High Performance Computing Exercise Sheet 13

FS 25
Dr. Douglas Potter

http://www.astro.uzh.ch/
Teaching Assistants:
Cesare Cozza, cesare.cozza@uzh.ch
Mark Eberlein, mark.eberlein@uzh.ch

Issued: 12.12.2025
Due: 19.12.2025

In this exercise session, we will practise how to use `Hadoop` to perform the basic data-handling operations such as map and reduce. We will go through a very similar procedure as the one discussed during the lecture.

Exercise 1 Hadoop - setting things up

Launch a VM instance on the UZH Science Cloud https://cloud.science-it.uzh.ch using the `hadoop` snapshot. Don't forget to delete the instances created during previous exercise sessions (you can create a snapshot in order to resume the status of previous instances). On your machine (your laptop), open the config file: `$HOME/.ssh/config` and add the following lines:

```
Host hadoop
  Hostname <VM instance IP address>
  User ubuntu
  ForwardX11 yes
  ForwardX11Trusted yes
  ForwardAgent yes
  LocalForward 9870 localhost:9870
  LocalForward 9864 localhost:9864
  LocalForward 8088 localhost:8088
  LocalForward 8888 localhost:8888
```

The above config entry will make an `hadoop` alias for ssh-ing into the VM image hadoop which has been created and set correctly the port forwarding. You can then connect to the created image via `ssh hadoop`. Once you are connected to Hadoop image, start distributed file system `start-dfs.sh` and start yarn `start-yarn.sh`. [1] Now, you should be able to visit

[1] These files are in `$PATH`, just type the name of the file on the shell. Type `start` on the shell and then `Tab Tab` and you will see them.

the Hadoop api from your web browser using the address `http://localhost:8088`.

Exercise 2 Map&reduce

Once you accessed hadoop image, you can perform map&reduce operation on the data files in the `exercise_session_13` folder of the course repository. Address/perform the following tasks:

- write a script to run the map&reduce procedure locally (not using `Hadoop` instance, e.g. on Eiger) on the files provided in the course repository. Note, that python scripts are compatible with Python 2.X, which is the default Python version after logging to Eiger.
- create a new directory on the Hadoop file system called `DonaldTrump` and copy the data files into it.
- run the script `pymapred.sh` to perform the map&reduce procedure on the Hadoop file system:

  ```
  hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar\
      -files mapper.py,reducer.py\
      -mapper mapper.py -reducer reducer.py\ -input DonaldTrump/* -output output
  ```

  How many map and reduction operations were performed? How many lines/entries were treated?
- which output folder(s)/file(s) were created? Find the file containing the information about the word counts and have a look on the results of map&reduce operation.
- visit the `hadoop` api from your local machine. Can you find the output files?

Commit: Push your answers and an output file to your repository.