first as the Unix timestamp is given in milliseconds. Using the ToDate function with the specified format I convert it to that format, and then change it back to a string for easy reading. In B I only store the time and the bytes for each date. This is simply to keep things simple as we only need those two headers. In D I group all my tuples by time so that in the next line I can add the bytes for each date given. I then write the results into hdfs storage for easy reading. The result we can see below. For each Date we can see the totalled bytes.

Text

Description automatically generated

Text

Description automatically generated

My final code:

Text

Description automatically generated

Then after having all these terms I order them by the header: total\_bytes which contains the totalled bytes for each date in descending order. This will give us the most accessed dates at the top and go downward. Using the LIMIT function we can get the 3 most accessed dates. I then also store the result to a file.

The 3 most accessed dates are:

Text

Description automatically generated