COMP4651

Lab3: Docker and Hadoop API

Yuheng ZHAO
yzhaoep@connect.ust.hk
Spring 2023

THE DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

Why Docker?

Ready-to-use, plug-and-play

- installing and deploying a Hadoop cluster can be a challenging job for beginners
- you will try it in Lab 5

A consistent development environment

- no “work on my machine” excuse

Convenient for coding and debugging

VM based solutions have some suitability problem, especially VirtualBox may have problem running on M1 Mac

Prerequisites

If you have problem installing or running docker on your own device, try following Appendix here to try our AWS solution.

Install Docker & Run Docker Service

- Getting start with Docker, choose the right version of Docker Desktop for your machine.

Pull Docker Image

- TA has packed useful docker image with Hadoop service already configured
- Use the following command to pull the required docker image:
  `docker pull qpswwww/quickstarts:v0`

Download Exercise Code

- Clone the code for exercise 1

Start Container

Windows users plz press win+x and pick cmd.exe under administrator mode, and don't need sudo to require root permissions

Use the following command to run a container:

sudo docker run --hostname=quickstart.cloudera --restart unless-stopped --privileged=true -it -v your_directory_of_exercise_code:/share -p 8888:8888 -p 80:80 -p 8020:8020 -p 7180:7180 -p 50070:50070 -p 10086:22 qpswwww/quickstarts:v0 /bin/sh

You can also use the following command to find a running container and access with exec command:

docker container ls

sudo docker exec -it YOUR_CONTAINER_ID /bin/bash

```
farland@Farlands-MacBook-Pro-6 ~ % docker container ls
CONTAINER ID   IMAGE                          COMMAND                  CREATED       STATUS         PORTS
059986608a03   qpswwww/quickstarts:v0         "/bin/bash"              5 hours ago   Up 7 minutes   0.0.0.0:80->8
cd653a4ab91a   a422e0e98235                   "/metrics-sidecar"       5 hours ago   Up 5 hours
craper-64bcc67c9c-h9jm2_kubernetes-dashboard_2303a1ce-f354-4fb0-ad07-a54c3f0c94eb_10
396d5d7fd959   kindest/node:v1.25.3           "/usr/local/bin/entr…"   8 weeks ago   Up 5 hours
1f5d49fc739e   kindest/node:v1.25.3           "/usr/local/bin/entr…"   8 weeks ago   Up 5 hours     127.0.0.1:5046
faca593fa203   kindest/node:v1.25.3           "/usr/local/bin/entr…"   8 weeks ago   Up 5 hours
1b1d43fe7599   ghcr.io/k3d-io/k3d-proxy:5.4.6 "/bin/sh -c nginx-pr…"   8 weeks ago   Up 5 hours     80/tcp, 0.0.0.
587b0c547a6a   rancher/k3s:v1.24.4-k3s1       "/bin/k3d-entrypoint…"   8 weeks ago   Up 5 hours
farland@Farlands-MacBook-Pro-6 ~ % sudo docker exec -it 059986608a03 /bin/bash

[root@quickstart /]# echo "Now We're In"
Now We're In
[root@quickstart /]#
```

Restart Service

Use the following command to restart the service we already deployed:

sudo service hadoop-yarn-resourcemanager restart

sudo service hadoop-hdfs-namenode restart

sudo service hadoop-hdfs-datanode restart

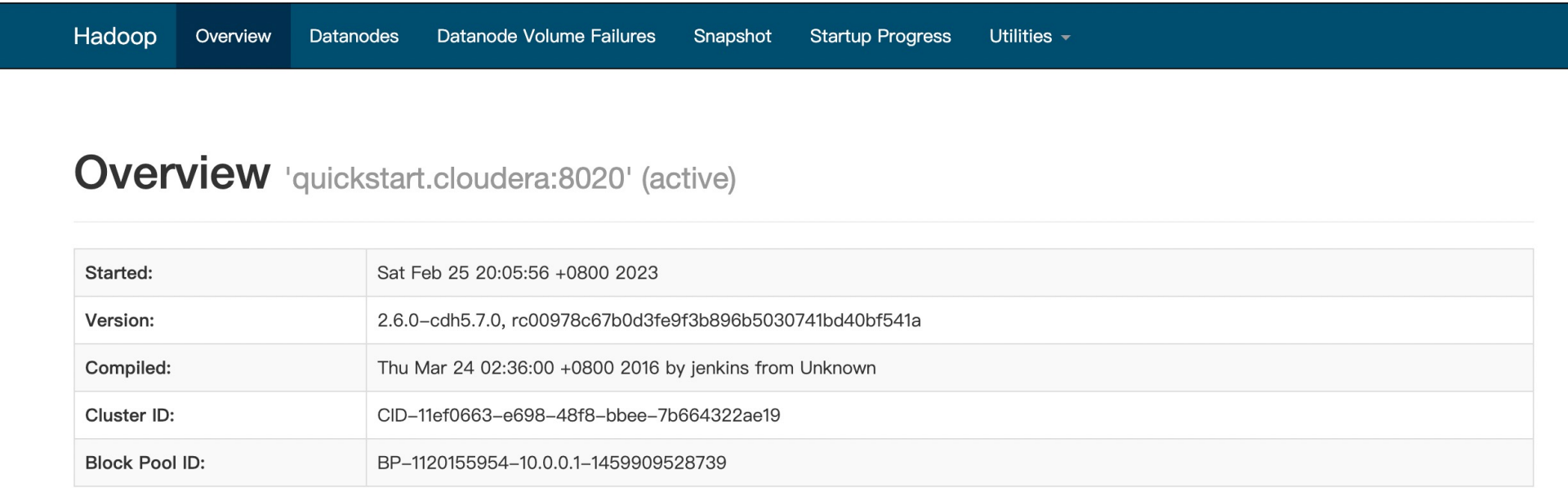
Always restart the service when starting container from a docker image

```
[root@quickstart /]# sudo service hadoop-yarn-resourcemanager restart
no resourcemanager to stop
Stopped Hadoop resourcemanager:                            [  OK  ]
starting resourcemanager, logging to /var/log/hadoop-yarn/yarn-yarn-resourcemanager-quickstart.cloudera.out
Started Hadoop resourcemanager:                            [  OK  ]
[root@quickstart /]# sudo  service hadoop-hdfs-namenode restart
no namenode to stop
Stopped Hadoop namenode:                                   [  OK  ]
starting namenode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-namenode-quickstart.cloudera.out
Started Hadoop namenode:                                   [  OK  ]
[root@quickstart /]# sudo service hadoop-hdfs-datanode restart
no datanode to stop
Stopped Hadoop datanode:                                   [  OK  ]
starting datanode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-datanode-quickstart.cloudera.out
Started Hadoop datanode (hadoop-hdfs-datanode):            [  OK  ]
[root@quickstart /]#
```

HDFS Web UI

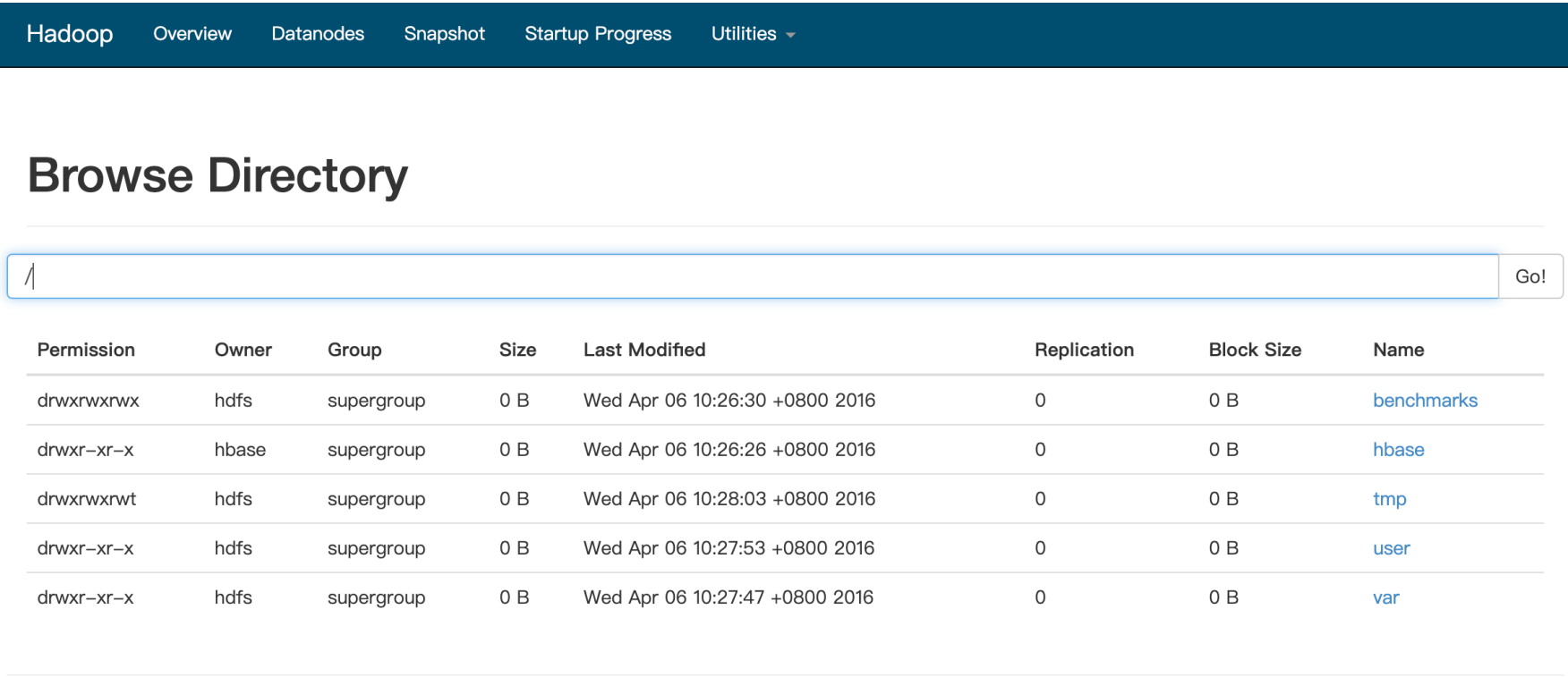
In the browser, enter http://localhost:50070

- HDFS NameNode, Secondary NameNode, DataNode



NameNode Web UI

Browse HDFS filesystem through NameNode UI (Utilities)



Datanode Information

In service Down Decommissioned Decommissioned & dead

In operation

Show 25 entries Search:

Node	Last contact	Capacity	Blocks	Block pool used	Version
quickstart.cloudera (172.17.0.2:50010)	Sat Feb 25 20:16:16 +0800 2023	58.37 GB	903	809.5 MB (1.35%)	2.6.0-cdh5.7.0

Showing 1 to 1 of 1 entries Previous 1 Next

Decommissioning

Node	Under replicated blocks	Blocks with no live replicas	Under Replicated Blocks In files under construction

HDFS Command Line

See what you have under these directories, and their permissions, owners, last modification time, etc:

- home: hadoop fs -ls
- root: hadoop fs -ls /
- var: hadoop fs –ls /var

You may compare the results with Web UI

You can perform the same operation on your local filesystem: **hadoop fs -ls file:///localDir**

Copy a local file to HDFS

Randomly generate a 10MB dummy file:

- head -c 10M < /dev/urandom > dummy

Copy it to HDFS:

- hadoop fs -put dummy .

See what you have now in your home directory:

- hadoop fs -ls

Get it back to your local disk

Use the following command:

- hadoop fs -get dummy dummy2

Check if the copy is the same thing as the original by comparing their md5sum:

- md5sum dummy
- md5sum dummy2

Try other commands

hadoop fs -cat fileName

hadoop fs -mkdir dirName

hadoop fs -rmdir dirName

hadoop fs -rm fileName

Exercise-1

You are required to write a simple code that copies a file from HDFS to the local disk, using HDFS APIs.

- This exercise gets you familiar with Hadoop APIs
- Find the detailed information here: hkust-comp4651-23s/Exercise-1: Hadoop exercise following lab 3 (github.com)

THE END

Play with the exercise code on your own