Docker-MR.md

Data Intensive Computing

CSE 4/587 - Spring 2019

Docker image for MR over Hadoop and Running a simple Wordcount program

Overview

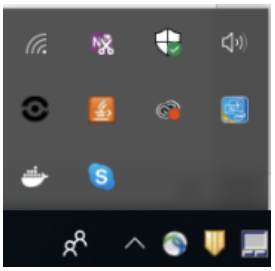
In this document you will learn about downloading and installing a Docker image for Hadoop and execute a simple map-reduce program to illustrate how to run MR programs on Hadoop.

1. Install Docker

- Go to docker installation page and select the appropriate OS that you are currently using and follow the instructions

2. Increase RAM size for Docker

On Windows:
On the right end of your task bar, right click on Docker image → Settings → Advanced and select the RAM size to 8GB. Close and Restart your system. Can you spot the docker icon in the picture below?



On Mac On the top of your menu bar you can check docker icon, click on that → Preferences → Advanced → select the RAM size to 8GB. Close and Restart your Docker. Can you spot the docker icon in the picture below?

3. Check Installation

- Go to your terminal / powershell and run the following command

```
$ docker run hello-world
```

- You should get the following output, if you don't check your installation:

```
Hello from Docker!
This message shows that your installation appears to be working correctly.
To generate this message, Docker took the following steps:
1. The Docker client contacted the Docker daemon.
2. The Docker daemon pulled the "hello-world" image from the Docker Hub.
(amd64)
3. The Docker daemon created a new container from that image which runs the
executable that produces the output you are currently reading.
4. The Docker daemon streamed that output to the Docker client, which sent it
to your terminal.
To try something more ambitious, you can run an Ubuntu container with:
$ docker run -it ubuntu bash
Share images, automate workflows, and more with a free Docker ID:
https://hub.docker.com/
For more examples and ideas, visit:
https://docs.docker.com/get-started/
```

- You can also check the images docker on your machine has by running. You should see `hello-world` image and other images if any in the directory. `$ docker images`

4. Pull the Cloudera-Quickstart docker Image

- This docker image manages all your dependencies and you can run all your MapReduce programs easily.
- Run `$ docker pull cloudera/quickstart:latest` this should take a while as the docker container is of about 4 GBs

5. Create local directory (on your laptop) for your MR solution and data

- Create a local directory in your C: drive on Windows (for example `C:\Users\bina\Documents`), ~ or CSE487 or CSE586 on Linux or Mac, and create a directory `dockerMR` for example. This is where you will add your local data, and map and reduce programs of Lab2. This is your local workspace. Later you will map it to your docker and Hadoop file system. The “dockerMR” for example has,

```
mapper.py
reducer.py
data
```

6. Configure docker directory and map to local workspace

- Run the following command. If it errors out, restart docker on your machine and then try again.

```
$ docker run --hostname=quickstart.cloudera --privileged=true -t -i -v localpath:/src --publish-
all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart
```

replace `localpath` with a location you want to map

for example
On a Unix/Linux system:

```
$ docker run --hostname=quickstart.cloudera --privileged=true -t -i -v
/Users/xyz/Documents/SomeFolder:/src --publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-
quickstart
```