Hadoop Docker

Clone this repo with the below command

git clone https://github.com/ppfenning/docker-hadoop.git

Supported Hadoop Versions

See repository branches for supported hadoop versions (using 3.2.3)

Pre-requisites

Git: https://git-scm.com/book/en/v2/Getting-Started-Installing-Git
Docker: https://docs.docker.com/engine/install/

Quick Start

To deploy an example HDFS cluster, run:

```
docker-compose up
# docker-compose up -d (if background run is desired)
```

Run example wordcount job:

```
make wordcount
```tack deploy -c docker-compose-v3.yml hadoop
```

docker-compose creates a docker network that can be found by running docker network list, e.g. dockerhadoop_default.

Run docker network inspect on the network (e.g. dockerhadoop_default) to find the IP the hadoop interfaces are published on. Access these interfaces with the following URLs:

- Namenode: http://<dockerhadoop_IP_address>:9870/dfshealth.html#tab-overview
- History server: http://<dockerhadoop_IP_address>:8188/applicationhistory
- Datanode(s): http://<dockerhadoop_IP_address>:$DATA_PORT/
- Nodemanager: http://<dockerhadoop_IP_address>:8042/node
- Resource manager: http://<dockerhadoop_IP_address>:8088/

Configure Environment Variables

The configuration parameters can be specified in the hadoop.env file or as environmental variables for specific services (e.g. namenode, datanode etc.):

```
CORE_CONF_fs_defaultFS=hdfs://namenode:8020
```

CORE_CONF corresponds to core-site.xml. fs_defaultFS=hdfs://namenode:8020 will be transformed into:

```
<property><name>fs.defaultFS</name><value>hdfs://namenode:8020</value></property>
```
```

To define dash inside a configuration parameter, use triple underscore, such as YARN_CONF_yarn_log___aggregation___enable=true (yarn-site.xml):

```
<property><name>yarn.log-aggregation-enable</name><value>true</value></property>
```

The available configurations are:

- /etc/hadoop/core-site.xml CORE_CONF
- /etc/hadoop/hdfs-site.xml HDFS_CONF
- /etc/hadoop/yarn-site.xml YARN_CONF
- /etc/hadoop/httpfs-site.xml HTTPFS_CONF
- /etc/hadoop/kms-site.xml KMS_CONF
- /etc/hadoop/mapred-site.xml MAPRED_CONF

If you need to extend some other configuration file, refer to base/entrypoint.sh bash script.

Note: This has been adapted from Big Data Europe's docker image