Brandeis University
Department of Computer Science
COSI 129a - Introduction to Big Data Analysis
Fall 2016

Installing Hadoop on a Docker Container

There are many ways to install Hadoop on your machine, and you may pick any way you want. But the simplest way, especially if you use Windows, is to use a Docker container.

1 Installing Docker

To install Docker get the installer at

```
https://www.docker.com/products/docker
```

Note that if you have a Windows version < Win 10 or a MacOX version < 10.3.3, then you need to get the Docker toolbox from

```
https://www.docker.com/products/docker-toolbox
```

2 Run Docker and install Hadoop

You can install a docker image by going to your terminal (for example, cmd on Windows, terminal on MacOS, or the Docker Quickstart terminal). You first need to pull the Docker Hadoop image:

```
docker pull sequenceiq/hadoop-docker:2.3.0
```

We are using version 2.3.0 because that is the version installed on the cluster. Let it finish downloading the Hadoop image, after which you can run the docker instance:

```
docker run -it sequenceiq/hadoop-docker:2.3.0 /etc/bootstrap.sh -bash
```

It will start a docker container and run bootstrap bash file on it. The bootstrap command basically just start hdfs and yarn. A standalone hadoop instance is up and running. Note that the installation contains a file named disk.vmdk which by default is about 20GB, let us know if this is a problem, there is a setting that controls the size.

If you like to use Kitematic in the Docker Toolbox, do the following:

- If you use Docker toolbox, install docker-image by open Kitematic (a component of Docker toolbox):
- If there is some error, try Run with Virtual Box on GUI. If the error persists, it is very likely that you need to go to BIOS of your machine (laptop or desktop) and enable virtualization.
- On the view to add a new container, search for sequenceiq/hadoop-docker
- Click the Create button.
- Now you should see a view (Container logs), click on Button (Exec), which is between Restart and Docs, it would open a terminal, move to step 3.

3 Working on your container from the command line

Let's go to the Hadoop directory and check if there is anything on hdfs:

```
$ cd $HADOOP_PREFIX
$ bin/hdfs dfs -ls
```

This should print one directory named "input" on hdfs. Now run a jar file on the hadoop instance (note that the following is really one line that spills over):

```
$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.3.0.jar grep
input output 'dfs[a-z.]+'
```

This runs the Grep class inside the jar file, which starts a MapReduce job that takes the "input" directory as input and outputs into a new directory "output" on hdfs. You can check the output by doing:

```
$ bin/hdfs dfs -cat output/*
```

Your output should look like

```
6 dfs.audit.logger 4 dfs.class
3 dfs.server.namenode.
2 dfs.period
2 dfs.audit.log.maxfilesize
2 dfs.audit.log.maxbackupindex
1 dfsmetrics.log
1 dfsadmin
1 dfs.servers
1 dfs.replication
1 dfs.file
```