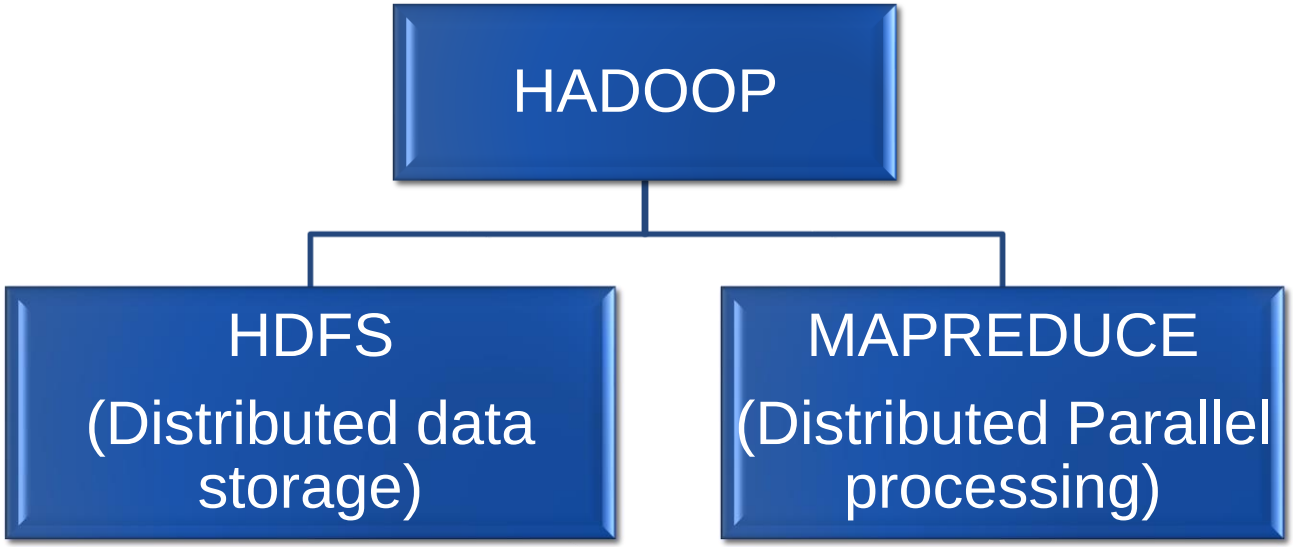
CLOUDERA

A Quick Overview

by Suchitra Jayaprakash
suchitra@cmi.ac.in

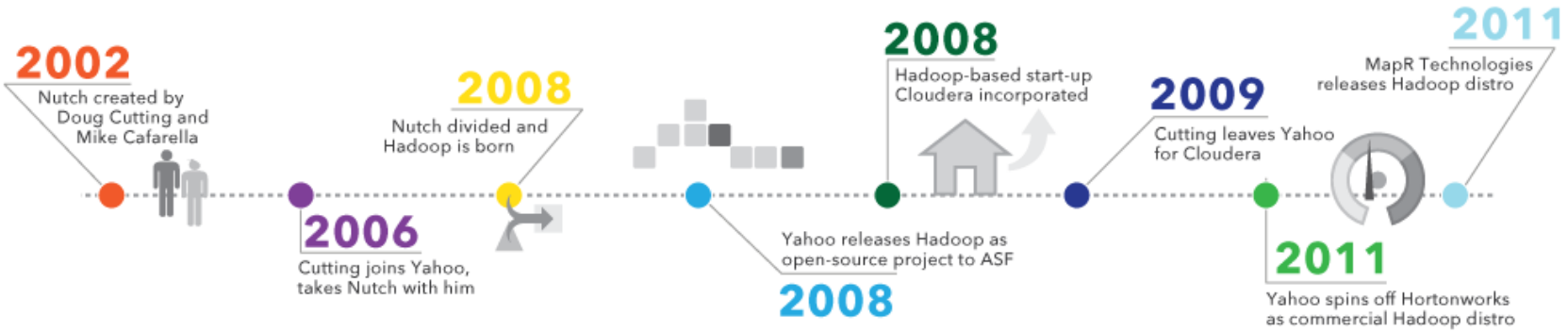
Apache Hadoop

- Hadoop is open source software framework used for processing data on distributed commodity computing environment.



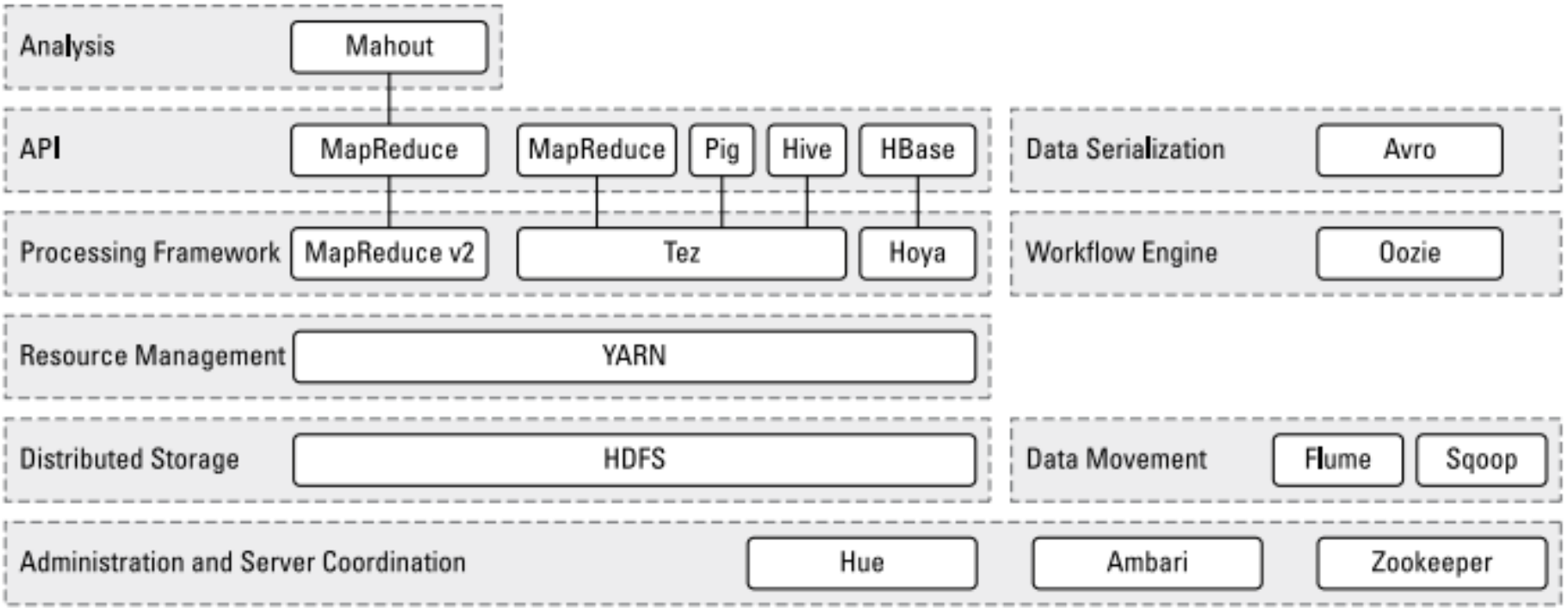
Apache Hadoop

- 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.



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

Hadoop Ecosystem



(source: Hadoop for Dummies)

HADOOP DISTRIBUTION

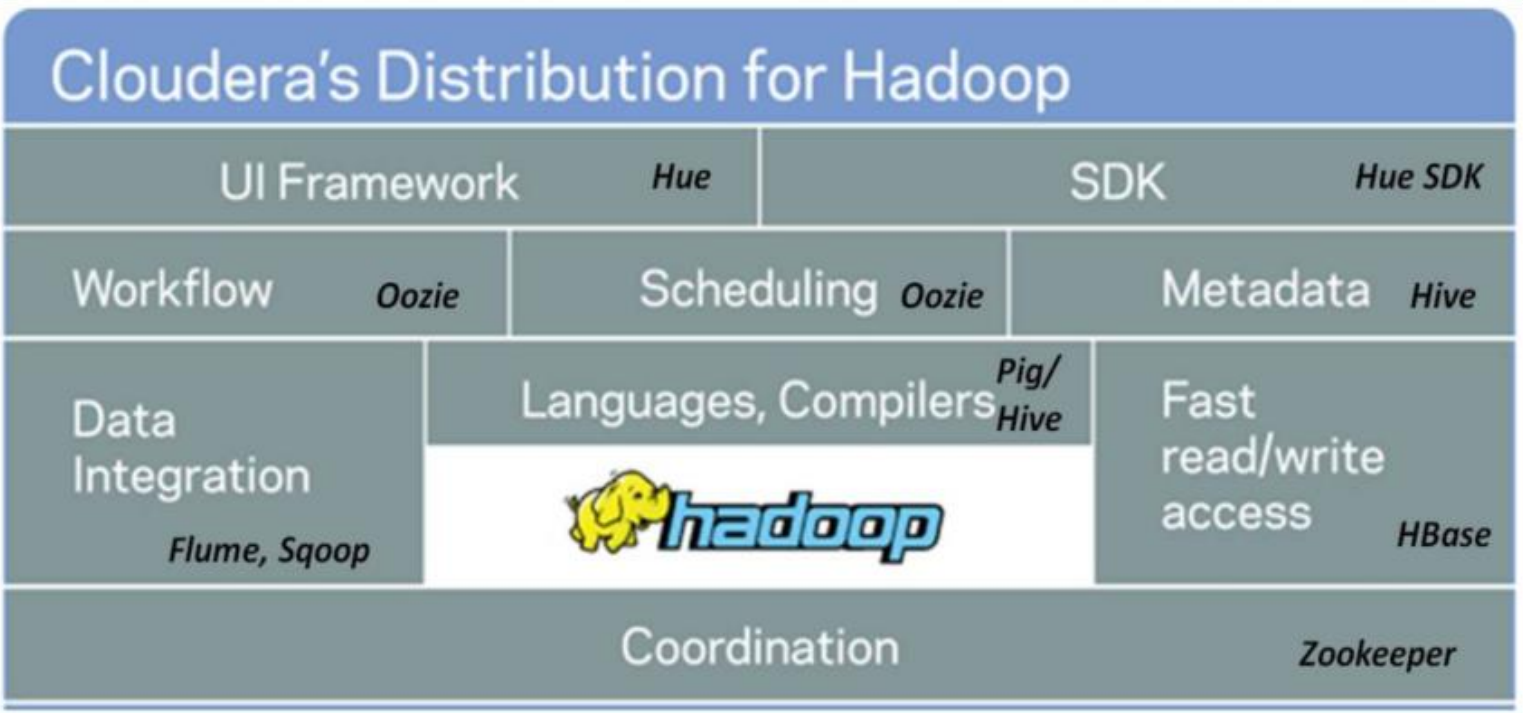
- Customisation for industry needs resulted in emergence of commercial distribution.
- Base version Apache Hadoop + features (UI , Security , Monitoring , logging, Support).
- Top Vendors offering Big Data Hadoop solution :
  - Cloudera
  - Hortonworks
  - MapR
  - Amazon Web Services Elastic MapReduce 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



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)

Quiz

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.

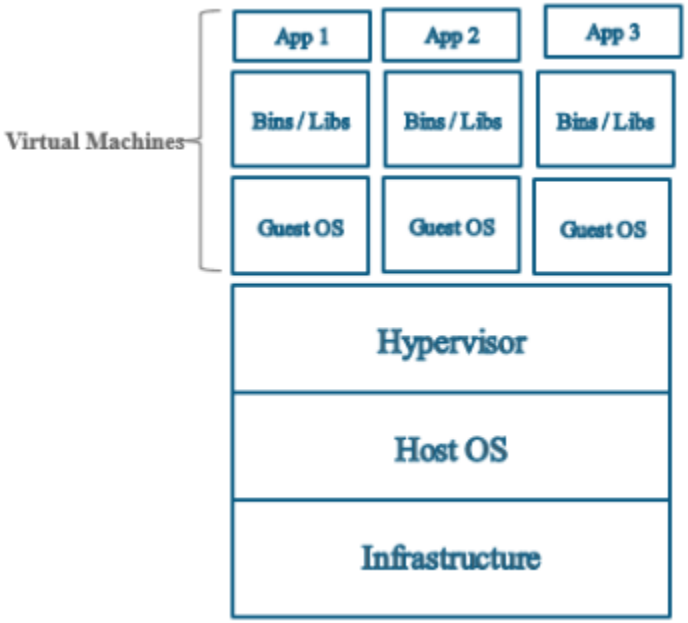
C. CDH contains all the products and frameworks belonging to the hadoop ecosystem.

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

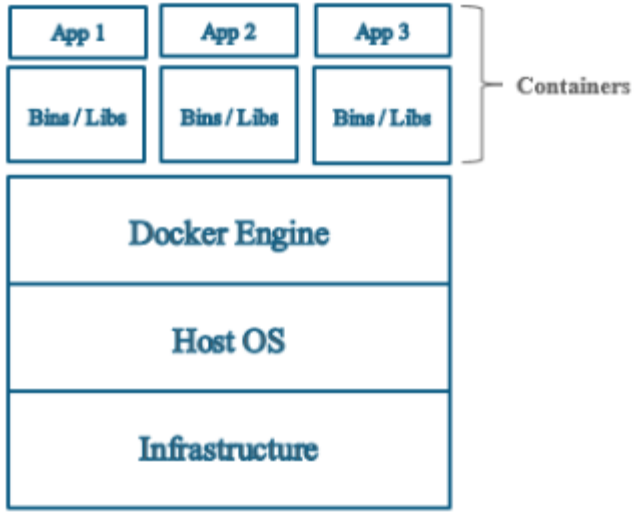
DEPLOYMENT MODES - DOCKER

- 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



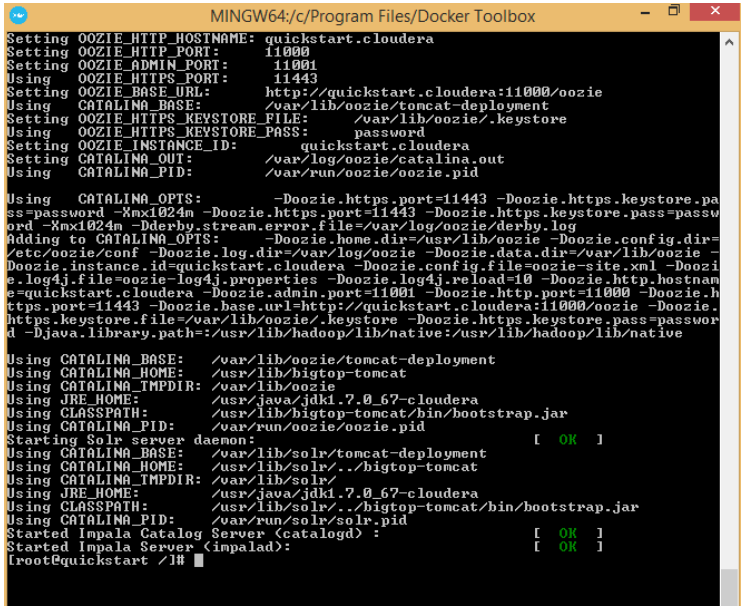
Virtual Machine / Virtual Box



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



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.

QUICKSTART : DOCKER INSTALL

- **Installation Steps for Windows** :

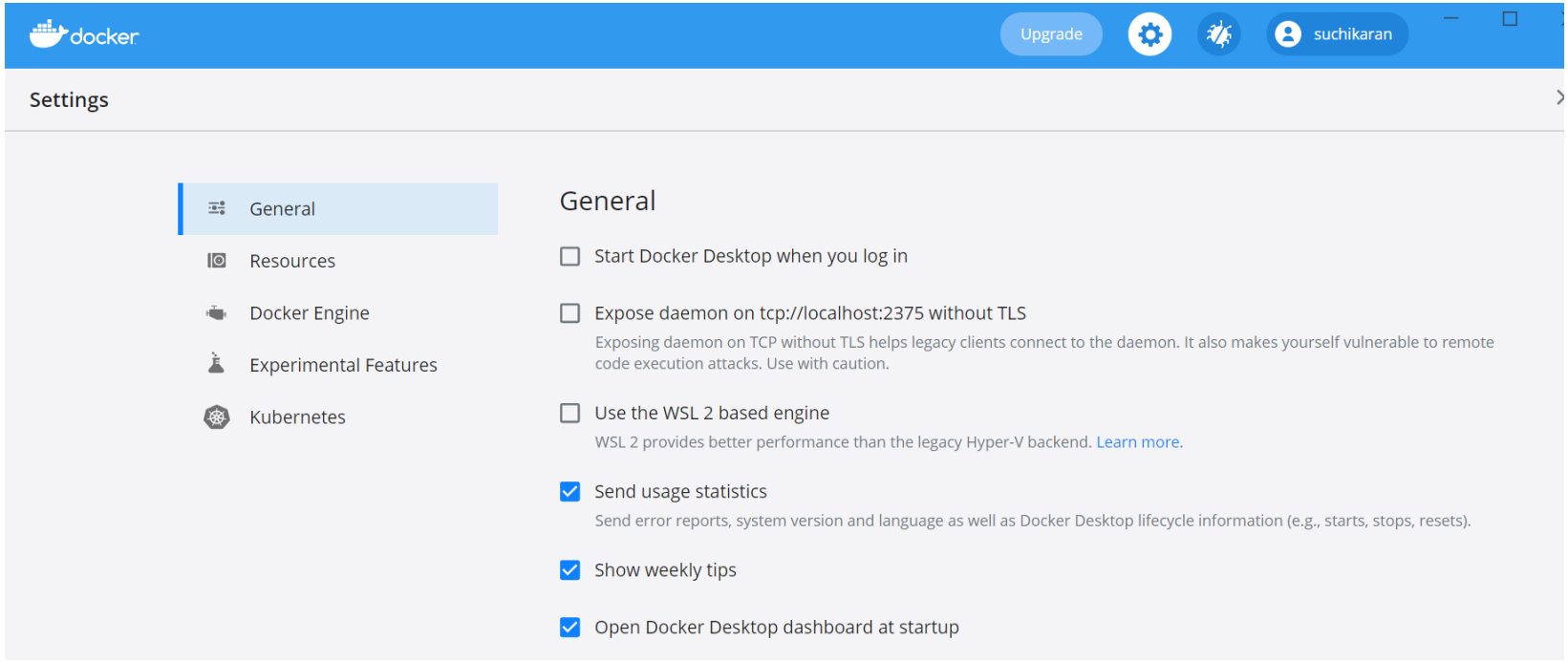
1. **Install Docker :**

    - Sign up to https://docs.docker.com/
    - Follow instructions at https://docs.docker.com/docker-for-windows/install/
    - For Windows 10 64-bit Home , Pro, Enterprise, or Education (Build 15063 or later) : Install Docker Desktop.
    - For Other Windows OS :

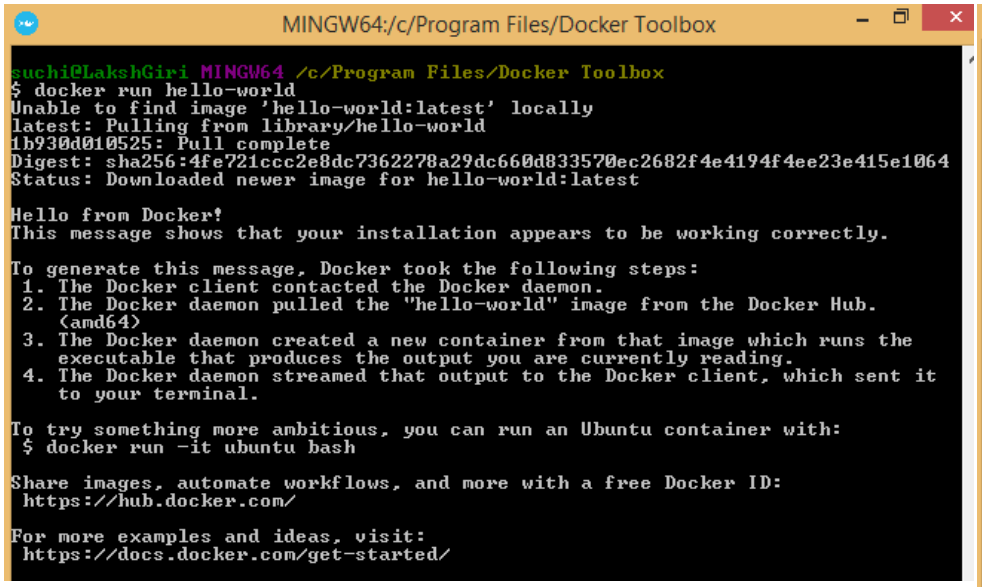
      Install Docker Toolbox (refer below link for instructions. https://docs.docker.com/toolbox/toolbox_install_windows/)

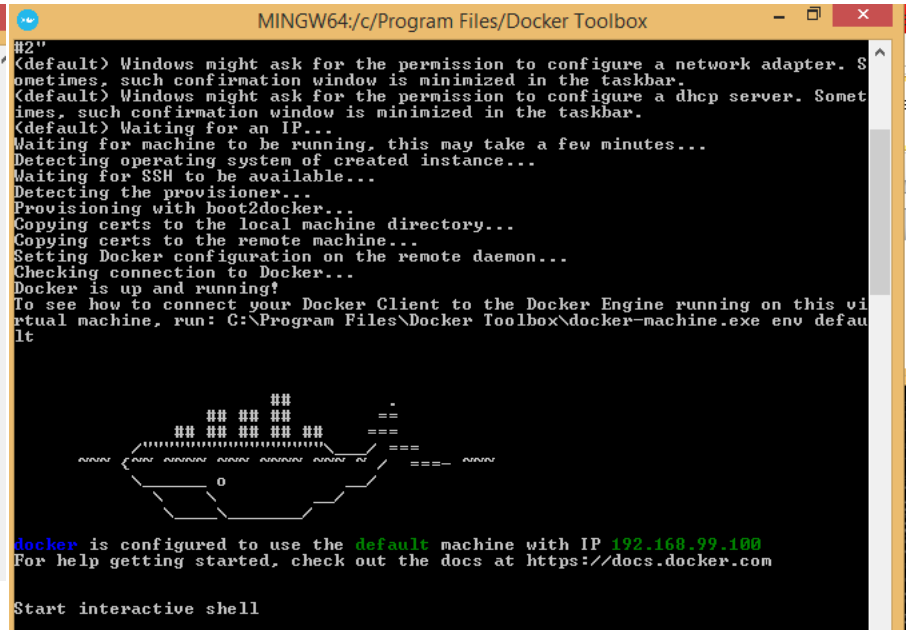
QUICKSTART : DOCKER INSTALL

- Don’t select WSL2 while installing docker. Cloudera Quick start VM is not compatible.



QUICKSTART : DOCKER INSTALL



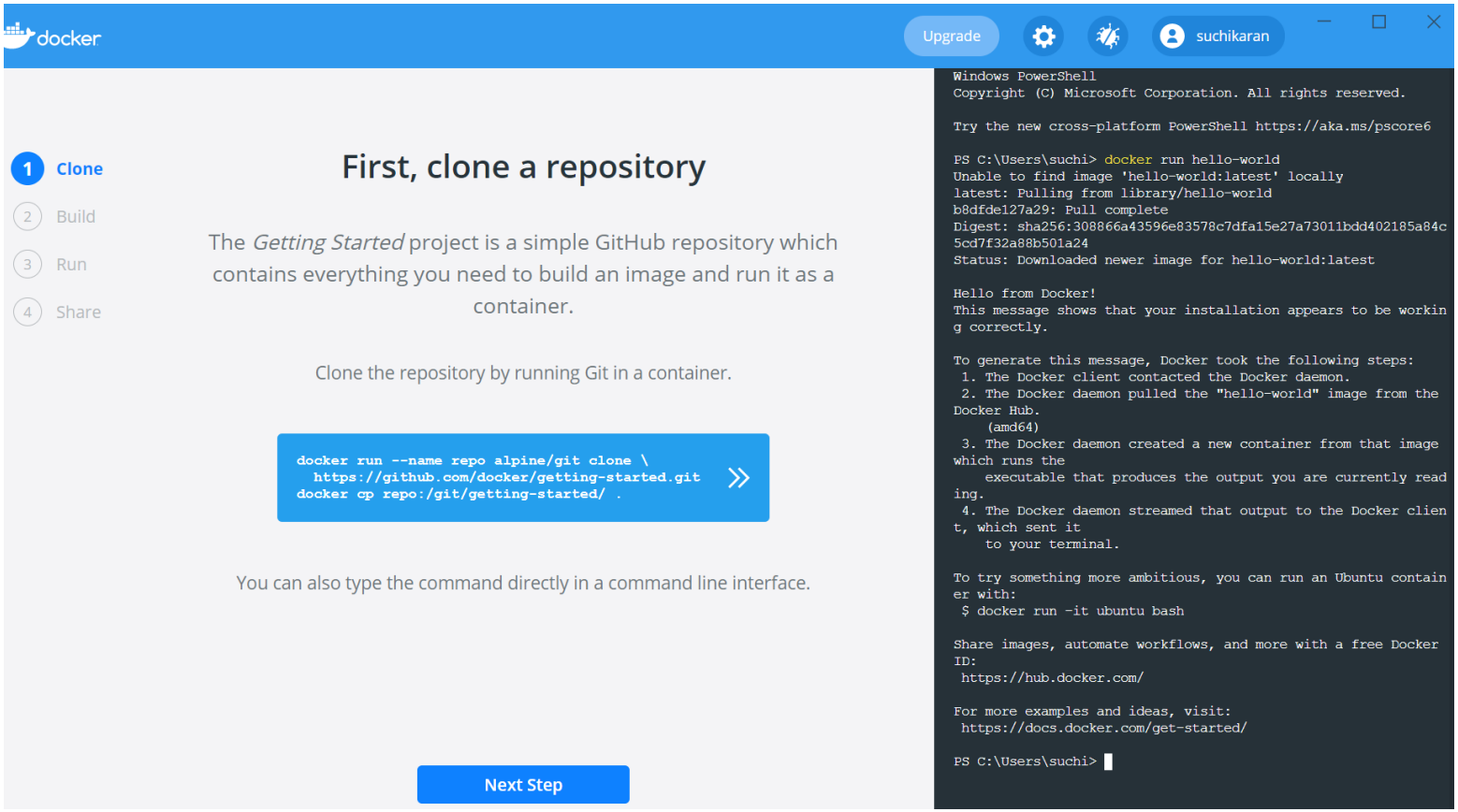


- To check docker installation is proper , type below command in docker terminal.

  docker run hello-world

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

QUICKSTART : DOCKER INSTALL

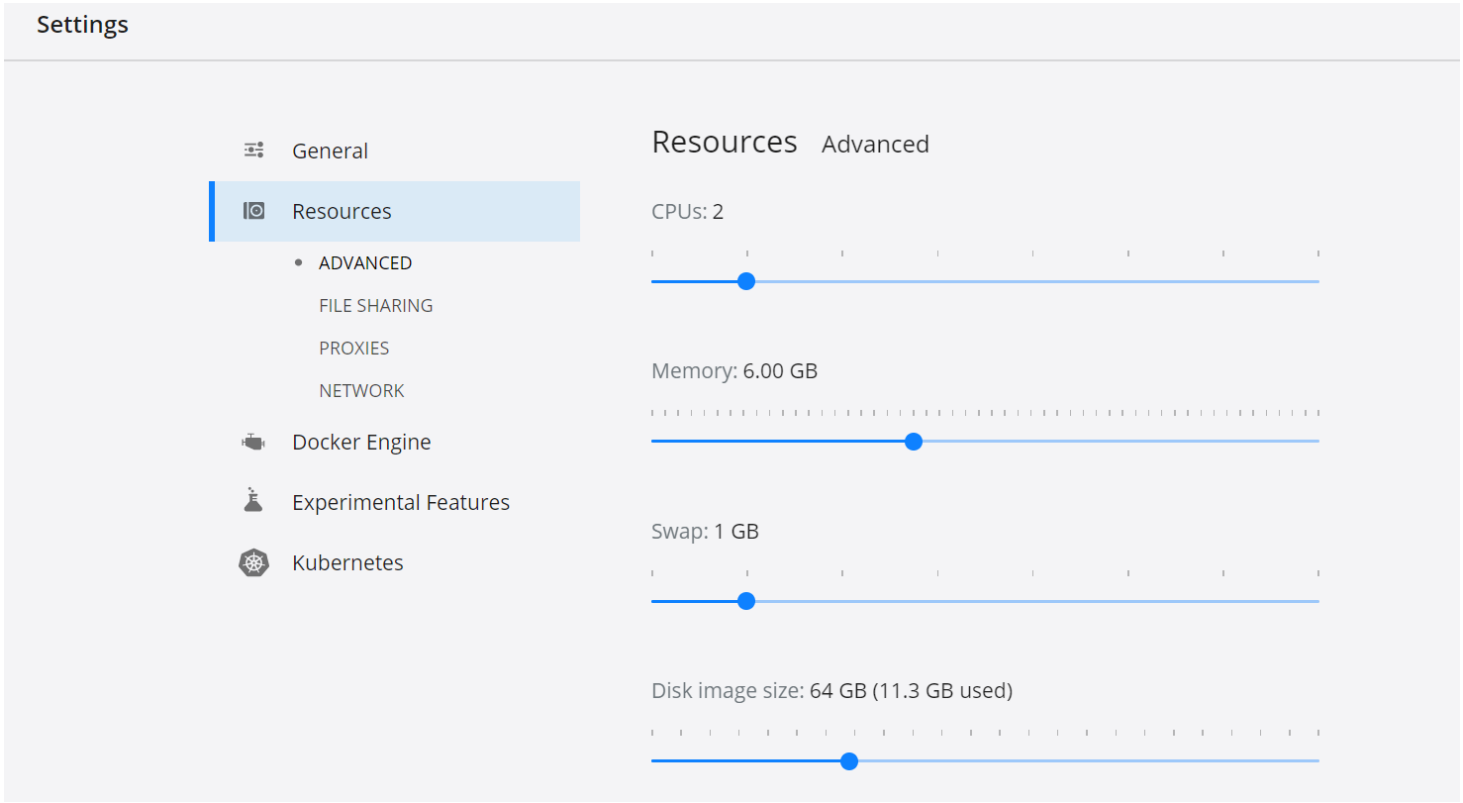


- Docker for Desktop output
- For windows 10 : Run docker command in powershell or command prompt

QUICKSTART : DOCKER INSTALL

2. **Docker Desktop: Update Docker memory**

   Under setting select Resources and update CPU & Memory as mentioned below:



QUICKSTART : DOCKER INSTALL

2. **Docker Toolbox** : **Update Docker memory (optional)**
   2. Create a new VM with 1 CPUs and 4GB of memory (recommended).

   3. 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

3. **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

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 8080:8080 -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
50070	Name node web interface
8088	job tracker :- yarn
7180	Cloudera manager
80	Cloudera examples

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

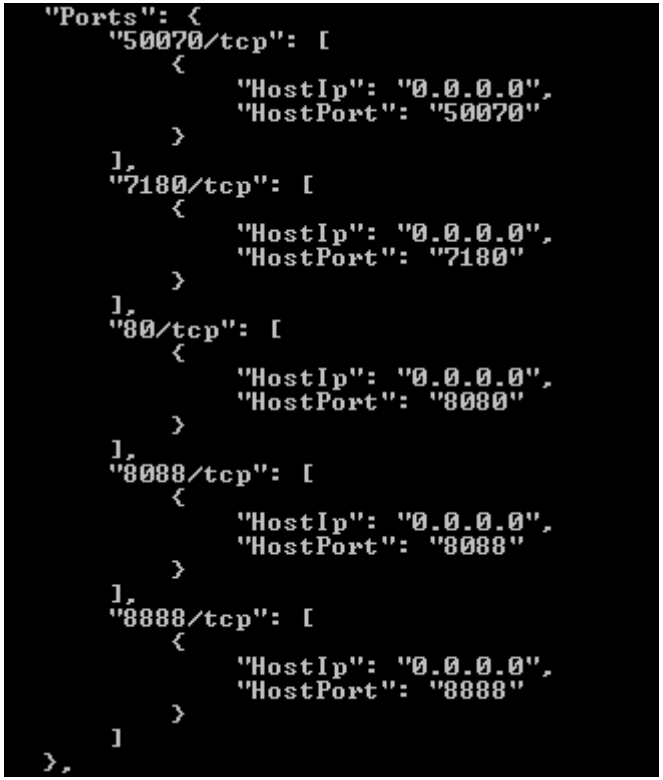
  docker stats [CONTAINER ID]

```
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]



- 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://192.168.99.100:50070/

For other docker install use localhost

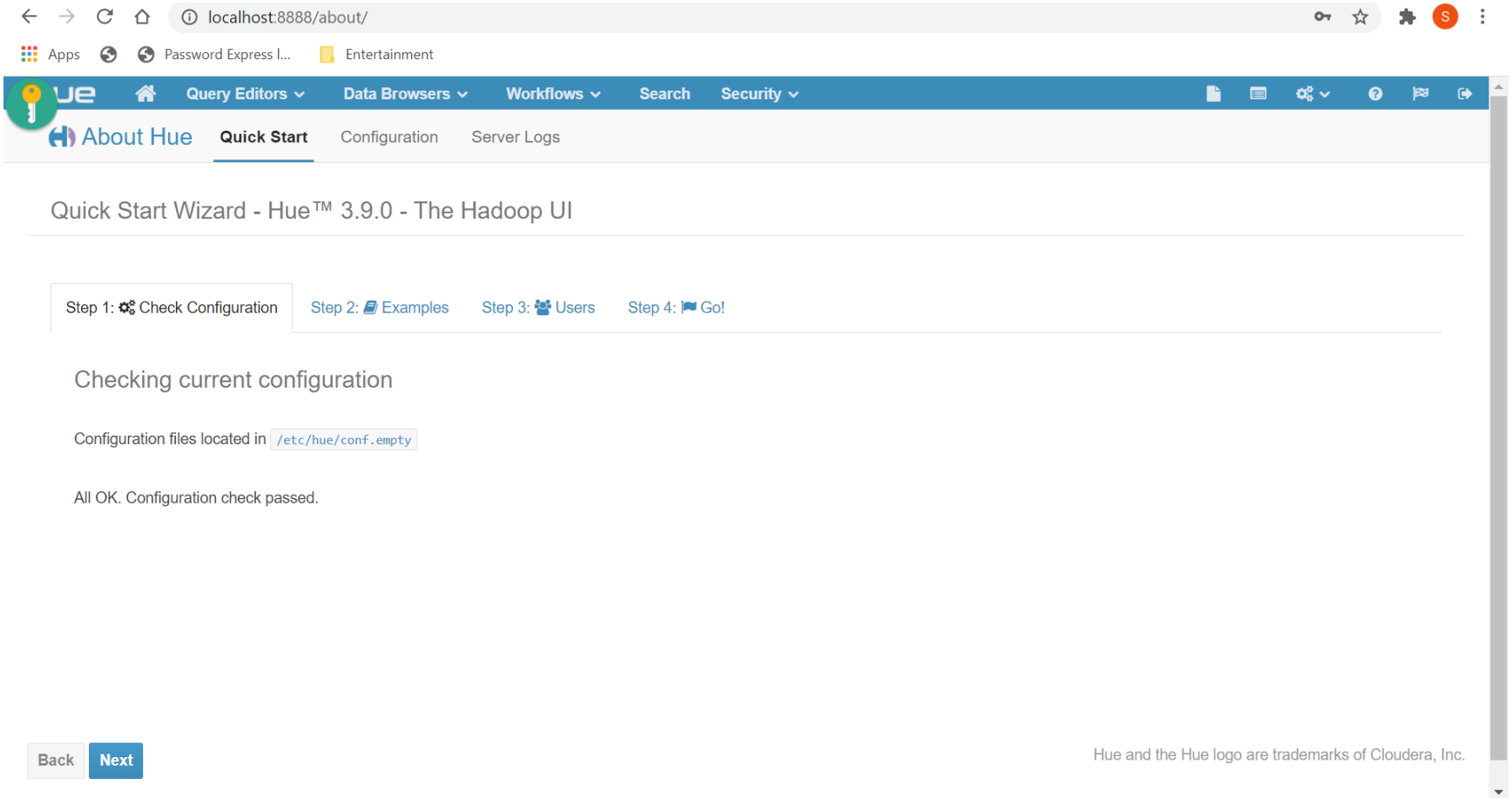
http://localhost:50070/

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

QUICKSTART : DOCKER INSTALL

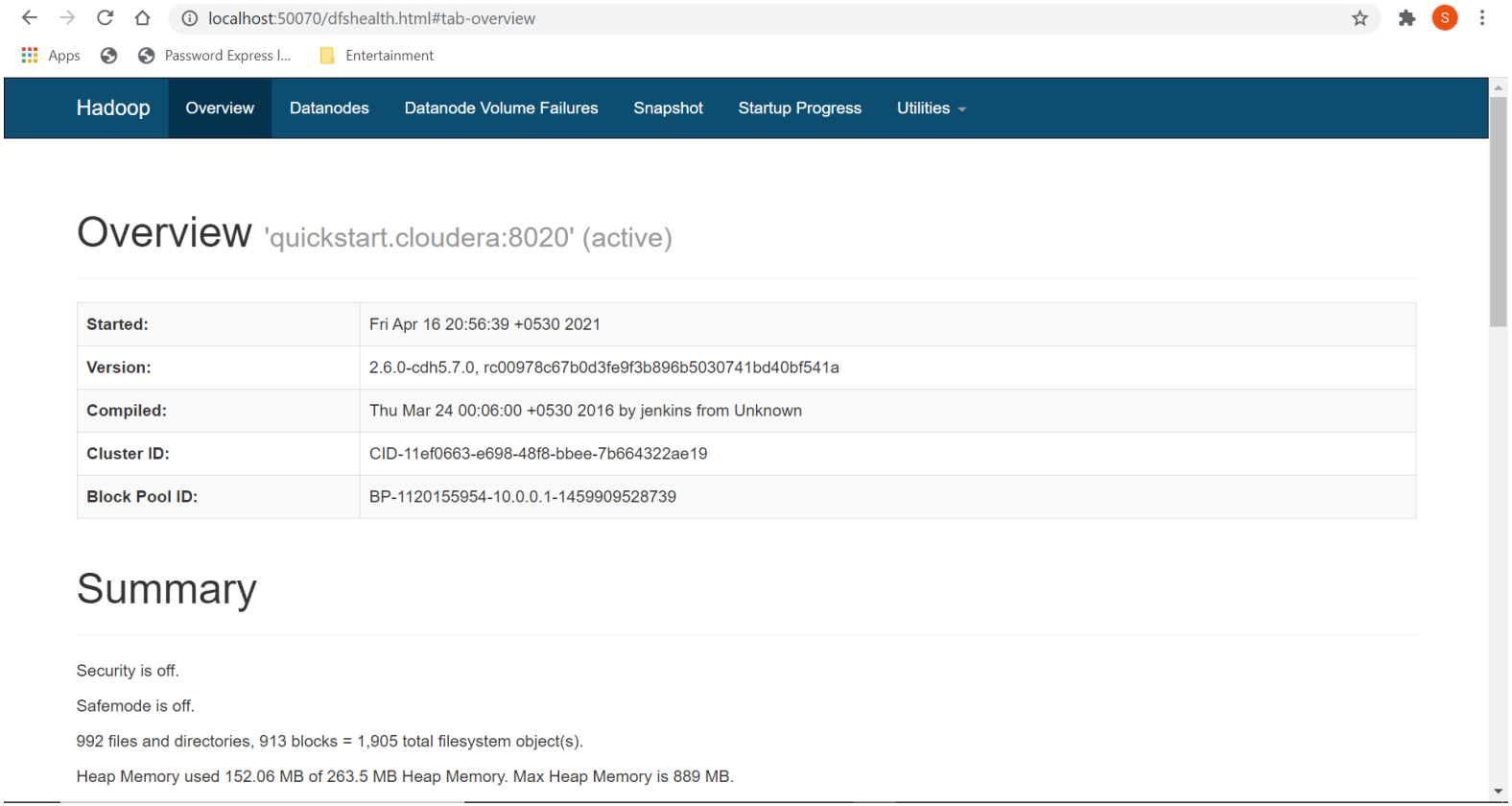
HUE- http://localhost:8888/

Default username / password : cloudera / cloudera



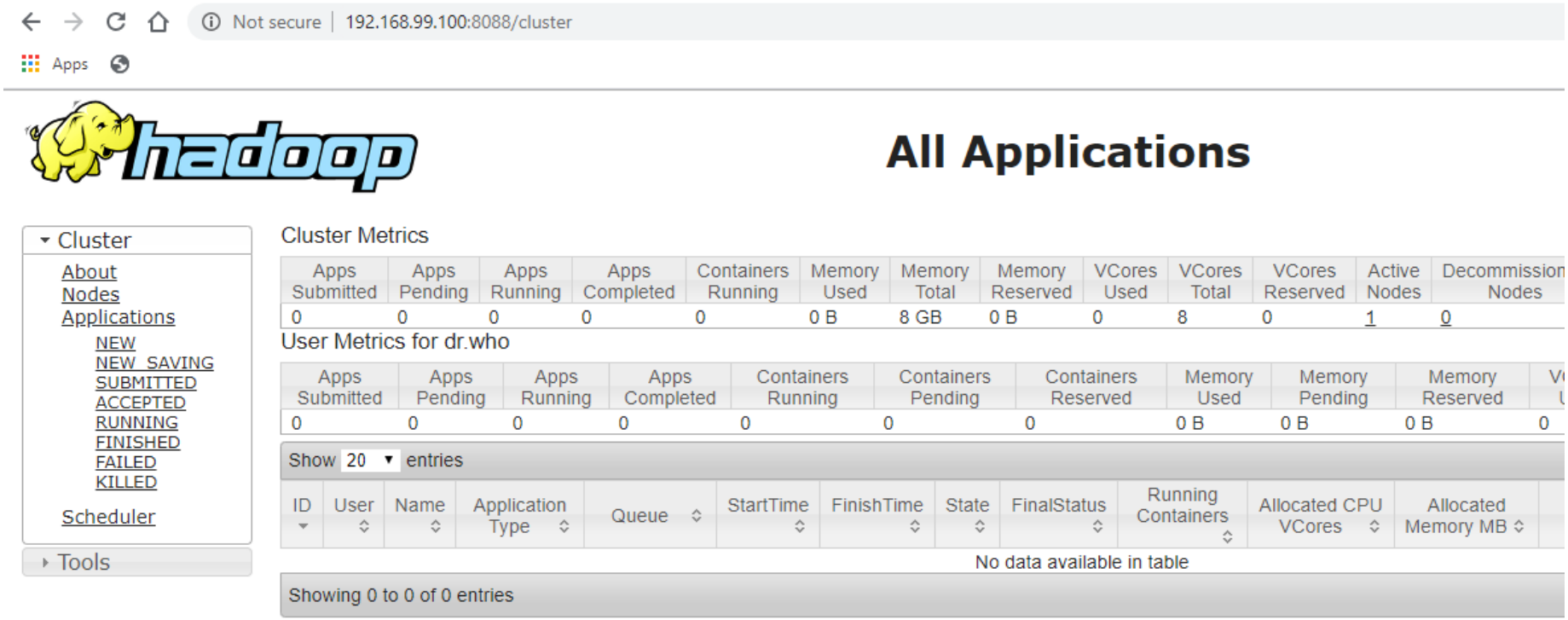
QUICKSTART : DOCKER INSTALL

Name Node - http://localhost:50070/



QUICKSTART : DOCKER INSTALL

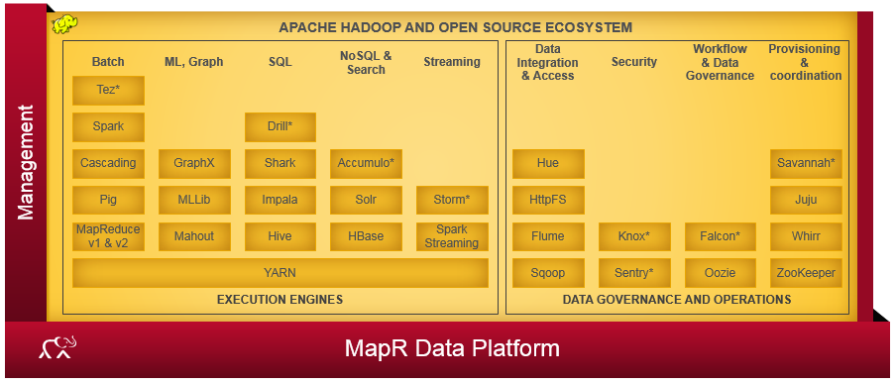
Yarn page - http://192.168.99.100:8088/



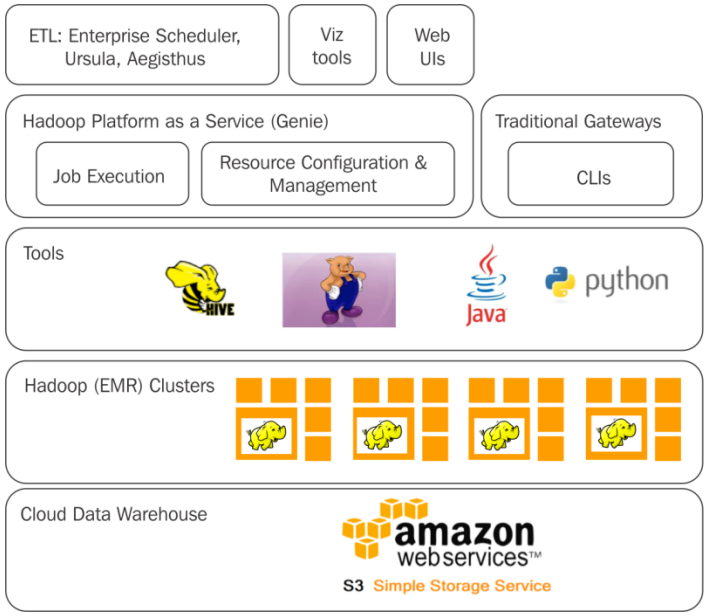
Yarn is resource management layer of Apache Hadoop ecosystem.

Other Vendors

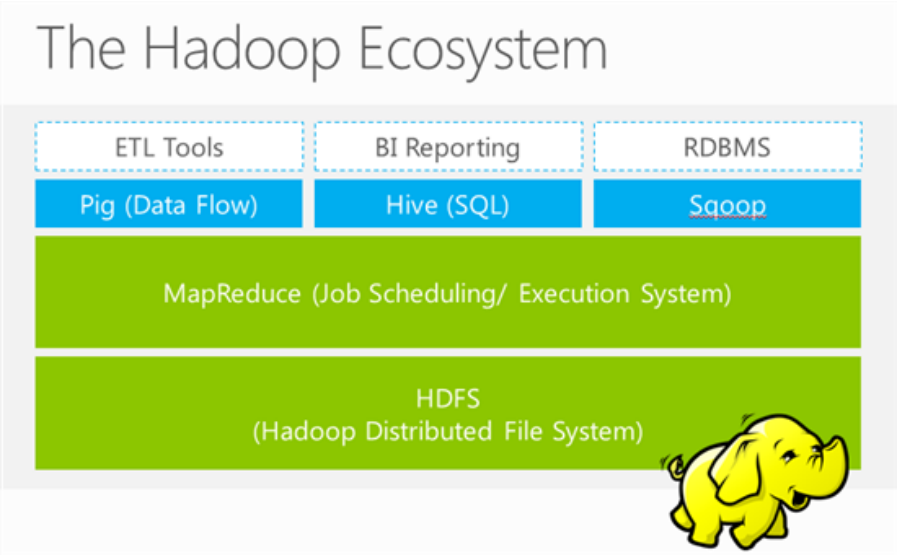
MapR Distribution for Hadoop



AWS EMR



Windows Azure HDInsight



THANK YOU