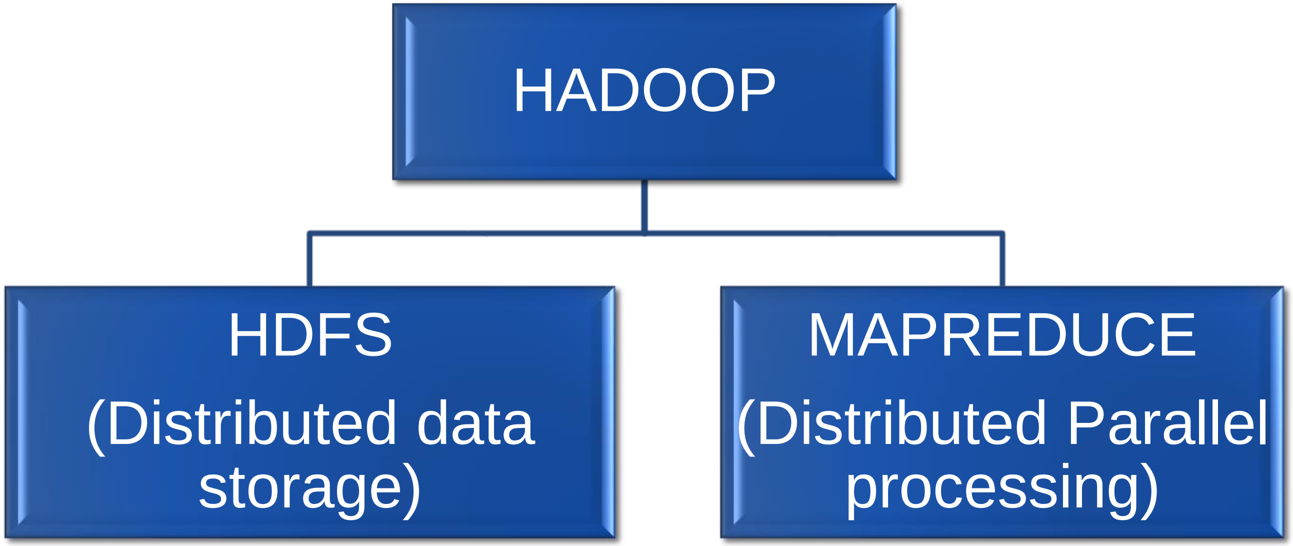
CLOUDERA

A Quick Overview

by Suchitra Jayaprakash
suchitra@cmi.ac.in

Apache Hadoop


HADOOP

HDFS

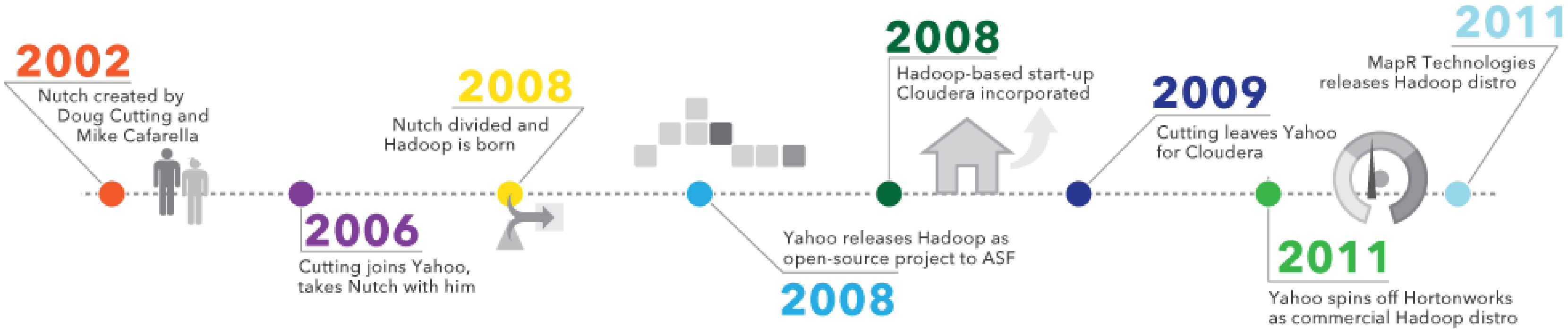
(Distributed data storage)

MAPREDUCE

(Distributed Parallel processing)

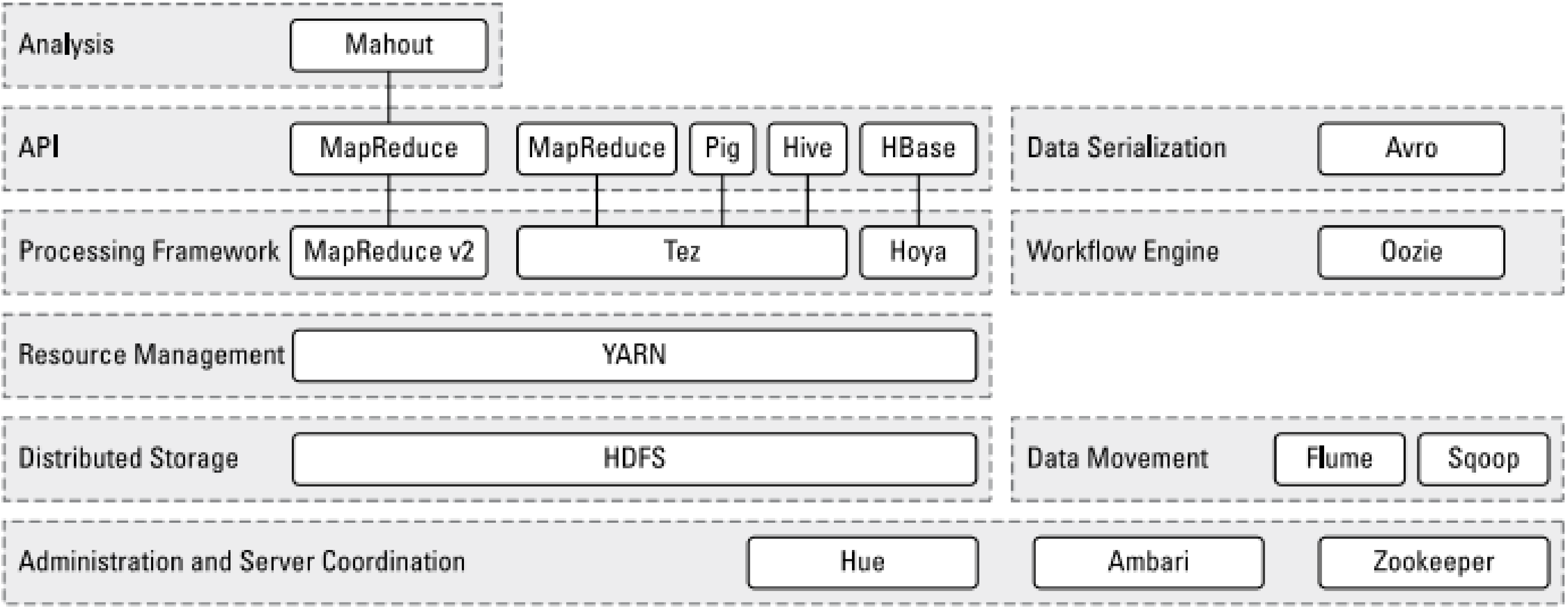
Apache Hadoop

- Hadoop is open source software framework used for processing data on distributed commodity computing environment.
- It is a java based software managed by Apache Software Foundation.
- Hadoop is designed to scale up from single server to thousands of machines.
- Doug Cutting & Mike Cafarella are co-founders of Hadoop. It is based on google's white paper on Google File System & mapreduce.


2002
Nutch created by Doug Cutting and Mike Cafarella
2006
Cutting joins Yahoo, takes Nutch with him
2008
Nutch divided and Hadoop is born
2008
Yahoo releases Hadoop as open-source project to ASF
2008
Hadoop-based start-up Cloudera incorporated
2009
Cutting leaves Yahoo for Cloudera
2011
Yahoo spins off Hortonworks as commercial Hadoop distro
2011
MapR Technologies releases Hadoop distro

(source: https://www.sas.com/en_in/insights/big-data/hadoop.html)

Hadoop Ecosystem



(source: Hadoop for Dummies)

HADOOP DISTRIBUTION

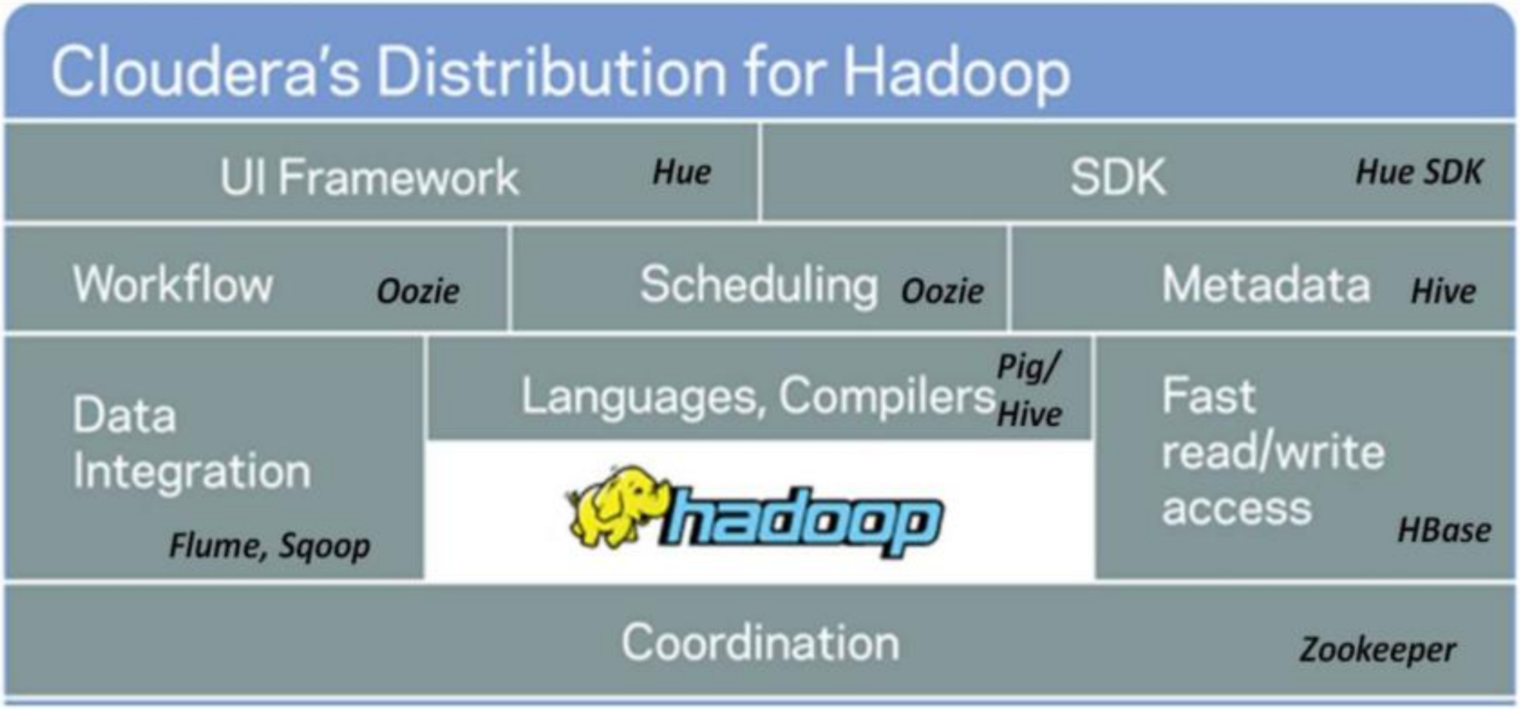
- Customisation for industry needs resulted in emergence of commercial distribution.
- Base version Apache Hadoop + features (UI , Security , Monitoring , logging, Support).
- Top Vendors offering Big Data Hadoop solution :
  - Cloudera
  - Hortonworks
  - MapR
  - Amazon Web Services Elastic MapReduce Hadoop Distribution
  - Microsoft Azure's HDInsight -Cloud based Hadoop Distrbution
  - IBM InfoSphere Insights

CLOUDERA

- Founded in 2008 by three engineers from Google, Yahoo! and Facebook (Christophe Bisciglia, Amr Awadallah and Jeff Hammerbacher).
- Major code contributor of Apache Hadoop ecosystem.
- First company to develop and distribute Apache Hadoop based software in March 2009.
- Additional feature includes user interface, security, interface for third party application integration.
- Offers customer support for installing , configuring , optimising Cloudera distribution through its enterprise subscription service.
- Provides a proprietary Cloudera Manager for easy installation , monitoring & trouble shooting.
- In 2016, Cloudera was ranked #5 on the Forbes Cloud 100 list

(source: Cloudera wiki)

CLOUDERA DISTRIBUTION



Cloudera's Distribution for Hadoop

UI Framework *Hue*

SDK *Hue SDK*

Workflow *Oozie*

Scheduling *Oozie*

Metadata *Hive*

Data Integration *Flume, Sqoop*

Languages, Compilers *Pig/ Hive*

Fast read/write access *HBase*

Coordination *Zookeeper*

An illustration of Cloudera's open-source Hadoop distribution (source: cloudera website).

CLOUDERA QUICKSTART

- Cloudera QuickStart VM is a sandbox environment of CDH.

- It gives a hands-on experience with CDH for demo and self-learning purposes.

- CDH deployed via Docker containers or VMs, are not intended for production use. Latest version is QuickStarts for CDH 5.13.

- System Requirement: Cloudera's 64-bit VMs require a 64-bit host OS and a virtualization product that can support a 64-bit guest.

- The amount of RAM required by the VM (separate from system RAM) varies by the run-time option you choose:

CDH and Cloudera Manager Version	RAM Required by VM
CDH 5 (default)	4+ GiB*
Cloudera Express	8+ GiB*
Cloudera Enterprise (trial)	12+ GiB*

*Minimum recommended memory.

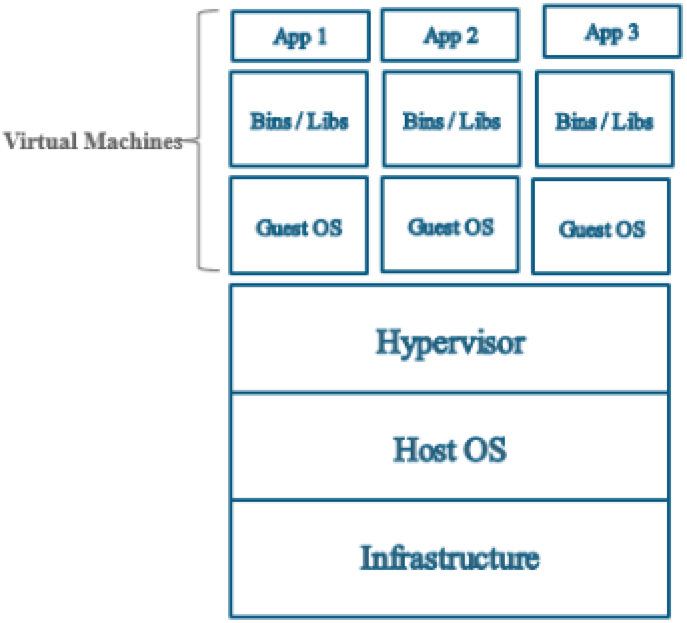
(source: Cloudera website)

DEPLOYMENT MODES - DOCKER

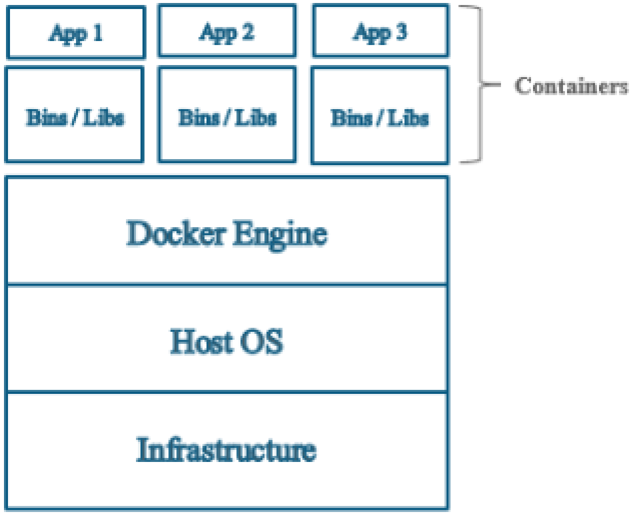
- Docker is an open source tool that uses containers to create, deploy, and manage distributed applications.

- Developers use containers to create packages for applications that include all libraries that are needed to run the application in isolation.

DEPLOYMENT MODES : VM vs DOCKER



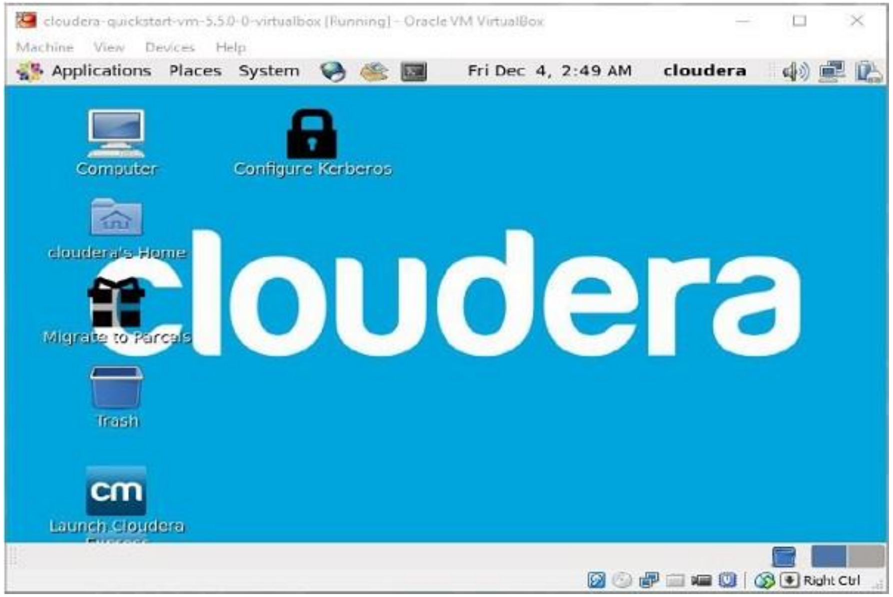
Virtual Machine / Virtual Box

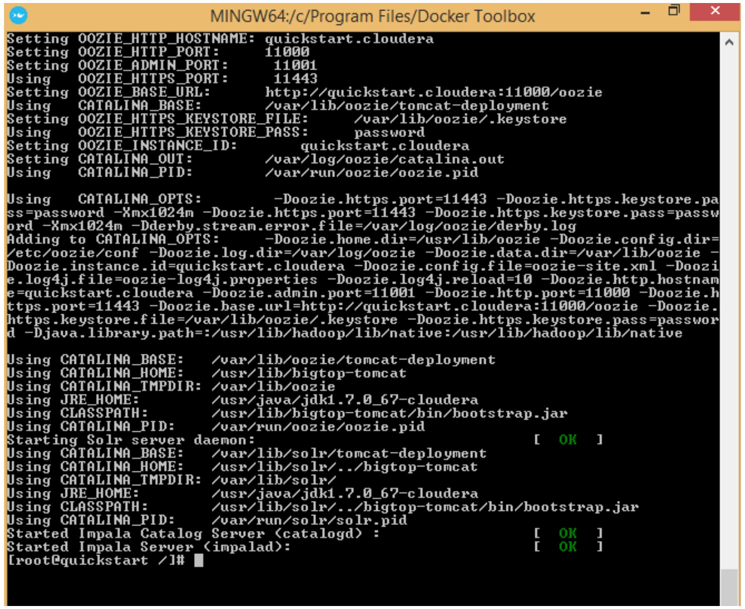


Docker Container

- Virtual machine has its guest operating system above the host operating system.
- Docker containers share the host operating system.

Virtual Machine vs Docker Container





```
Setting OOZIE_HTTP_HOSTNAME: quickstart.cloudera
Setting OOZIE_HTTP_PORT:     11000
Setting OOZIE_ADMIN_PORT:     11001
Using   OOZIE_HTTPS_PORT:     11443
Setting OOZIE_BASE_URL:      http://quickstart.cloudera:11000/oozie
Using   CATALINA_BASE:       /var/lib/oozie/tomcat-deployment
Setting OOZIE_HTTPS_KEYSTORE_FILE:     /var/lib/oozie/.keystore
Using   OOZIE_HTTPS_KEYSTORE_PASS:     password
Setting OOZIE_INSTANCE_ID:       quickstart.cloudera
Setting CATALINA_OUT:        /var/log/oozie/catalina.out
Using   CATALINA_PID:        /var/run/oozie/oozie.pid

Using   CATALINA_OPTS:        -Doozie.https.port=11443 -Doozie.https.keystore.pa
ss=password -Xmx1024m -Doozie.https.port=11443 -Doozie.https.keystore.pass=passw
ord -Xmx1024m -Dderby.stream.error.file=/var/log/oozie/derby.log
Adding to CATALINA_OPTS:     -Doozie.home.dir=/usr/lib/oozie -Doozie.config.dir=
/etc/oozie/conf -Doozie.log.dir=/var/log/oozie -Doozie.data.dir=/var/lib/oozie -
Doozie.instance.id=quickstart.cloudera -Doozie.config.file=oozie-site.xml -Doozi
e.log4j.file=oozie-log4j.properties -Doozie.log4j.reload=10 -Doozie.http.hostnam
e=quickstart.cloudera -Doozie.admin.port=11001 -Doozie.http.port=11000 -Doozie.h
ttps.port=11443 -Doozie.base.url=http://quickstart.cloudera:11000/oozie -Doozie.
https.keystore.file=/var/lib/oozie/.keystore -Doozie.https.keystore.pass=passwor
d -Djava.library.path=:/usr/lib/hadoop/lib/native:/usr/lib/hadoop/lib/native

Using CATALINA_BASE:   /var/lib/oozie/tomcat-deployment
Using CATALINA_HOME:   /usr/lib/bigtop-tomcat
Using CATALINA_TMPDIR: /var/lib/oozie
Using JRE_HOME:        /usr/java/jdk1.7.0_67-cloudera
Using CLASSPATH:       /usr/lib/bigtop-tomcat/bin/bootstrap.jar
Using CATALINA_PID:    /var/run/oozie/oozie.pid
Starting Solr server daemon:                                [  OK  ]
Using CATALINA_BASE:   /var/lib/solr/tomcat-deployment
Using CATALINA_HOME:   /usr/lib/solr/../bigtop-tomcat
Using CATALINA_TMPDIR: /var/lib/solr/
Using JRE_HOME:        /usr/java/jdk1.7.0_67-cloudera
Using CLASSPATH:       /usr/lib/solr/../bigtop-tomcat/bin/bootstrap.jar
Using CATALINA_PID:    /var/run/solr/solr.pid
Started Impala Catalog Server (catalogd) :                  [  OK  ]
Started Impala Server (impalad):                            [  OK  ]
[root@quickstart /]#
```

QUICKSTART : DOCKER INSTALL

- The Cloudera Docker image is a single-host deployment of the Cloudera open-source distribution.

- Single Node Hadoop Cluster has only a single machine
  - DataNode, NameNode run on the same machine

- Multi-Node Hadoop Cluster will have more than one machine
  - DataNode, NameNode run on different machines.

- Follows instructions in below link for Quickstart docker installation, https://docs.cloudera.com/documentation/enterprise/5-13-x/topics/quickstart_docker_container.html

QUICKSTART : DOCKER INSTALL

- **Installation Steps for Windows** :

1. **Install Docker :**

   - Sign up to https://docs.docker.com/

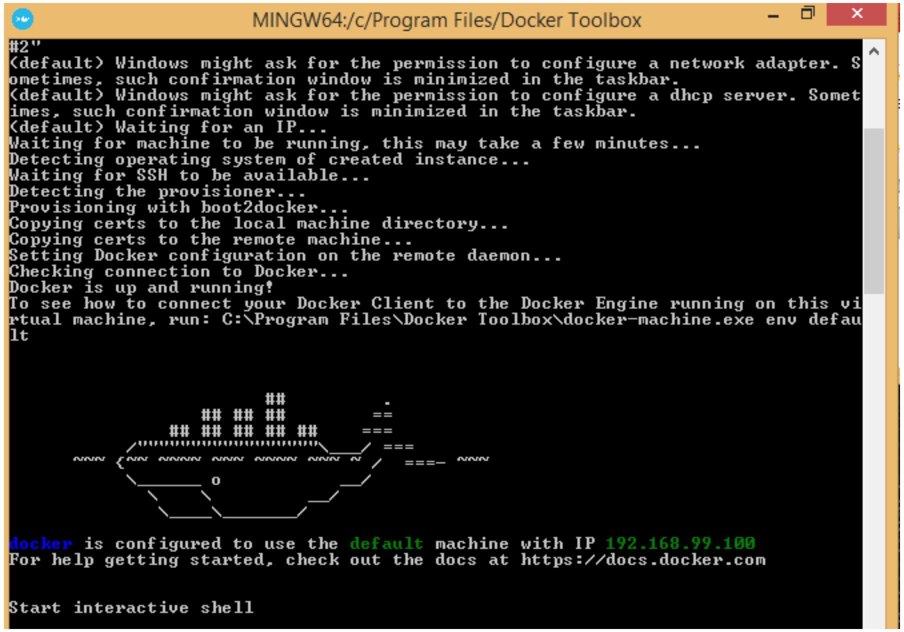
   - Follow instructions at https://docs.docker.com/docker-for-windows/install/

   - For Windows 10 64-bit Pro, Enterprise, or Education (Build 15063 or later) : Install Docker Desktop.

   - For Other Windows OS :

     Install Docker Toolbox (refer below link for instructions. https://docs.docker.com/toolbox/toolbox_install_windows/)

QUICKSTART : DOCKER INSTALL



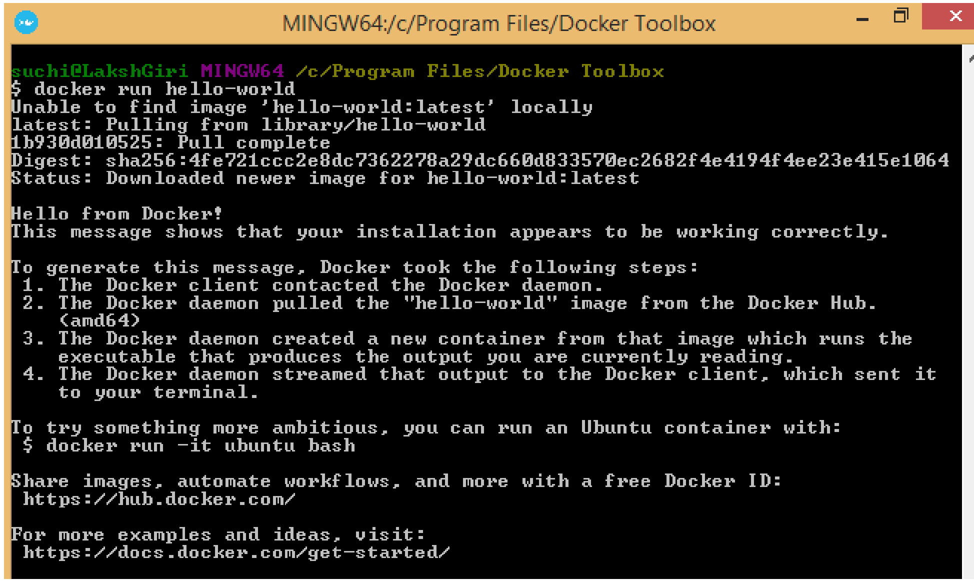
```
#2"
(default) Windows might ask for the permission to configure a network adapter. S
ometimes, such confirmation window is minimized in the taskbar.
(default) Windows might ask for the permission to configure a dhcp server. Somet
imes, such confirmation window is minimized in the taskbar.
(default) Waiting for an IP...
Waiting for machine to be running, this may take a few minutes...
Detecting operating system of created instance...
Waiting for SSH to be available...
Detecting the provisioner...
Provisioning with boot2docker...
Copying certs to the local machine directory...
Copying certs to the remote machine...
Setting Docker configuration on the remote daemon...
Checking connection to Docker...
Docker is up and running!
To see how to connect your Docker Client to the Docker Engine running on this vi
rtual machine, run: C:\Program Files\Docker Toolbox\docker-machine.exe env defau
lt



                        ##         .
                  ## ## ##        ==
               ## ## ## ## ##    ===
           /"""""""""""""""""\___/ ===
      ~~~ {~~ ~~~~ ~~~ ~~~~ ~~~ ~ /  ===- ~~~
           \______ o           __/
             \    \         __/
              \____\_______/

docker is configured to use the default machine with IP 192.168.99.100
For help getting started, check out the docs at https://docs.docker.com


Start interactive shell
```



```
suchi@LakshGiri MINGW64 /c/Program Files/Docker Toolbox
$ docker run hello-world
Unable to find image 'hello-world:latest' locally
latest: Pulling from library/hello-world
1b930d010525: Pull complete
Digest: sha256:4fe721ccc2e8dc7362278a29dc660d833570ec2682f4e4194f4ee23e415e1064
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/
```

- To check docker installation is proper , type below command in docker terminal.

  docker run hello-world

- If you get above ouput in the terminal then docker installation is fine.

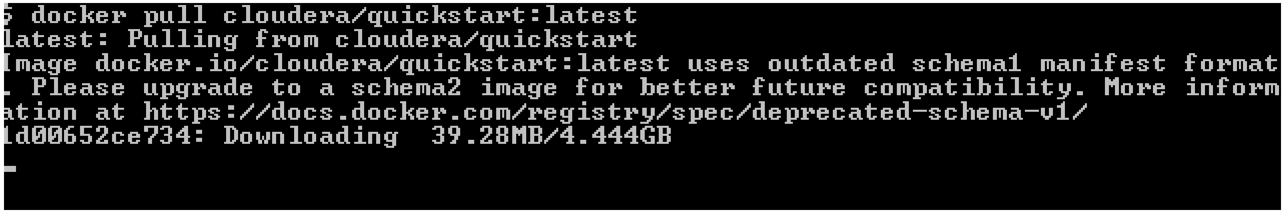
QUICKSTART : DOCKER INSTALL

2. **Install Cloudera Quickstart:**

Type following command in the docker terminal to import Cloudera Quickstart image from Docker Hub:

docker pull cloudera/quickstart:latest

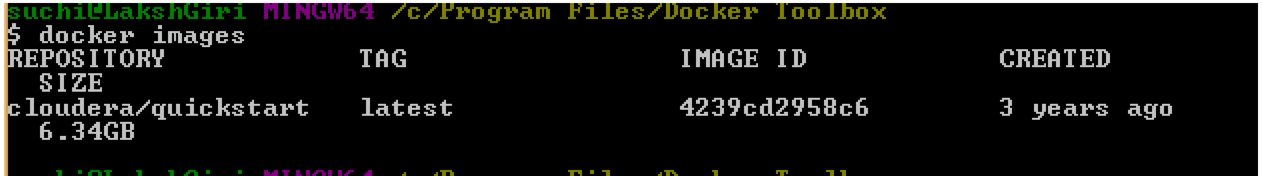
(refer link https://hub.docker.com/r/cloudera/quickstart)



```
$ docker pull cloudera/quickstart:latest
latest: Pulling from cloudera/quickstart
Image docker.io/cloudera/quickstart:latest uses outdated schema1 manifest format
. Please upgrade to a schema2 image for better future compatibility. More inform
ation at https://docs.docker.com/registry/spec/deprecated-schema-v1/
1d00652ce734: Downloading  39.28MB/4.444GB
_
```

Cloudera quickstart download will take a while to complete. After download is complete , type following in terminal :

docker images



```
suchi@LakshGiri MINGW64 /c/Program Files/Docker Toolbox
$ docker images
REPOSITORY          TAG                 IMAGE ID            CREATED
  SIZE
cloudera/quickstart   latest              4239cd2958c6        3 years ago
  6.34GB
```

QUICKSTART : DOCKER INSTALL

3. **Update Docker memory (optional)** :
   - Create a new VM with 1 CPUs and 4GB of memory (recommended).
   - Run the following command in docker terminal:
   - Remove the default vm.

     ```
     docker-machine rm default
     ```

   - Re-create the default vm.

     ```
     docker-machine create -d virtualbox --virtualbox-cpu-count=1 --virtualbox-memory=4096 --virtualbox-disk-size=50000 default
     ```

options	Description
--virtualbox-cpu-count	number of cpus
--virtualbox-memory	amount of RAM
-virtualbox-disk-size	amount of disk space

QUICKSTART : DOCKER INSTALL

4. Run Cloudera Quickstart container

- Click on “Docker Quickstart Terminal” Icon and Type below command in docker termimal to start Cloudera Quickstart

```
docker run --hostname=quickstart.cloudera --privileged=true -t -i  -p 8888:8888 -p 80:80 -p 8088:8088 -p 7180:7180 -p 50070:50070 cloudera/quickstart /usr/bin/docker-quickstart
```

Options	Required	Description
--hostname=quickstart.cloudera	Yes	Pseudo-distributed configuration assumes this as hostname.
--privileged=true	Yes	For HBase, MySQL-backed Hive metastore, Hue, Oozie, Sentry, and Cloudera Manager.
-t	Yes	Allocate a pseudoterminal. Once services are started, a Bash shell takes over. This switch starts a terminal emulator to run the services.
-i	Yes	Enable interactive terminal i.e. If you want to use the terminal, either immediately or connect to the terminal later.
--publish-all=true	No	opens up all the host ports to the docker ports
-p 8888	Yes - Recommended	Map the Hue port in the guest to port on the host.
-p [PORT]	No	Map any other ports in the guest to port on the host.
cloudera/quickstart	Yes	Name of image which run as new container
/usr/bin/docker-quickstart	Yes	Start all CDH services, and then run a Bash shell.

QUICKSTART : DOCKER INSTALL

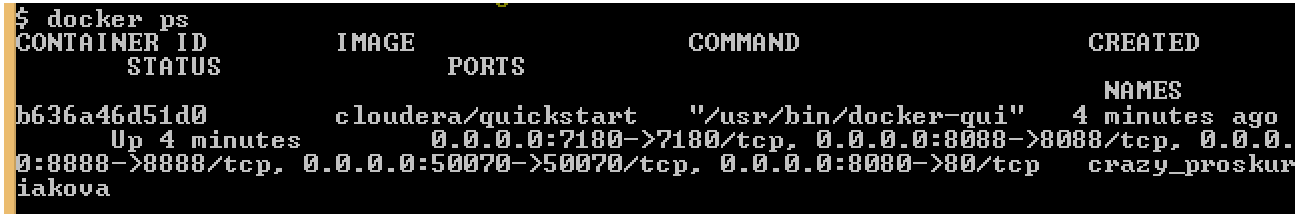
List of common ports used in Cloudera :

Port	Purpose
8888	Hue web interface
7180	Cloudera manager
80	Cloudera examples
50070	Name node web interface
8088	job tracker :- yarn

5. **Host – Guest port mapping**

- Open new docker terminal & type below command.

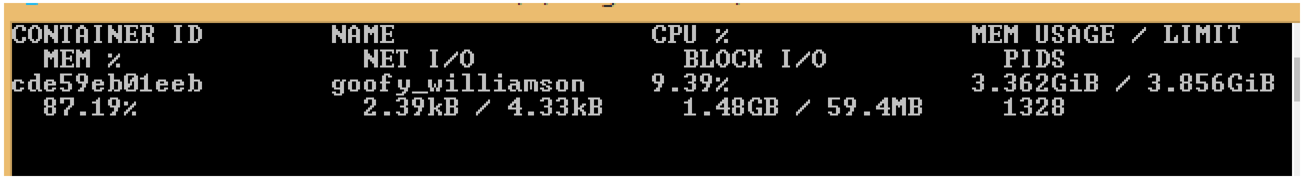
docker ps



```
$ docker ps
CONTAINER ID        IMAGE                          COMMAND                  CREATED
       STATUS              PORTS
                                                                             NAMES
b636a46d51d0        cloudera/quickstart   "/usr/bin/docker-qui"   4 minutes ago
      Up 4 minutes         0.0.0.0:7180->7180/tcp, 0.0.0.0:8088->8088/tcp, 0.0.0.
0:8888->8888/tcp, 0.0.0.0:50070->50070/tcp, 0.0.0.0:8080->80/tcp   crazy_proskur
iakova
```

- Copy the docker container ID.
- Type below to check memory allocation

docker stats [CONTAINER ID]

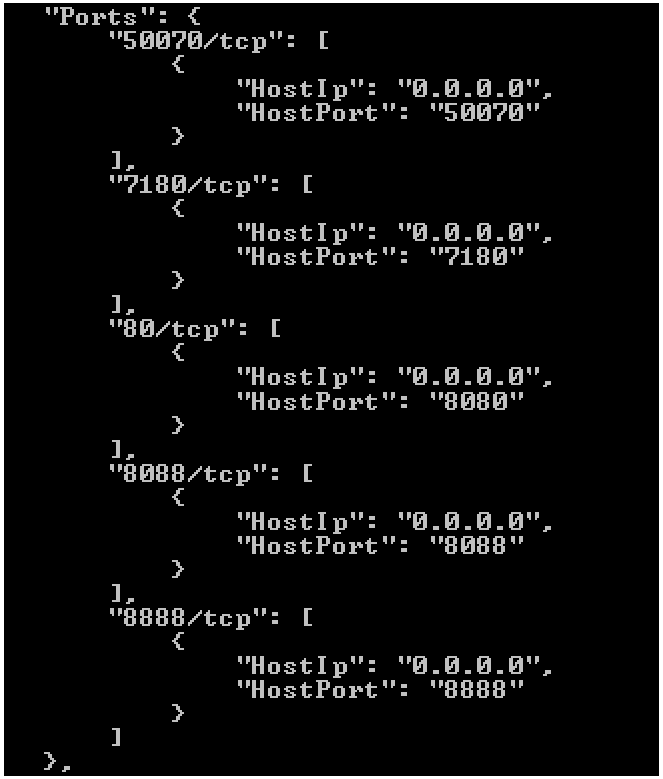


```
CONTAINER ID        NAME                CPU %               MEM USAGE / LIMIT
  MEM %               NET I/O             BLOCK I/O           PIDS
cde59eb01eeb        goofy_williamson    9.39%               3.362GiB / 3.856GiB
  87.19%              2.39kB / 4.33kB     1.48GB / 59.4MB     1328
```

QUICKSTART : DOCKER INSTALL

- Type below command and get see which Host port Hue and YARN are working.

  docker inspect [CONTAINER ID]



```
"Ports": {
    "50070/tcp": [
        {
            "HostIp": "0.0.0.0",
            "HostPort": "50070"
        }
    ],
    "7180/tcp": [
        {
            "HostIp": "0.0.0.0",
            "HostPort": "7180"
        }
    ],
    "80/tcp": [
        {
            "HostIp": "0.0.0.0",
            "HostPort": "8080"
        }
    ],
    "8088/tcp": [
        {
            "HostIp": "0.0.0.0",
            "HostPort": "8088"
        }
    ],
    "8888/tcp": [
        {
            "HostIp": "0.0.0.0",
            "HostPort": "8888"
        }
    ]
},
```

- YARN is working on port

  8088 inside the docker machine

  8088 outside on host machine

Note : in case of docker tool box, host machine is mapped to ip address 192.168.99.100. Use url

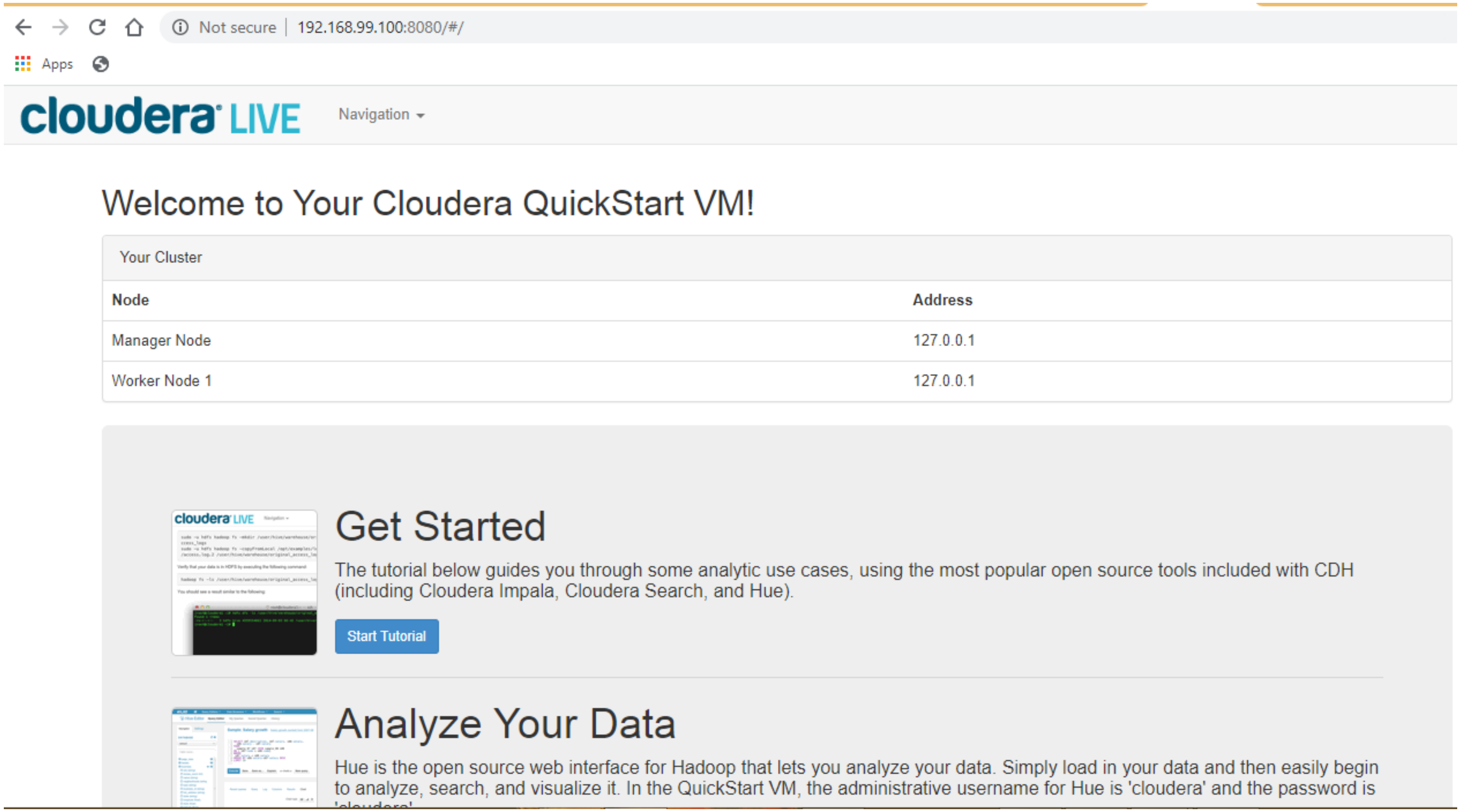
http://192.168.99.100:8080/

For other docker install use localhost
http://localhost:8080/

- **Installation Steps for Ubuntu** : https://medium.com/@dataakkadian/how-to-install-and-running-cloudera-docker-container-on-ubuntu-b7c77f147e03

QUICKSTART : DOCKER INSTALL

Tutorial page



Welcome to Your Cloudera QuickStart VM!

Your Cluster

Node	Address
Manager Node	127.0.0.1
Worker Node 1	127.0.0.1

Get Started

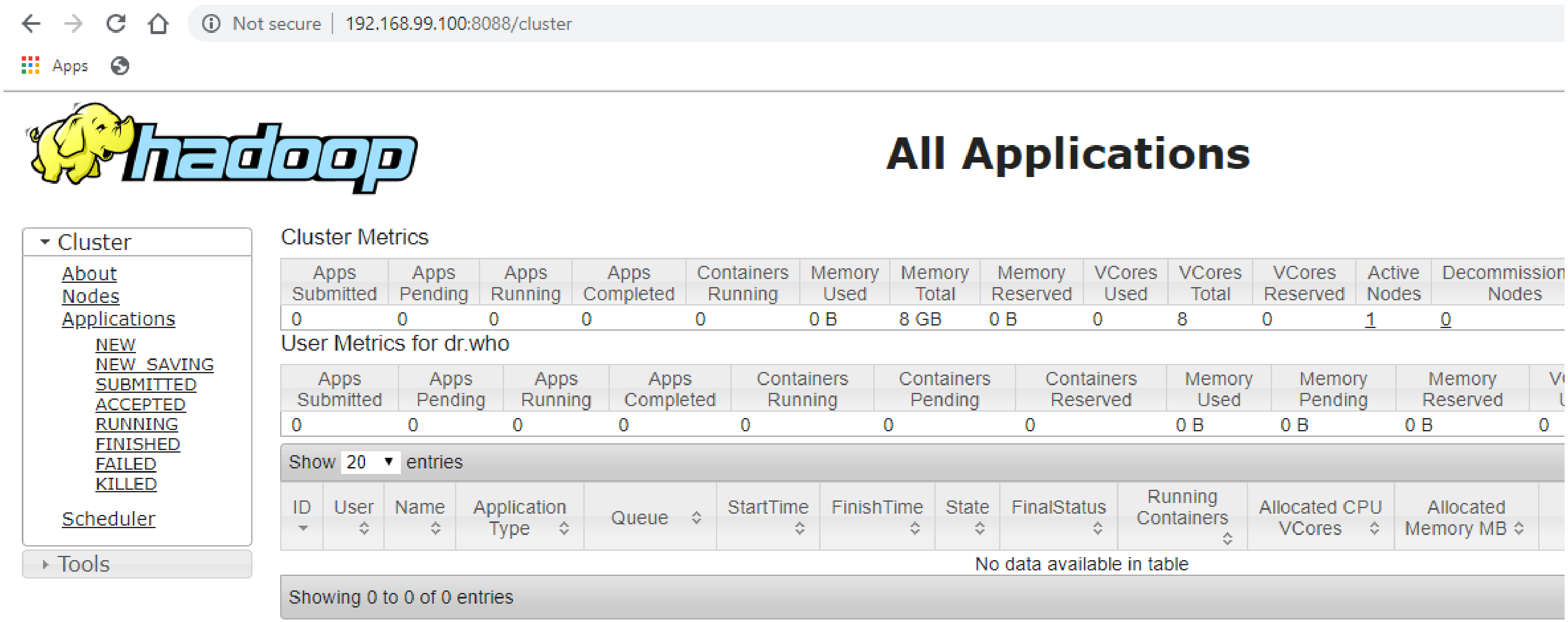
The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).

Analyze Your Data

Hue is the open source web interface for Hadoop that lets you analyze your data. Simply load in your data and then easily begin to analyze, search, and visualize it. In the QuickStart VM, the administrative username for Hue is 'cloudera' and the password is 'cloudera'.

QUICKSTART : DOCKER INSTALL

Yarn page - http://192.168.99.100:8088/



Not secure | 192.168.99.100:8088/cluster

All Applications

- Cluster
  - About
  - Nodes
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommission Nodes
0	0	0	0	0	0 B	8 GB	0 B	0	8	0	1	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	V U
0	0	0	0	0	0	0	0 B	0 B	0 B	0

Show 20 ▾ entries

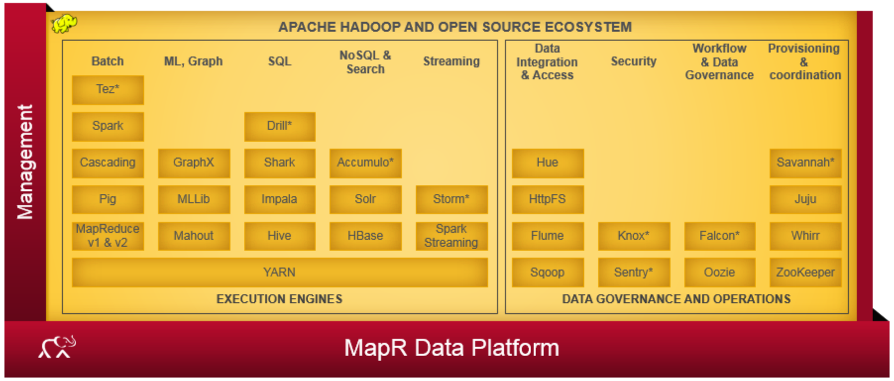
ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB
No data available in table											

Showing 0 to 0 of 0 entries

Yarn is resource management layer of Apache Hadoop ecosystem.

Other Vendors

MapR Distribution for Hadoop



APACHE HADOOP AND OPEN SOURCE ECOSYSTEM

Management

Batch: Tez*, Spark, Cascading, Pig, MapReduce v1 & v2

ML, Graph: GraphX, MLLib, Mahout

SQL: Drill*, Shark, Impala, Hive

NoSQL & Search: Accumulo*, Solr, HBase

Streaming: Storm*, Spark Streaming

YARN

EXECUTION ENGINES

Data Integration & Access: Hue, HttpFS, Flume, Sqoop

Security: Knox*, Sentry*

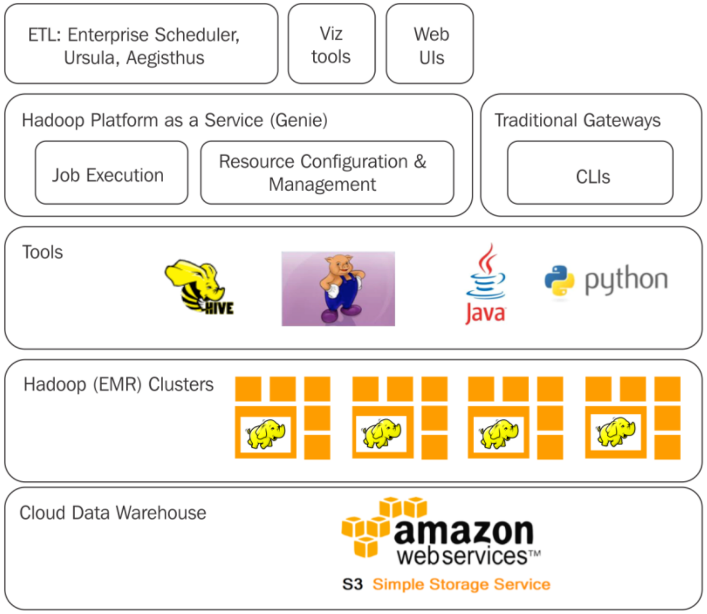
Workflow & Data Governance: Falcon*, Oozie

Provisioning & coordination: Savannah*, Juju, Whirr, ZooKeeper

DATA GOVERNANCE AND OPERATIONS

MapR Data Platform

AWS EMR



ETL: Enterprise Scheduler, Ursula, Aegisthus

Viz tools

Web UIs

Hadoop Platform as a Service (Genie)

Job Execution

Resource Configuration & Management

Traditional Gateways

CLIs

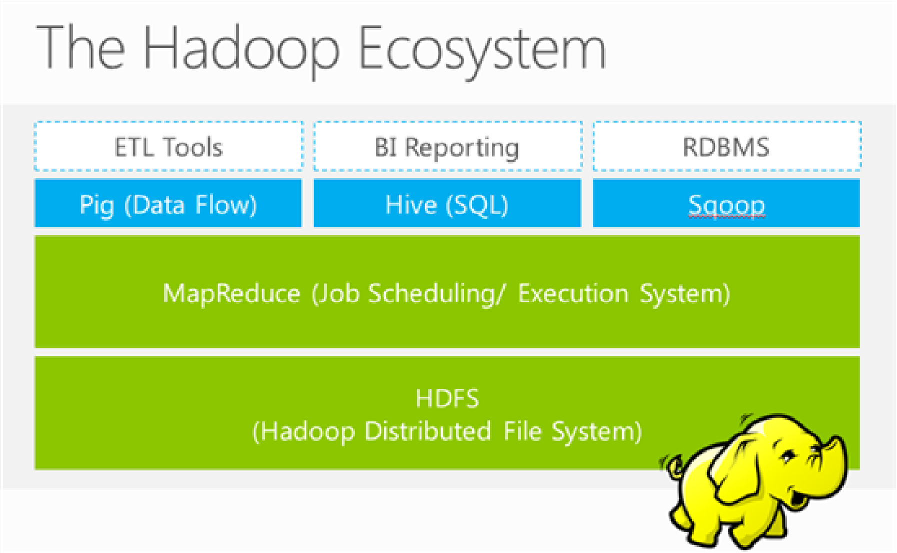
Tools

Hadoop (EMR) Clusters

Cloud Data Warehouse

S3 Simple Storage Service

Windows Azure HDInsight



The Hadoop Ecosystem

ETL Tools | BI Reporting | RDBMS

Pig (Data Flow) | Hive (SQL) | Sqoop

MapReduce (Job Scheduling/ Execution System)

HDFS (Hadoop Distributed File System)

Quiz 1

Q) Which of the following is false?

A. Cloudera products and solutions enable you to deploy and manage Apache Hadoop and related projects.

B. Cloudera QuickStart VM is a sandbox environment of CDH.

C. CDH contains all the products and frameworks belonging to the hadoop ecosystem.

D. Hadoop is open source software framework used for processing data on distributed commodity hardware.

THANK YOU