CSE291E-lab3

Running Hadoop map reduce in docker

– Zhe Wang (A53097553) zhw176@cs.ucsd.edu
– Yuning Hui (A53091905) y3hui@eng.ucsd.edu

Input file (pseudo distributed mode)

Input.txt:

```
line1. line1. line1!
line2, line2, line2
line3 line3..
```

Input file (cluster mode)

Input1.txt:

```
Hello Hadoop! xixixi
Hello Docker! hehehe
```

Input2.txt

```
Hello Docker!
Hello World!
```

Input3.txt

```
Hello Hadoop!
Hello World!
```

How to run the demo in a pre-configured

server (For CSE291E staff, available until June 9)

This project depends on some third-party docker images and need `sudo` privilege. If you don't want to download those images or running in `sudo` mode on you machine, you can run it on a server that I've already configured.

- log in the server: `ssh course_staff@291elab3.philosopherwang.me`, password is `cse291e`.
- `cd lab3`
- Run the demo in single node pseudo distributed mode:
    - `cd single`
    - `sudo ./build.sh`, and see the result, compared with input file (pseudo distributed mode) above
    - `cd ..`

- Run the demo in a docker cluster:
    - `cd cluster`
    - `sudo ./bootstrap.sh`, you'll login to the master node of the cluster
        - To see all cluster members `serf members`
        - To run word count demo `/data/wordcount.sh`, compared with input file (cluster mode) above.
        - To run bigram demo `/data/bigrams.sh`, compared with input file (cluster mode) above.

    - Exit session with master node `exit`

- Exit the session with the server: `exit`

How to configure and run the project on your own machine

Before running, change mode:

- Change current directory to the project root directory `cd ...lab3`
- Run a script that recursively change mode of any `.sh` and `.py` files in current directory
    - `chmod a+x change_mode.sh`
    - `./change_mode.sh`

How to run word count in a single docker

container using Hadoop pseudo distributed mode

- Change current directory: `cd /XXX.../lab3/single/` (change `XXX...` according to your machine)
- Build docker image and run the task: `sudo single/build.sh`, wait until finish, then you will see the result.
- To clean up, execute `sudo ./clean.sh`

How to run Hadoop in cluster mode using docker

- Note: this section is based on the work of Kai LIU: https://github.com/kiwenlau/hadoop-cluster-docker
- Configure the environment
    - Change current directory `cd cluster`
    - Pull the dependent docker image `sudo ./boost.sh`
    - Build the data volume image `sudo ./build-image.sh data-volume`, which includes most our code.
        - WARN: this script will remove all running / waiting containers, in order to work properly.
    - Start cluster containers `sudo ./bootstrap.sh`, now you will be in a session with the master node of a Hadoop cluster with 4 slave nodes.
- Run the same word count program as last section in cluster mode
    - `/data/wordcount.sh`
- Run the bigram count program in cluster mode
    - `/data/bigrams.sh`

Reference

- Single node pseudo distributed mode is based on: https://hub.docker.com/r/sequenceiq/hadoop-docker/
- Cluster mode is based on: https://github.com/kiwenlau/hadoop-cluster-docker
- Word count example: https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#MapReduce_Tutorial