Cluster

We have a small Hadoop cluster for this course, based on Cloudera *(http://www.cloudera.com/)* express.

Connecting Remotely

The goal here is to connect to `cluster.cs.sfu.ca` by SSH. Since you can't connect directly from the outside world, it's not completely straightforward.

Off-Campus Prerequisite: VPN

If you are on-campus (in the lab, or on campus wi-fi), this is not necessary.

If you are off-campus, you need to activate the SFU VPN *(https://www.sfu.ca/information-systems/services/sfu-vpn.html)* to get access to the cluster (and other network resources that are restricted to campus-only access).

Option 1: the right way

If you don't already have one, create an SSH key so you can log in without a password. The command will be like this, and **accept the default output filename**:

```
ssh-keygen -t ed25519 -N ""
```

Then copy your public key to the server:

```
ssh-copy-id -p24 <USERID>@cluster.cs.sfu.ca
```

Create or add to the `~/.ssh/config` (on your local computer, not the cluster gateway) this configuration that will let you connect to the cluster by SSH. Then you can simply `ssh cluster.cs.sfu.ca` to connect.

```
Host cluster.cs.sfu.ca
  User <USERID>
  Port 24
  LocalForward 8088 controller.local:8088
  LocalForward 9870 controller.local:9870
  LocalForward 18080 controller.local:18080
```

With this configuration, port forwards will let you connect (in a limited unauthenticated way) to the web interfaces:

- HDFS namenode: http://localhost:9870/ *(http://localhost:9870/)*
- YARN application master: http://localhost:8088/ *(http://localhost:8088/)*
- Spark job history server: http://localhost:18080/ *(http://localhost:18080/)*

Once it's set up, you should be able to copy files and connect remotely quickly:

```
scp wordcount.jar cluster.cs.sfu.ca:
ssh cluster.cs.sfu.ca
```

Option 2: just get it working

You will be connecting to the cluster a lot: you will want to get things set up more nicely to make your life easier later. But, this should at least *work*.

You generally just need to SSH to `cluster.cs.sfu.ca` (substituting whatever SSH method you use on your computer):

```
[yourcomputer]$ ssh -p24 <USERID>@cluster.cs.sfu.ca
[gateway] $
```

Once you're connected to the Hadoop gateway, you can start running `hdfs` and `yarn` commands.

You will also frequently need to copy files to the cluster:

```
[yourcomputer]$ scp -P24 assignment.jar <USERID>@cluster.cs.sfu.ca:
```

If you need access to the web frontends in the cluster, you can do the initial SSH with a much longer command including a bunch of port forwards:

```
ssh -p24 -L 8088:controller.local:8088 -L 9870:controller.local:9870 <USERID>@cluster.cs.sfu.c
```

Job Logs

If you have set up your SSH config file as in the Cluster instructions, you can see the list of jobs that have run on the cluster at http://localhost:8088/ *(http://localhost:8088/)* .

Then at the command line, use the application ID from that list to get the logs like this:

```
yarn logs -applicationId application_1234567890123_0001 | less
```

Cleaning Up

If you have unnecessary files sitting around (especially large files created as part of an assignment), please clean them up with a command like this:

```
hdfs dfs -rm -r /user/<USERID>/output*
```

It is possible that you have jobs running and consuming resources without knowing: maybe you created an infinite loop or otherwise have a job burning memory or CPU. You can list jobs running on the cluster like this:

```
yarn application -list
```

And kill a specific job:

```
yarn application -kill <APPLICATION_ID>
```

Updated Fri Aug. 25 2023, 15:55 by ggbaker.