Using Hadoop at Télécom Paris

1 Connecting to the HADOOP at Télécom

1.1 Accessing HADOOP

Accessing HADOOP is done by connecting to the machine `lame11.enst.fr`. For security reasons, `lame11` is not directly reachable from the internet. You can access the machine from inside the Télécom network or you can proxy with VPN (or eventuall with ssh through `ssh.enst.fr`).

1.2 Connecting to lame 11

On Unix (Linux/Mac) with ssh installed this can be done with the following command (where login is replaced by your login)

```
ssh login@lame11.enst.fr
```

On windows, you can either use the ssh command if it is installed or download and use KiTTY. With KiTTY you need to set the host to `lame11.enst.fr`

1.3 Initialize the environment

When you open a new connection to `lame11`, the HADOOP configuration is not loaded. Therefore HADOOP cannot be directly accessed. To load HADOOP configuration you need to enter the following command :

```
source /infres/ir510/hadoop/bin/hadoop_env
```

2 Playing with HADOOP

2.1 Browing the filesystem

HADOOP filesystem is organized as a Unix filesystem. This means that files and folders form a treelike hierarchy starting from the root and branching at each folder. You can access a file with its *path*. For instance the path `/datasets/ban` means that you start from the root (noted `/`) then you go down to the folder `datasets` then you access the folder `ban`.

To show the files in a folder located at path `a/b/c` you can use the command `hadoop fs -ls /a/b/c`
For instance, to see the files below the path `/datasets/ban` :

```
hadoop fs -ls /datasets/ban
```

and the first lines it prints are :

```
Found 102 items
-rw-r--r--   3 root supergroup  146170315 2019-09-04 17:32 /datasets/ban/ban-01.csv
-rw-r--r--   3 root supergroup  140079632 2019-09-04 17:32 /datasets/ban/ban-02.csv
-rw-r--r--   3 root supergroup  105002484 2019-09-04 17:32 /datasets/ban/ban-03.csv
-rw-r--r--   3 root supergroup   46720937 2019-09-04 17:32 /datasets/ban/ban-04.csv
-rw-r--r--   3 root supergroup   32827000 2019-09-04 17:32 /datasets/ban/ban-05.csv
-rw-r--r--   3 root supergroup  125327386 2019-09-04 17:32 /datasets/ban/ban-06.csv
-rw-r--r--   3 root supergroup  113968989 2019-09-04 17:32 /datasets/ban/ban-07.csv
```

Here, each line describes a file. In the first line, for example, the part `-rw-r-r-` describes the permission : who can read or write the file ? `root` describes the owner of the file, `146170315` is the size (in bytes) of the file. `2019-09-04 17:32` is the last modification date of the file and finally `/datasets/ban/ban-01.csv` is the path of the file.

2.2 Handling files from the command line

Reading a file *that is not a folder* can be done by using the `cat` command. For instance, to read the file `/datasets/ban/ban-01.csv` :

```
hadoop fs -cat /datasets/ban/ban-01.csv
```

Note that this command will output the whole file and it will be transmitted to your computer. Therefore only use cat for very small files (here the file is pretty big with 140MB). You can look at the beginning of a file with the command `head` :

```
hadoop fs -head /datasets/ban/ban-01.csv
```

Similarly, you can look at the end of a file with the command `tail` :

```
hadoop fs -tail /datasets/ban/ban-01.csv
```

2.3 Moving data in and out of the HADOOP cluster

The command `copyToLocal` takes a path on the HADOOP filesystem and copies the folder (or file) to the local machine. The second argument to the command specifies the name of the file you downloaded, for instance :

```
hadoop fs -copyToLocal /tmp/distantFile localName
```

copies the file `/tmp/distantFile` on the HADOOP cluster into the file `localName`.

Conversely putting data on HADOOP can be done using the command `copyFromLocal` whose first argument is the local file and whose second argument is the name of the copied file on HADOOP :

```
hadoop fs -copyFromLocal localFile /tmp/distantFile
```

2.4 Moving files in and out of the `lame11` machine

To run your code on HADOOP you need to file upload it to `lame11`. In order to do so you can use a graphical interface with SFTP (such as filezilla, recommended) or use the command scp.

3 Submitting a job to the HADOOP

3.1 With HADOOP streaming

The HADOOP streaming allows only for one MapReduce job at a time. The command should specify the input and output files, the command for the mapper and for the reducer. The command can also specify which files to distribute (command line generally fit in one line but when they don't you can split them on several line using a \) :

```
mapred streaming -input /path/in -output /path/out \
       -mapper 'mapper command' -reducer 'reducer command' \
       -file localFile1 -file localFile2
```

3.2 With Java program

First you need to compile your Java code with a classpath containing the HADOOP classpath. Unless you installed HADOOP locally you therefore need to compile on `lame11`.

```
hadoop jar JAR_FILE.jar my.classe.java.Class arg[0] arg[1] ...
```