Cloudera Quickstart

## Create EC2 Instance

* R5a.large storage 80G
* Ubuntu 20.04
* Firewall
  + ~~All traffic 161.246.0.0/16 or My IP~~
  + All traffic 158.108.0.0/16 or My IP

## Install Docker

$ sudo apt upgrade

$ sudo apt update

$ sudo apt install docker.io

$ sudo docker --version

$ sudo docker info

$ sudo docker images

$ sudo docker ps

$ sudo docker ps -a

[A Docker Tutorial for Beginners (docker-curriculum.com)](https://docker-curriculum.com/)

<https://docs.docker.com/get-started/>

## Run hello-world

$ sudo docker run hello-world

$ sudo docker run feb5d9fea6a5

## Pulling Cloudera Quickstart

$ sudo docker pull cloudera/quickstart:latest

$ sudo docker images

## 

## Run Cloudera Quickstart

$ sudo docker run --hostname=quickstart.cloudera --privileged=true -t -i --publish-all=true-p 8888:8888 -p 7180:7180 -p 80:80 4239cd2958c6 /usr/bin/docker-quickstart

$ sudo docker run --hostname=quickstart.cloudera --privileged=true -t -i --publish-all=true -p 8888:8888 -p 7180:7180 -p 80:80 cloudera/quickstart /usr/bin/docker-quickstart

~~$ sudo docker run -m 12g --hostname=quickstart.cloudera --privileged=true -t -i --publish-all=true -p 8888:8888 -p 7180:7180 -p 80:80 cloudera/quickstart /usr/bin/docker-quickstart~~

$ sudo docker stop cloudera/quickstart

$ sudo docker rm

$ sudo docker exec -it <container\_id> bash

[Installing Cloudera Quickstart VM using Docker Hub (on Mac M1) | by Gopidi Rakesh Reddy | Medium](https://medium.com/@rakeshgopidi/installing-cloudera-quickstart-vm-through-docker-hub-on-mac-m1-879f4a3d0fd4)

[(177) Cloudera QuickStart - Installation Using Docker Container on Ubuntu - Part 1 - YouTube](https://www.youtube.com/watch?v=elXTmZbNNmE)

## Hue

root@quickstart$ jps

root@quickstart$ service --status-all

root@quickstart$ hadoop version

root@quickstart$ hdfs fsck/

Public IP : 8888

## 

## 

## Bay Area Bike Share - Hive

root@quickstart$ cd /tmp

root@quickstart$ mkdir bike

root@quickstart$ cd bike

root@quickstart$ curl -O <https://s3.amazonaws.com/babs-open-data/babs_open_data_year_1.zip>

root@quickstart$ ls

root@quickstart$ unzip babs\_open\_data\_year\_1.zip

root@quickstart$ cd 201402\_babs\_open\_data/

root@quickstart$ more README.txt

root@quickstart$ hadoop fs -ls /

root@quickstart$ hadoop fs -ls /user/

root@quickstart$ hadoop fs -ls /user/cloudera/

root@quickstart$ hadoop fs -put 201402\_trip\_data.csv /user/cloudera/

root@quickstart$ hadoop fs -ls /user/cloudera/

## In Hive Query Editor

SELECT \* FROM trip

SELECT \* FROM trip LIMIT 10

SELECT \* FROM trip ORDER BY startterminal LIMIT 10

SELECT \* FROM trip ORDER BY DESC startterminal LIMIT 10

#### Top 10 most popular start stations based on the trip data

SELECT startterminal, startstation, COUNT(1) AS count FROM trip

GROUP BY startterminal, startstation ORDER BY count DESC LIMIT 10

#### 

#### 

#### Find the total number of trips and average duration (in minutes) of those trips, grouped by hour

SELECT

hour,

COUNT(1) AS trips,

ROUND(AVG(duration) / 60) AS avg\_duration

FROM (

SELECT

CAST(SPLIT(SPLIT(t.startdate, ' ')[1], ':')[0] AS INT) AS hour, t.duration AS duration

FROM trip t

WHERE

t.startterminal = 70

AND

t.duration IS NOT NULL

) r

GROUP BY hour

ORDER BY hour ASC;

## MovieLens Dataset

root@quickstart$ cd /tmp

root@quickstart$ mkdir movielens

root@quickstart$ cd movielens

root@quickstart$ curl -O <http://files.grouplens.org/datasets/movielens/ml-100k.zip>

root@quickstart$ unzip ml-100k.zip

root@quickstart$ more ml-100k/u.user

root@quickstart$ cd ml-100k

root@quickstart$ hadoop fs -mkdir /user/cloudera/movielens

root@quickstart$ hadoop fs -put u.user /user/cloudera/movielens

root@quickstart$ hadoop fs -ls /user/cloudera/movielens

#### Hive CLI

CREATE EXTERNAL TABLE users (userid INT, age INT, gender STRING, occupation STRING, zipcode STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘|’ STORED AS TEXTFILE LOCATION ‘/user/cloudera/movielens’;

SELECT \* FROM users;

#### 

#### 

#### 

#### 

#### Stop Sevices

#! /usr/bin/env bash

/etc/init.d/zookeeper-server stop

/etc/init.d/hadoop-hdfs-datanode stop

/etc/init.d/hadoop-hdfs-journalnode stop

/etc/init.d/hadoop-hdfs-namenode stop

/etc/init.d/hadoop-hdfs-secondarynamenode stop

/etc/init.d/hadoop-httpfs stop

/etc/init.d/hadoop-mapreduce-historyserver stop

/etc/init.d/hadoop-yarn-nodemanager stop

/etc/init.d/hadoop-yarn-resourcemanager stop

/etc/init.d/hbase-master stop

/etc/init.d/hbase-rest stop

/etc/init.d/hbase-thrift stop

/etc/init.d/hive-metastore stop

/etc/init.d/hive-server2 stop

/etc/init.d/sqoop2-server stop

/etc/init.d/spark-history-server stop

/etc/init.d/hbase-regionserver stop

/etc/init.d/hue stop

/etc/init.d/impala-state-store stop

/etc/init.d/oozie stop

/etc/init.d/solr-server stop

/etc/init.d/impala-catalog stop

/etc/init.d/impala-server stop

#### Stop Container

root@quickstart$ exit

or

$ sudo docker stop <container\_id>