How to set up okeanos machines Hadoop and Spark

First, we setup our machines, the master, and the slave, following the guide Okeanos_setup.pdf. After we have completed the setup we have a Public IP, in our case it is 83.212.80.104, and for each machine we a hostname and a password.

For our machines those are:

master:	slave:
hostname: snf-34535.ok-kno.grnetcloud.net	hostname: snf-34536.ok-kno.grnetcloud.net
password: Ap1grlfJ1O	password: gWywEflj75

We connect to master through ssh and follow this guide https://sparkbyexamples.com/hadoop/apache-hadoop-installation/ in order to first enable passwordless login between master, which will be our Name Node, and slave, which will be our Data Node. Now we can connect to the slave by running the command: ssh slave in the masters' terminal. Then we continue following the guide and install JDK, Hadoop in both machines, the configure the Hadoop Cluster, format HDFS on the master (Name Node) and start the HDFS cluster. By running jps command we can see that everything is setup properly and can now see the web UI by accessing http://83.212.80.104:9870/

Then we move on to the python and Spark installation by following the guide Spark_Install_instructions.pdf skipping the deployment of workers with custom recourses. We deploy one worker per machine and now if we run jps command we will see that there is a worker running. Also we can now access the web UI http://83.212.80.104:8080/

In order to run the script, which should be located at the folder /home/user/spark-3.1.3-bin-hadoop2.7/bin, by running the command *python3.8 script.py*. The result will be saved at the HDFS and we can copy them to a local folder using the command *hdfs dfs -get results*, where results is the hdfs folders where all the output files have been saved.