

## HW 3: Hadoop using Docker

### HW 3A: Setting up Hadoop inside Docker

#### Steps:

1. Install docker and run a simple hello-world example.

```
lakshmi@lakshmi-VirtualBox:~$ sudo docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS
PORTS              NAMES
lakshmi@lakshmi-VirtualBox:~$ sudo docker run hello-world
Unable to find image 'hello-world:latest' locally
latest: Pulling from library/hello-world
1b930d010525: Pull complete
Digest: sha256:2557e3c07ed1e38f26e389462d03ed943586f744621577a99efb77324b0fe535
Status: Downloaded newer image for hello-world:latest

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/

lakshmi@lakshmi-VirtualBox:~$
```

```
File Edit View Search Terminal Help
lakshmi@lakshmi-VirtualBox:~$ sudo docker images
REPOSITORY          TAG                 IMAGE ID            CREATED             SIZE
hello-world         latest             fce289e99eb9       5 weeks ago        1.84kB
lakshmi@lakshmi-VirtualBox:~$ sudo docker ps -a
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS
PORTS              NAMES
11f6c8d6d1d4       hello-world        "/hello"           3 minutes ago      Exited (0) 3 minu
tes ago
priceless_spence
lakshmi@lakshmi-VirtualBox:~$
```

2. To install hadoop-2.7.1 docker image in a docker container and check whether the hadoop docker image got downloaded correctly.

\$ docker pull sequenceiq/hadoop-docker:2.7.1

\$ docker images

```
File Edit View Search Terminal Help
timeout] [-u user] file ...
lakshmi@lakshmi-VirtualBox:~$ sudo r
sudo: r: command not found
lakshmi@lakshmi-VirtualBox:~$ clear

lakshmi@lakshmi-VirtualBox:~$ sudo docker images
REPOSITORY          TAG                 IMAGE ID            CREATED             SIZE
hello-world         latest             fce289e99eb9       5 weeks ago        1.84kB
lakshmi@lakshmi-VirtualBox:~$ sudo docker ps -a
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS              PORTS
11f6c8d6d1d4       hello-world        "/hello"           3 minutes ago      Exited (0) 3 minutes ago
riceless_spence

lakshmi@lakshmi-VirtualBox:~$ sudo docker pull sequenceiq/hadoop-docker:2.7.1
2.7.1: Pulling from sequenceiq/hadoop-docker
6253335dcf03: Pulling fs layer
a3ed95caeb02: Pulling fs layer
69623ef05416: Pulling fs layer
8d2023764774: Waiting
9c3c0ff61963: Waiting
ff0696749bf6: Waiting
72accdc282f3: Waiting
5298ddb3b339: Waiting
f252bbba6bda: Waiting
5298ddb3b339: Pull complete
26343a20fa29: Download complete
f3e272e0e801: Download complete
ad78a593ca62: Download complete
673712aa7667: Download complete
9af06cd0aa6e: Download complete
fed9c9377250: Download complete
44385c519f63: Download complete
49ca93868354: Downloading [=====>] 94.1MB/126MB
98e62c38a660: Downloading [=====>] 138.7MB/154.4MB
3679d1cf91a0: Download complete
31ae294be02b: Download complete
13605254d8c3: Download complete
a54805751dfa: Downloading [=====>] 63.52MB/211.7MB
38537e9c387f: Waiting

11f6c8d6d1d4: Pull complete
c91b10bf3a44: Pull complete
adede6edfea0: Pull complete
4afb2f219fa7: Pull complete
0335bc4000de: Pull complete
e6c5265506dc: Pull complete
3bb2b06400be: Pull complete
d9665143ac9a: Pull complete
2a1a28b12647: Pull complete
5c175609cbf3: Pull complete
e2a7d6798159: Pull complete
88d87e462c71: Pull complete
3a404fc6437e: Pull complete
5517052ef612: Pull complete
fa61c616ddd1: Pull complete
d4ab0c19cb91: Pull complete
9aa826a9ca93: Pull complete
b2ecd44f6d78: Pull complete
824658b0b14c: Pull complete
e5c31d8cbce: Pull complete
Digest: sha256:2da37e4eeea57bc99dd64987391ce9e1384c63b4fa56b7525a60849a758fb950
Status: Downloaded newer image for sequenceiq/hadoop-docker:2.7.1
lakshmi@lakshmi-VirtualBox:~$ sudo docker images
REPOSITORY          TAG                 IMAGE ID            CREATED             SIZE
hello-world         latest             fce289e99eb9       5 weeks ago        1.84kB
sequenceiq/hadoop-docker  2.7.1             42efa33d1fa3       3 years ago        1.76GB
lakshmi@lakshmi-VirtualBox:~$
```

3. Create a docker container where Hadoop 2.7.1 will run and run jps command to see if the Hadoop services are up and running

```
$ docker run -it -p 50070:50070 sequenceiq/hadoop-docker:2.7.1 /etc/bootstrap.sh -bash
# jps
# ifconfig
```

```
lakshmi@lakshmi-VirtualBox:~$ sudo docker run -it -p 50070:50070 sequenceiq/hadoop-docker:2.7.1 /etc/bootstrap.sh -bash
/
Starting sshd: [ OK ]
19/02/09 17:35:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
Starting namenodes on [160a2bc6b259]
160a2bc6b259: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-160a2bc6b259.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-160a2bc6b259.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-160a2bc6b259.out
19/02/09 17:36:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-160a2bc6b259.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-160a2bc6b259.out
bash-4.1# jps
236 DataNode
547 ResourceManager
939 Jps
396 SecondaryNameNode
637 NodeManager
121 NameNode
bash-4.1#
```

```
bash-4.1# ifconfig
eth0      Link encap:Ethernet  HWaddr 02:42:AC:11:00:02
          inet addr:172.17.0.2  Bcast:172.17.255.255  Mask:255.255.0.0
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:44 errors:0 dropped:0 overruns:0 frame:0
          TX packets:19 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:5804 (5.6 KiB)  TX bytes:1444 (1.4 KiB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:347 errors:0 dropped:0 overruns:0 frame:0
          TX packets:347 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:57876 (56.5 KiB)  TX bytes:57876 (56.5 KiB)

bash-4.1#
```

4. As seen in the previous steps since all services are running, we can check namenode ui on the browser by going to <http://localhost:50070>.

Namenode information - Mozilla Firefox

Fwd: Cloud Computing

docker for ubuntu 18.10

How to Setup a Single N

How to fix docker: Got p

Namenode information

localhost:50070/dfshealth.html#tab-overview

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

## Overview '160a2bc6b259:9000' (active)

Started:	Sat Feb 09 17:35:51 EST 2019
Version:	2.7.1, r15ecc87ccf4a0228f35af08fc56de536e6ce657a
Compiled:	2015-06-29T06:04Z by jenkins from (detached from 15ecc87)
Cluster ID:	CID-5e691286-4de5-4dde-800b-c02a7a8bf44a
Block Pool ID:	BP-1961412683-172.17.0.32-1450036414523

## Summary

Security is off.  
Safemode is off.  
35 files and directories, 31 blocks = 66 total filesystem object(s).  
Heap Memory used 80.22 MB of 170 MB Heap Memory. Max Heap Memory is 889 MB.  
Non Heap Memory used 29.62 MB of 30.94 MB Committed Non Heap Memory. Max Non Heap Memory is 130 MB.

Configured Capacity:	39.12 GB
DFS Used:	320 KB (0%)
Non DFS Used:	10.21 GB

wd: Cloud Computing

docker for ubuntu 18.10

How to Setup a Single N

How to fix docker: Got p

Namenode information

localhost:50070/dfshealth.html#tab-datanode

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

## Datanode Information

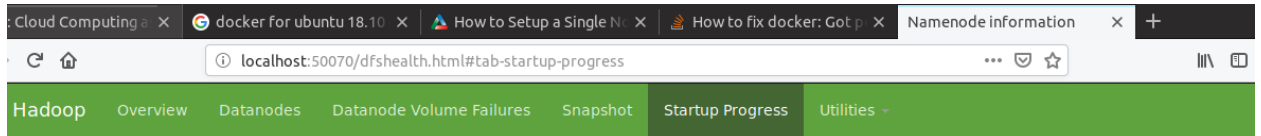
### In operation

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
160a2bc6b259:50010 (172.17.0.2:50010)	2	In Service	39.12 GB	320 KB	10.21 GB	28.91 GB	31	320 KB (0%)	0	2.7.1

### Decomissioning

Node	Last contact	Under replicated blocks	Blocks with no live replicas	Under Replicated Blocks In files under construction
------	--------------	-------------------------	------------------------------	--

Hadoop, 2015.



## Startup Progress

Elapsed Time: 35 sec, Percent Complete: 100%

Phase	Completion	Elapsed Time
<b>Loading fsimage /tmp/hadoop-root/dfs/name/current/fsimage_00000000000000000000 351 B</b>	<b>100%</b>	<b>0 sec</b>
inodes (0/0)	100%	
delegation tokens (0/0)	100%	
cache pools (0/0)	100%	
<b>Loading edits</b>	<b>100%</b>	<b>0 sec</b>
/tmp/hadoop-root/dfs/name/current/edits_00000000000000000001-00000000000000000003 1 MB (3/3)	100%	
/tmp/hadoop-root/dfs/name/current/edits_00000000000000000004-00000000000000000191 1 MB (188/188)	100%	
<b>Saving checkpoint</b>	<b>100%</b>	<b>0 sec</b>
inodes /tmp/hadoop-root/dfs/name/current/fsimage.ckpt_00000000000000000191 (0/0)	100%	
delegation tokens /tmp/hadoop-root/dfs/name/current/fsimage.ckpt_00000000000000000191 (0/0)	100%	
cache pools /tmp/hadoop-root/dfs/name/current/fsimage.ckpt_00000000000000000191 (0/0)	100%	
<b>Safe mode</b>	<b>100%</b>	<b>34 sec</b>
awaiting reported blocks (31/31)	100%	

- To run a Hadoop mapreduce in the docker container follow the below commands.

```
bash-4.1# cd $HADOOP_PREFIX
```

```
bash-4.1# bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar grep input output 'dfs[a-z.]+'
```

```
bash-4.1# cd SHADOOP_PREFIX
bash-4.1# bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-
Not a valid JAR: /usr/local/hadoop-2.7.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-
bash-4.1# bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar grep input output 'dfs[a-z.]+'
19/02/09 17:39:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
19/02/09 17:39:57 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/02/09 17:39:58 INFO input.FileInputFormat: Total input paths to process : 31
19/02/09 17:39:58 INFO mapreduce.JobSubmitter: number of splits:31
19/02/09 17:39:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1549751770411_0001
19/02/09 17:39:59 INFO impl.YarnClientImpl: Submitted application application_1549751770411_0001
19/02/09 17:39:59 INFO mapreduce.Job: The url to track the job: http://160a2bc6b259:8088/proxy/application_1549751770411_0001/
19/02/09 17:39:59 INFO mapreduce.Job: Running job: job_1549751770411_0001
19/02/09 17:40:08 INFO mapreduce.Job: Job job_1549751770411_0001 running in uber mode : false
19/02/09 17:40:08 INFO mapreduce.Job: map 0% reduce 0%
19/02/09 17:40:33 INFO mapreduce.Job: map 19% reduce 0%
19/02/09 17:40:53 INFO mapreduce.Job: map 32% reduce 0%
19/02/09 17:40:54 INFO mapreduce.Job: map 39% reduce 0%
19/02/09 17:41:13 INFO mapreduce.Job: map 39% reduce 13%
19/02/09 17:41:14 INFO mapreduce.Job: map 45% reduce 13%
19/02/09 17:41:15 INFO mapreduce.Job: map 55% reduce 13%
19/02/09 17:41:16 INFO mapreduce.Job: map 55% reduce 18%
19/02/09 17:41:28 INFO mapreduce.Job: map 58% reduce 18%
19/02/09 17:41:31 INFO mapreduce.Job: map 61% reduce 19%
19/02/09 17:41:33 INFO mapreduce.Job: map 71% reduce 19%
19/02/09 17:41:34 INFO mapreduce.Job: map 71% reduce 22%
19/02/09 17:41:37 INFO mapreduce.Job: map 71% reduce 24%
19/02/09 17:41:45 INFO mapreduce.Job: map 74% reduce 24%
19/02/09 17:41:46 INFO mapreduce.Job: map 74% reduce 25%
19/02/09 17:41:50 INFO mapreduce.Job: map 77% reduce 25%
19/02/09 17:41:52 INFO mapreduce.Job: map 81% reduce 26%
19/02/09 17:41:53 INFO mapreduce.Job: map 84% reduce 26%
19/02/09 17:41:54 INFO mapreduce.Job: map 87% reduce 26%
19/02/09 17:41:55 INFO mapreduce.Job: map 87% reduce 29%
19/02/09 17:42:04 INFO mapreduce.Job: map 90% reduce 29%
19/02/09 17:42:07 INFO mapreduce.Job: map 90% reduce 30%
```

```

lakshmi@lakshmi-VirtualBox: ~
File Edit View Search Terminal Help
19/02/09 17:41:55 INFO mapreduce.Job: map 87% reduce 29%
19/02/09 17:42:04 INFO mapreduce.Job: map 90% reduce 29%
19/02/09 17:42:07 INFO mapreduce.Job: map 90% reduce 30%
19/02/09 17:42:09 INFO mapreduce.Job: map 100% reduce 30%
19/02/09 17:42:11 INFO mapreduce.Job: map 100% reduce 100%
19/02/09 17:42:12 INFO mapreduce.Job: Job job_1549751770411_0001 completed successfully
19/02/09 17:42:12 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=345
    FILE: Number of bytes written=3722308
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=80529
    HDFS: Number of bytes written=437
    HDFS: Number of read operations=96
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=31
    Launched reduce tasks=1
    Data-local map tasks=31
    Total time spent by all maps in occupied slots (ms)=581545
    Total time spent by all reduces in occupied slots (ms)=77117
    Total time spent by all map tasks (ms)=581545
    Total time spent by all reduce tasks (ms)=77117
    Total vcore-seconds taken by all map tasks=581545
    Total vcore-seconds taken by all reduce tasks=77117
    Total megabyte-seconds taken by all map tasks=595502080
    Total megabyte-seconds taken by all reduce tasks=78967808
  Map-Reduce Framework
    Map input records=2060
    Map output records=24
    Map output bytes=590
    Map output materialized bytes=525
    Input split bytes=3812
    Combine input records=24
    Combine output records=13
    Reduce input groups=11

```

```

lakshmi@lakshmi-VirtualBox: ~
File Edit View Search Terminal Help
  Map output records=24
  Map output bytes=590
  Map output materialized bytes=525
  Input split bytes=3812
  Combine input records=24
  Combine output records=13
  Reduce input groups=11
  Reduce shuffle bytes=525
  Reduce input records=13
  Reduce output records=11
  Spilled Records=26
  Shuffled Maps =31
  Failed Shuffles=0
  Merged Map outputs=31
  GC time elapsed (ms)=3947
  CPU time spent (ms)=14300
  Physical memory (bytes) snapshot=7621881856
  Virtual memory (bytes) snapshot=21615378432
  Total committed heap usage (bytes)=6037176320
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
  File Input Format Counters
    Bytes Read=76717
  File Output Format Counters
    Bytes Written=437
9/02/09 17:42:12 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
9/02/09 17:42:12 INFO input.FileInputFormat: Total input paths to process : 1
9/02/09 17:42:12 INFO mapreduce.JobSubmitter: number of splits:1
9/02/09 17:42:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1549751770411_0002
9/02/09 17:42:12 INFO impl.YarnClientImpl: Submitted application application_1549751770411_0002
9/02/09 17:42:12 INFO mapreduce.Job: The url to track the job: http://160a2bc6b259:8088/proxy/application_1549751770411_0002/
9/02/09 17:42:12 INFO mapreduce.Job: Running job: job_1549751770411_0002
9/02/09 17:42:23 INFO mapreduce.Job: Job job_1549751770411_0002 running in uber mode : false
9/02/09 17:42:23 INFO mapreduce.Job: map 0% reduce 0%

```

File Edit View Search Terminal Help

```
19/02/09 17:42:12 INFO mapreduce.Job: Running job: job_1549751770411_0002
19/02/09 17:42:23 INFO mapreduce.Job: Job job_1549751770411_0002 running in uber mode : false
19/02/09 17:42:23 INFO mapreduce.Job: map 0% reduce 0%
19/02/09 17:42:30 INFO mapreduce.Job: map 100% reduce 0%
19/02/09 17:42:38 INFO mapreduce.Job: map 100% reduce 100%
19/02/09 17:42:38 INFO mapreduce.Job: Job job_1549751770411_0002 completed successfully
19/02/09 17:42:38 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=291
    FILE: Number of bytes written=232093
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=570
    HDFS: Number of bytes written=197
    HDFS: Number of read operations=7
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=4028
    Total time spent by all reduces in occupied slots (ms)=5172
    Total time spent by all map tasks (ms)=4028
    Total time spent by all reduce tasks (ms)=5172
    Total vcore-seconds taken by all map tasks=4028
    Total vcore-seconds taken by all reduce tasks=5172
    Total megabyte-seconds taken by all map tasks=4124672
    Total megabyte-seconds taken by all reduce tasks=5296128
  Map-Reduce Framework
    Map input records=11
    Map output records=11
    Map output bytes=263
    Map output materialized bytes=291
    Input split bytes=133
    Combine input records=0
    Combine output records=0
    Reduce input groups=5
```



```

Total Megabyte-seconds taken by all reduce tasks=5296128
Map-Reduce Framework
  Map input records=11
  Map output records=11
  Map output bytes=263
  Map output materialized bytes=291
  Input split bytes=133
  Combine input records=0
  Combine output records=0
  Reduce input groups=5
  Reduce shuffle bytes=291
  Reduce input records=11
  Reduce output records=11
  Spilled Records=22
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=86
  CPU time spent (ms)=1480
  Physical memory (bytes) snapshot=409325568
  Virtual memory (bytes) snapshot=1365807104
  Total committed heap usage (bytes)=325582848
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=437
File Output Format Counters
  Bytes Written=197
bash-4.1#

```

6. To check the output run the command:

**bash-4.1# bin/hdfs dfs -cat output/\***

```

bash-4.1# bin/hdfs dfs -cat output/*
19/02/09 17:45:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
6   dfs.audit.logger
4   dfs.class
3   dfs.server.namenode.
2   dfs.period
2   dfs.audit.log.maxfilesize
2   dfs.audit.log.maxbackupindex
1   dfsmetrics.log
1   dfsadmin
1   dfs.servers
1   dfs.replication
1   dfs.file
bash-4.1#

```

7. Hadoop Framework and Mapreduce program model:

## Hadoop Framework

Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop enables resilient, distributed processing of massive

unstructured data sets across commodity computer clusters, in which each node of the cluster includes its own storage

The project includes these modules:

- Hadoop Common: The common utilities that support the other Hadoop modules.
- Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

## **MapReduce**

MapReduce is a core component of the Apache Hadoop software framework. MapReduce serves two essential functions: It parcels out work to various nodes within the cluster or map, and it organizes and reduces the results from each node into a cohesive answer to a query.

MapReduce is composed following components:

- JobTracker - the master node that manages all jobs and resources in a cluster
- TaskTrackers - agents deployed to each machine in the cluster to run the map and reduce tasks
- JobHistoryServer - a component that tracks completed jobs and is typically deployed as a separate function or with JobTracker.