09: Docker Tutorial: Getting started with Hadoop Big Data on Cloudera quickstart

Posted on May 24, 2019

If you are not familiar with Docker get some hands-on experience at a series of step by step Docker tutorials with Java & Springboot examples.

Step 1: **https://hub.docker.com/** is a docker repository from where you can pull & push images. You can search for the images. You can see some of the below steps at: https://hub.docker.com/r/cloudera/quickstart.

300+ Java Interview FAQs

300+ Java FAQs 🔥

16+ Java Key Areas Q&As

150+ Java Architect FAQs

80+ Java Code Quality Q&As

150+ Java Coding Q&As

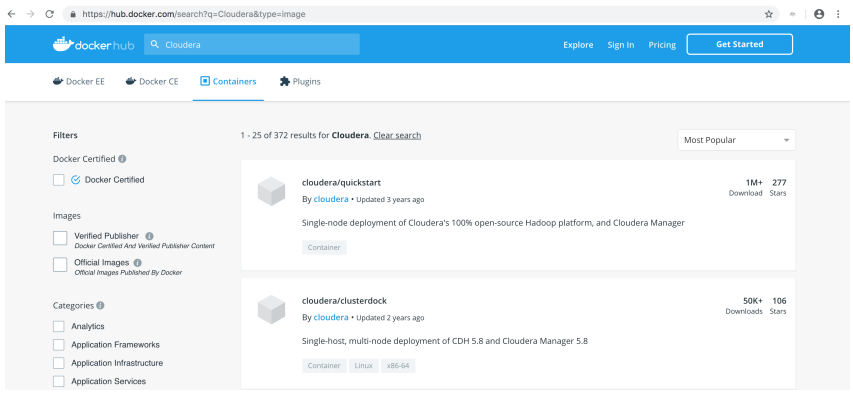
300+ Big Data Interview FAQs

300+ Big Data FAQs 🔥

Tutorials - Big Data

TUT - ⏯ Starting Big Data

TUT - Starting Spark & Scala



DockerHub – Cloudera Quickstart

```
~/projects/docker-hadoop]$ docker pull cloudera/qu
```

It will take a few minutes to download.

Step 2: Check the downloaded image with

```
~/projects/docker-hadoop]$ docker images
REPOSITORY              TAG                 IMAGE ID
cloudera/quickstart     latest              4239cd29
```

Step 3: Create the container from the image “cloudera/quickstart”.

```
~/projects/docker-hadoop]$ docker run --hostname=qu
```

Step 4: List the running containers and get the container id.

```
~/projects/docker-hadoop]$ docker ps
CONTAINER ID        IMAGE                    COMMAND
c6cbc91a69ad        cloudera/quickstart      "/usr/bi
```

TUT - Starting with Python
TUT - Kafka
TUT - Pig
TUT - Apache Storm
TUT - Spark Scala on Zeppelin
TUT - Cloudera
TUT - Cloudera on Docker
TUT - File Formats
TUT - Spark on Docker
TUT - Flume
TUT - Hadoop (HDFS)
TUT - HBase (NoSQL)
TUT - Hive (SQL)
TUT - Hadoop & Spark
TUT - MapReduce
TUT - Spark and Scala
TUT - Spark & Java
TUT - PySpark on Databricks
TUT - Zookeeper

You can see that 0.0.0.0:8888->8888/tcp, which means “0.0.0.0:8888” is the host ip address & port. You can also inspect the container by opening a new terminal and then type:

```
~/projects/docker-hadoop]$ docker inspect 7d636c3a
```

Where you can see:

```
.....  "NetworkSettings": {
            "Bridge": "",
            "SandboxID": "9f2fd6f8dc1b6a44b3da938
            "HairpinMode": false,
            "LinkLocalIPv6Address": "",
            "LinkLocalIPv6PrefixLen": 0,
            "Ports": {
                "7180/tcp": [
                    {
                        "HostIp": "0.0.0.0",
                        "HostPort": "7180"
                    }
                ],
                "80/tcp": [
                    {
                        "HostIp": "0.0.0.0",
                        "HostPort": "80"
                    }
                ],
                "8888/tcp": [
                    {
                        "HostIp": "0.0.0.0",
                        "HostPort": "8888"
                    }
                ]
            },
```

HUE

HUE stands for Hadoop User Experience where you can browse the files in HDFS and run SQL like queries against Hive & Impala tables.

Step 5: Open a browser and type "http://0.0.0.0:8888" to open HUE GUI.

username: cloudera
password: cloudera

HDFS commands

You can run a number HDFS commands on a command-line as shown below. These commands are Unix like:

Step 6: On the original terminal you can practice the Hadoop commands like

```
[root@quickstart /]# hdfs dfs -ls /user
Found 9 items
drwxr-xr-x   - cloudera cloudera          0 201
drwxr-xr-x   - hdfs     supergroup        0 201
drwxr-xr-x   - mapred   hadoop            0 201
drwxrwxrwx   - hive     supergroup        0 201
drwxrwxrwx   - hue      supergroup        0 201
drwxrwxrwx   - jenkins  supergroup        0 201
drwxrwxrwx   - oozie    supergroup        0 201
drwxrwxrwx   - root     supergroup        0 201
drwxr-xr-x   - hdfs     supergroup        0 201

```

You can also see this via the HUE GUI by clicking on the "**File Browser**" at the top right. Click on the "**/user**" to see the same folders as above.

mkdir

```
[root@quickstart /]# hdfs dfs -mkdir /user/java-su
```

ls

```
[root@quickstart /]# hdfs dfs -ls /user
```

```
ound 10 items
drwxr-xr-x   - cloudera cloudera          0 201
drwxr-xr-x   - hdfs     supergroup        0 201
drwxr-xr-x   - mapred   hadoop            0 201
drwxrwxrwx   - hive     supergroup        0 201
drwxrwxrwx   - hue      supergroup        0 201
drwxr-xr-x   - root     supergroup        0 201
drwxrwxrwx   - jenkins  supergroup        0 201
drwxrwxrwx   - oozie    supergroup        0 201
drwxrwxrwx   - root     supergroup        0 201
drwxr-xr-x   - hdfs     supergroup        0 201

```

touch

Create a file.

```
[root@quickstart /]# hdfs dfs -touchz /user/java-su
```

You can see what command options are available by typing:

```
[root@quickstart /]# hdfs dfs
```

Cloudera guide

You can look at the guide & examples via: “http://localhost:80” takes you to the Quickstart guide & tutorial.

Cloudera Manager

The Cloudera manager is not started by default. It requires at around 10 GB of RAM. Cloudera manager is a web UI to manage Hadoop cluster and services like Hive, Spark, Impala, HBase, etc. You can stop, start, and restart the services. You can modify the configuration values. You can monitor the jobs and their statuses.

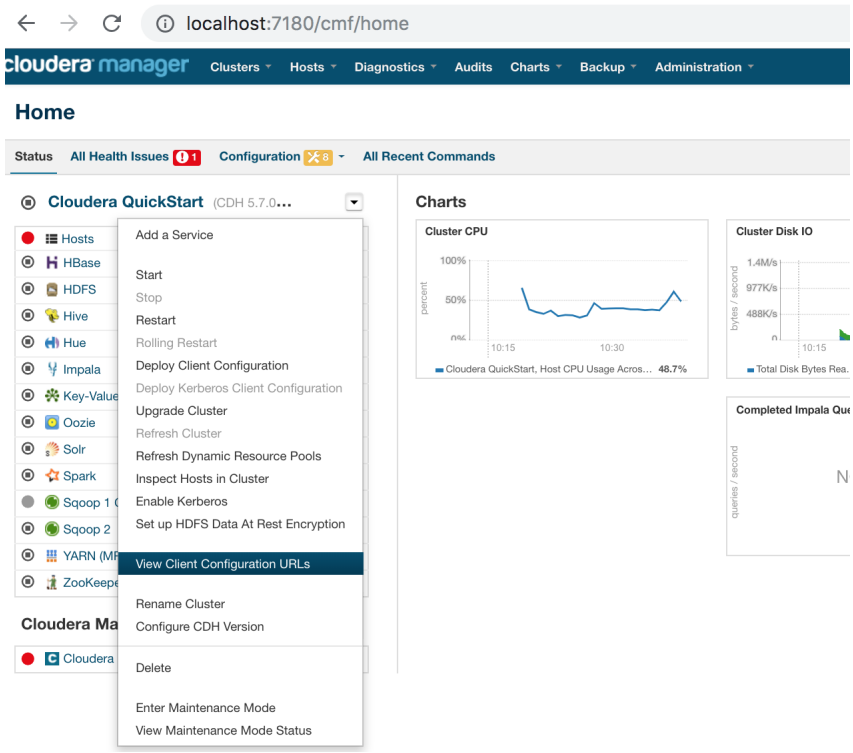
```
[root@quickstart ~]# cd /home/cloudera/
[root@quickstart cloudera]# ./cloudera-manager --e
.....
----------------------------------------------------

Success! You can now log into Cloudera Manager fr

    http://quickstart.cloudera:7180

    Username: cloudera
    Password: cloudera

[root@quickstart cloudera]#
```

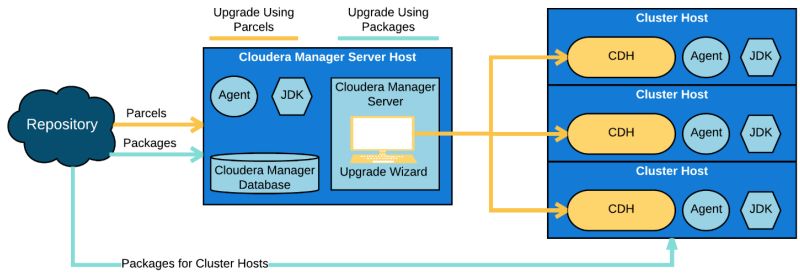


Cloudera Manager – manage cluster services & configurations

http://localhost:7180/ to access the **Cloudera manager**. You can start all the services you require or only the services that you require, but it can consume lots of resources. Services may be in bad health due to lack of resources.

The Cloudera Manager Server is the master service that manages the data model of the entire cluster in a database. The data model contains information

regarding the services, roles, and configurations assigned for each node in the cluster. You can also upgrade the services via parcels & packages.



Cloudera Manager – Server & Agents

CDH – stands for Cloudera Distribution Hadoop. CDH upgrades contain updated versions of the Hadoop software and other components. You can use Cloudera Manager to upgrade CDH for major, minor, and maintenance upgrades.

Do you want to open multiple terminal windows?

```
$ docker ps
CONTAINER ID        IMAGE                    COMMAND
bf645c6a2930        cloudera/quickstart    "/usr/bi
```

```
$ docker exec -it bf645c6a2930 /bin/bash
[root@quickstart /]#
```

How to stop the conatiner?

```
$ docker ps
CONTAINER ID        IMAGE                    COMMAND
bf645c6a2930        cloudera/quickstart    "/usr/bi
```

```
$ docker stop bf645c6a2930
```

```
2
```

How to remove all exited containers?

List all containers inclusive of stopped containers:

```
$ docker ps -a

```

Remove all containers inclusive of stopped containers:

```
$ docker rm $(docker ps -a -q)

```

How to list all the images?

```
$ docker images -a
REPOSITORY            TAG                 IMAGE ID
gdancik/cloudera      latest              49184101
cloudera/quickstart   latest              4239cd29

```

What is next?

In the next tutorials will drill into Cloudera Quickstart – Services, CLIs, config files, etc to get a good overview. This compliments ⏯ Getting started with BigData on Cloudera, which was on a Virtual Machine.

These tutorials are based on lighter **Docker containers**. 10: Docker Tutorial: BigData services & folders on Cloudera quickstart.

‹ 08: Docker Tutorial – client & server with docker-compose.yml

10: Docker Tutorial: Hadoop Big Data services & folders on Cloudera quickstart ›

Disclaimer