Apache Hadoop with Cloudera Tutorial 1

Venkatesh Vinayakarao

venkateshv@cmi.ac.in

http://vvtesh.co.in

Chennai Mathematical Institute

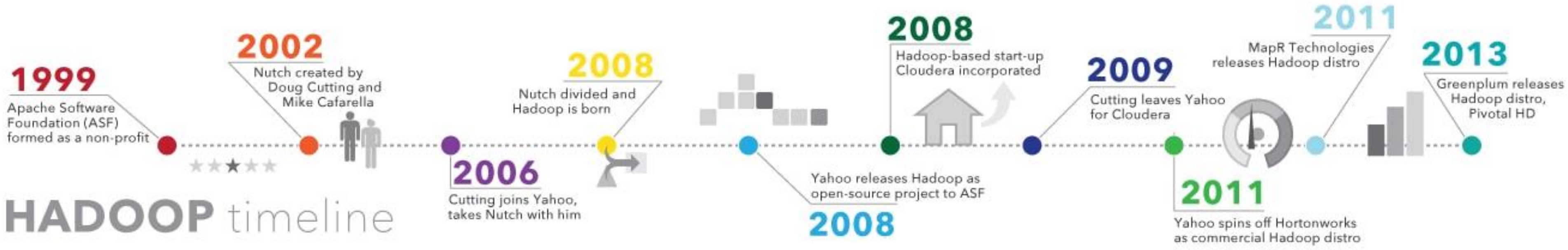
Based on slides by Suchitra Jayaprakash

Apache Hadoop

- Hadoop is open-source software framework used for processing data on distributed commodity computing environment.
- A Java based open-source software managed by Apache Software Foundation.
- Founded by Doug Cutting & Mike Cafarella.
- Based on Google’s white paper on Google File System & map-reduce.

Find Hadoop at https://hadoop.apache.org/

Hadoop - History

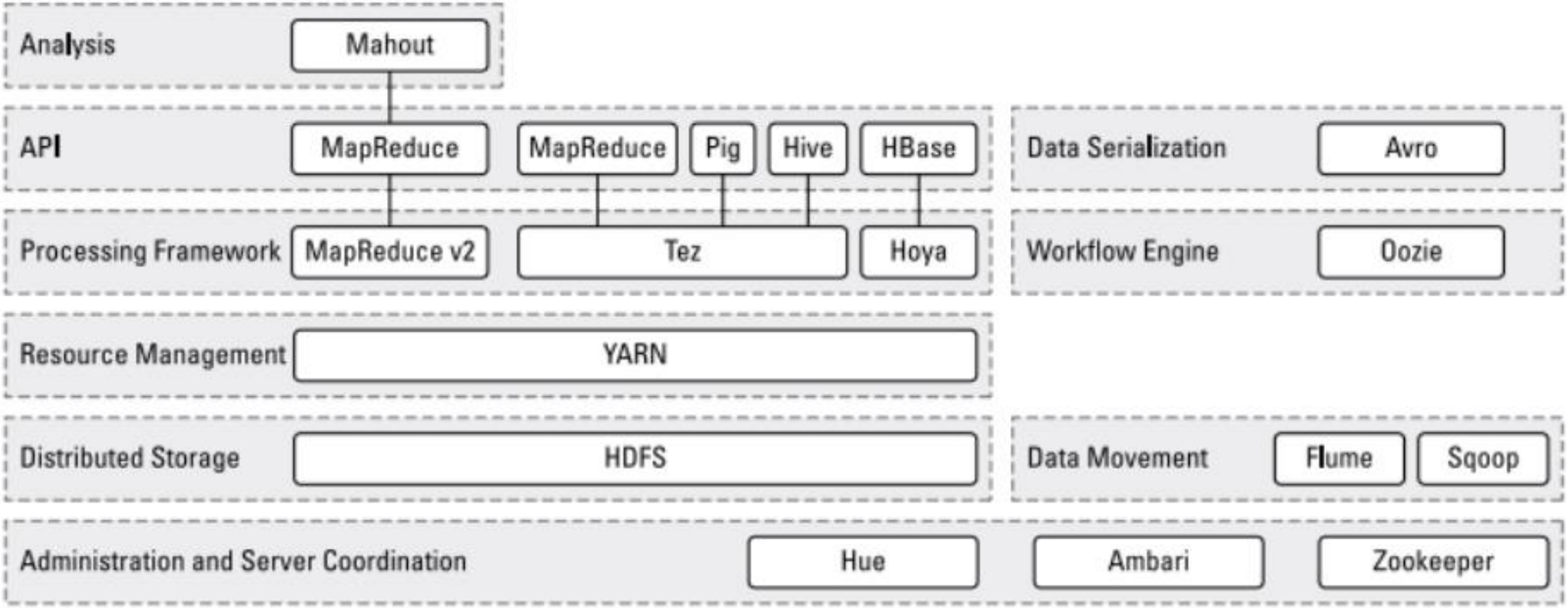

1999
Apache Software Foundation (ASF) formed as a non-profit
2002
Nutch created by Doug Cutting and Mike Cafarella
HADOOP timeline
2006
Cutting joins Yahoo, takes Nutch with him
2008
Nutch divided and Hadoop is born
2008
Yahoo releases Hadoop as open-source project to ASF
2008
Hadoop-based start-up Cloudera incorporated
2009
Cutting leaves Yahoo for Cloudera
2011
Yahoo spins off Hortonworks as commercial Hadoop distro
2011
MapR Technologies releases Hadoop distro
2013
Greenplum releases Hadoop distro, Pivotal HD

Hadoop History - Image Source: https://www.sas.com/en_in/insights/big-data/hadoop.html

Modules

- 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

Hadoop Ecosystem



Analysis: Mahout

API: MapReduce, MapReduce, Pig, Hive, HBase

Processing Framework: MapReduce v2, Tez, Hoya

Resource Management: YARN

Distributed Storage: HDFS

Data Serialization: Avro

Workflow Engine: Oozie

Data Movement: Flume, Sqoop

Administration and Server Coordination: Hue, Ambari, Zookeeper

(source: Hadoop for Dummies)

Hadoop Distribution

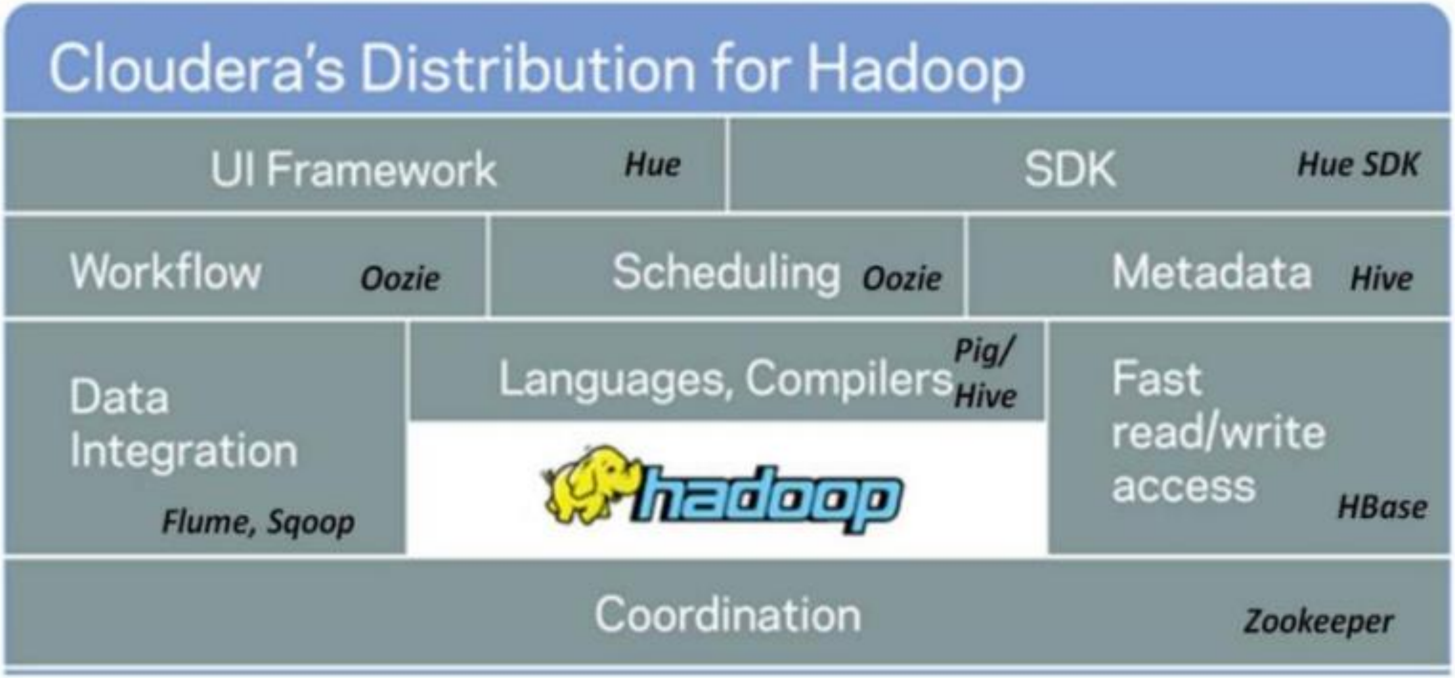
- Need for commercial distributions
  - Customization as per industry needs
  - Maintenance
  - Ease of installation and use
  - Reliability add-ons etc.
- Several vendors offer Hadoop distribution
  - We shall use cloudera for this tutorial

CLOUDERA Hortonworks MAPR amazon web services™ Microsoft Azure IBM Infosphere

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.

Cloudera



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

Tutorial - Objectives

- Understand and Install Cloudera distribution of Hadoop
  - Understand Containers
  - Understand Docker
  - Virtual Machines (VM) Vs. Docker
  - Docker Installation
  - Cloudera Quickstart
- Enjoy Hadoop!

Role of Containers

Why do we need such containers?

Docker

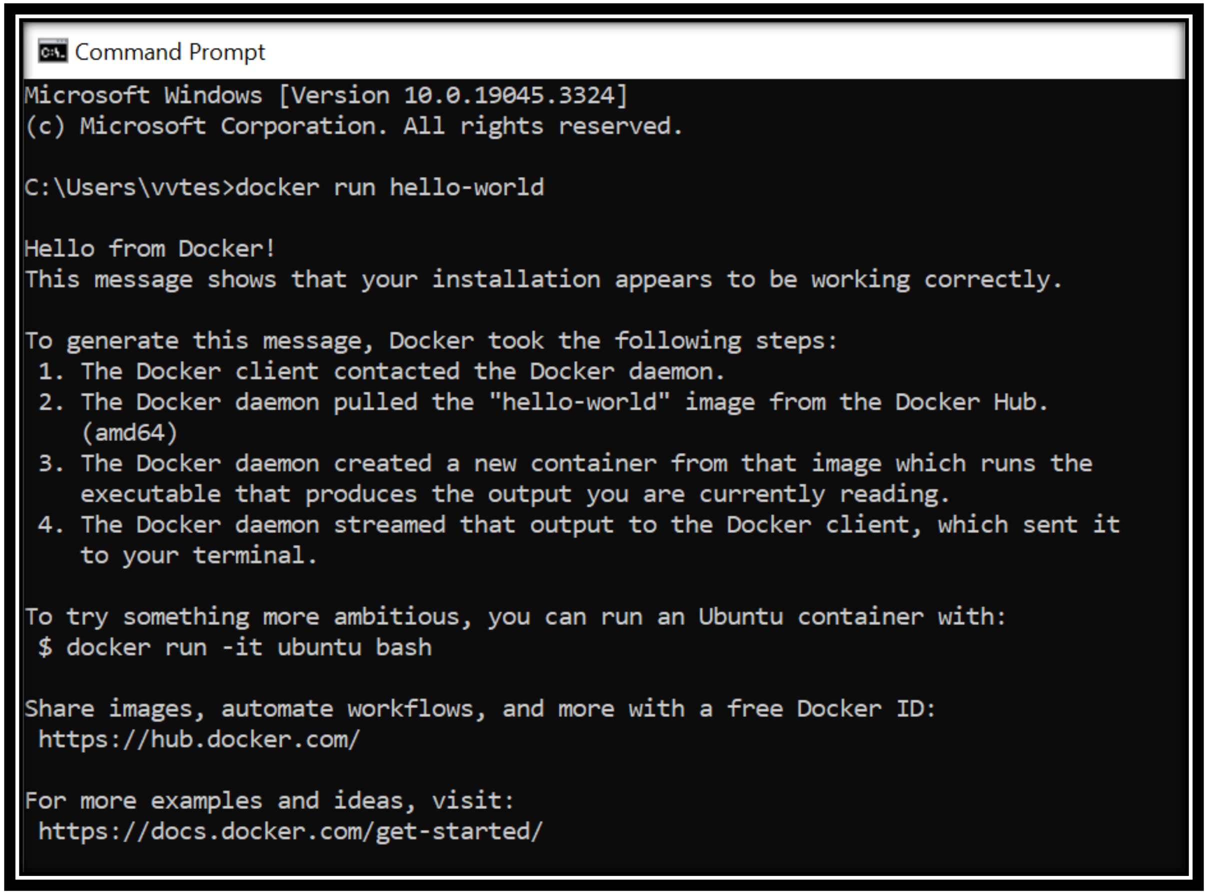
- Docker is an open-source tool that uses containers to create, deploy, and manage distributed applications.
- Developers use containers to create packages for applications that include all libraries that are needed to run the application in isolation.

Docker Installation

- Sign up to https://docs.docker.com/
- Follow instructions at https://docs.docker.com/docker-for-windows/install/
- For this tutorial, we install
  - Docker desktop on Windows
- To check docker installation is proper, open docker terminal (or Command Prompt on Windows) and type: docker run hello-world

Docker Installation

- Successful Installation!

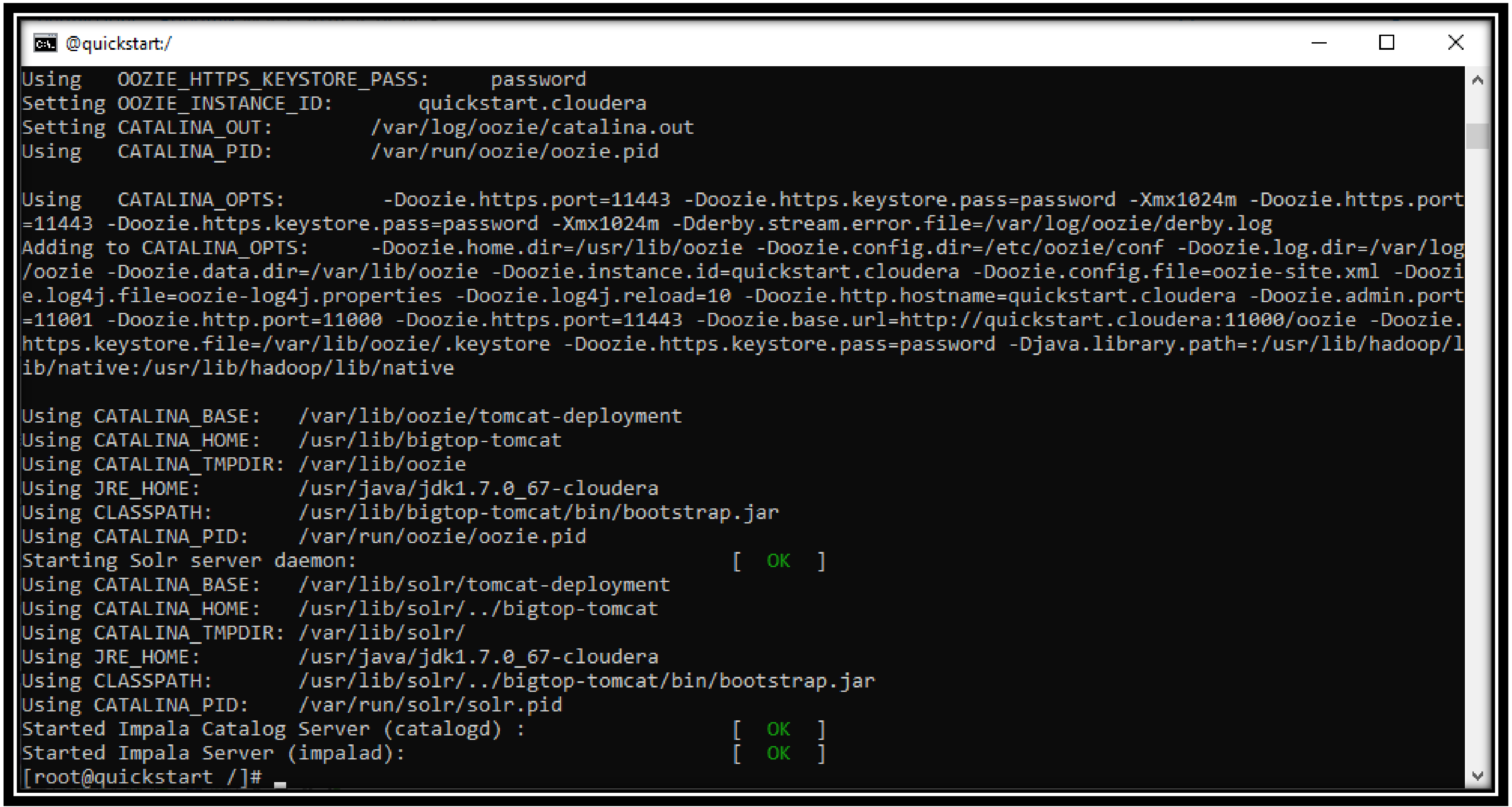


```
Microsoft Windows [Version 10.0.19045.3324]
(c) Microsoft Corporation. All rights reserved.

C:\Users\vvtes>docker run hello-world

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/
```

How to Run Cloudera Hadoop?

- Pull latest cloudera image
  - docker pull cloudera/quickstart:latest
- List the images we have
  - docker images
- Run Cloudera in a docker
  - 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

Cloudera Quickstart

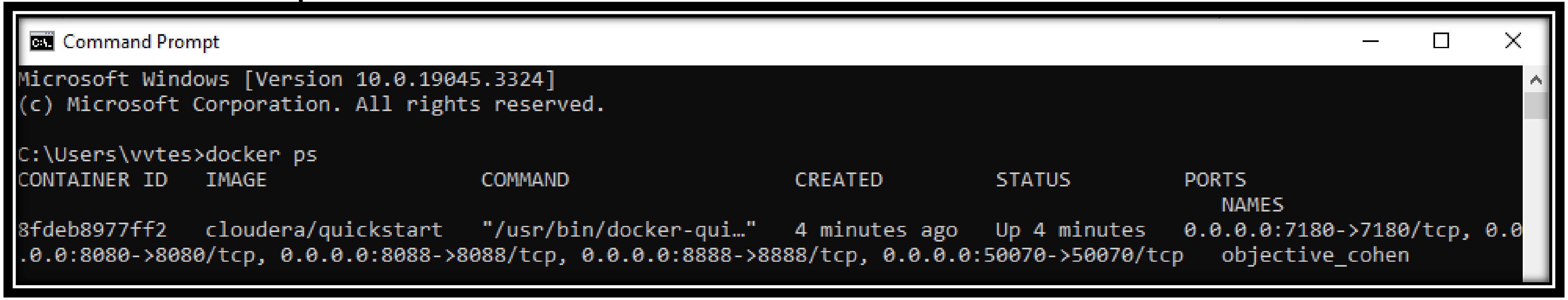


```
Using   OOZIE_HTTPS_KEYSTORE_PASS:     password
Setting OOZIE_INSTANCE_ID:       quickstart.cloudera
Setting CATALINA_OUT:        /var/log/oozie/catalina.out
Using   CATALINA_PID:        /var/run/oozie/oozie.pid

Using   CATALINA_OPTS:        -Doozie.https.port=11443 -Doozie.https.keystore.pass=password -Xmx1024m -Doozie.https.port
=11443 -Doozie.https.keystore.pass=password -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.hostname=quickstart.cloudera -Doozie.admin.port
=11001 -Doozie.http.port=11000 -Doozie.https.port=11443 -Doozie.base.url=http://quickstart.cloudera:11000/oozie -Doozie.
https.keystore.file=/var/lib/oozie/.keystore -Doozie.https.keystore.pass=password -Djava.library.path=:/usr/lib/hadoop/l
ib/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 /]# _
```

To exit, [root@quickstart /]# exit

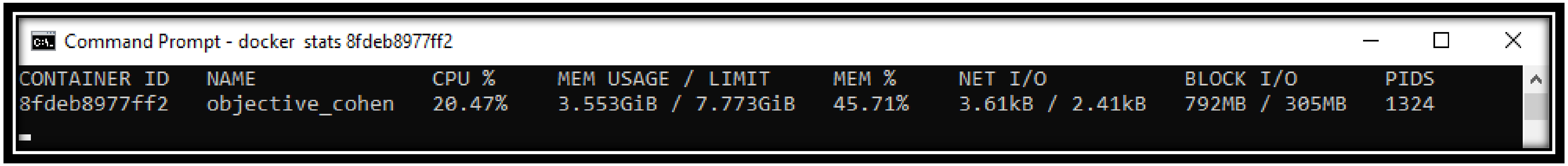
Some useful commands

- docker ps



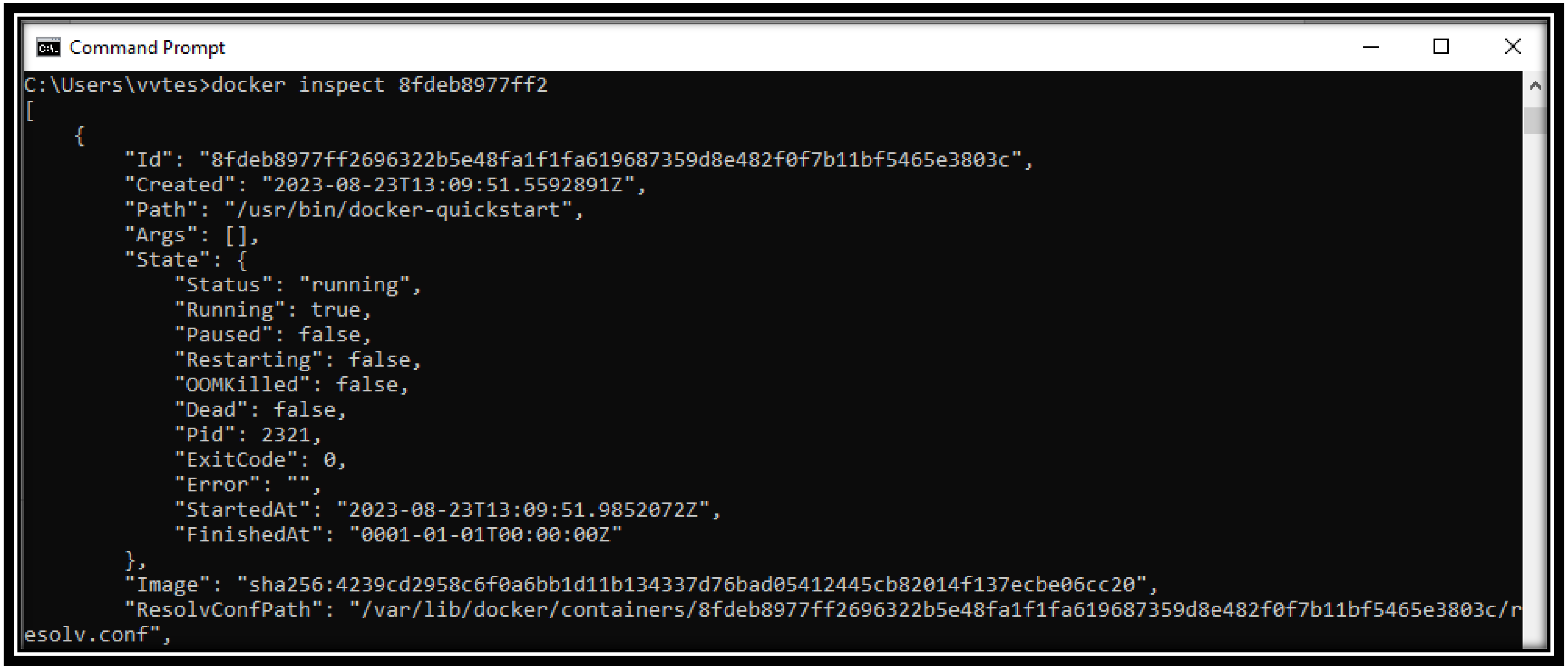
```
Microsoft Windows [Version 10.0.19045.3324]
(c) Microsoft Corporation. All rights reserved.

C:\Users\vvtes>docker ps
CONTAINER ID   IMAGE                  COMMAND                  CREATED          STATUS          PORTS
                                                                                                   NAMES
8fdeb8977ff2   cloudera/quickstart   "/usr/bin/docker-qui…"   4 minutes ago    Up 4 minutes    0.0.0.0:7180->7180/tcp, 0.0
.0.0:8080->8080/tcp, 0.0.0.0:8088->8088/tcp, 0.0.0.0:8888->8888/tcp, 0.0.0.0:50070->50070/tcp   objective_cohen
```

- docker stats [containerid]



```
CONTAINER ID   NAME              CPU %     MEM USAGE / LIMIT     MEM %     NET I/O             BLOCK I/O         PIDS
8fdeb8977ff2   objective_cohen   20.47%    3.553GiB / 7.773GiB   45.71%    3.61kB / 2.41kB     792MB / 305MB     1324
```

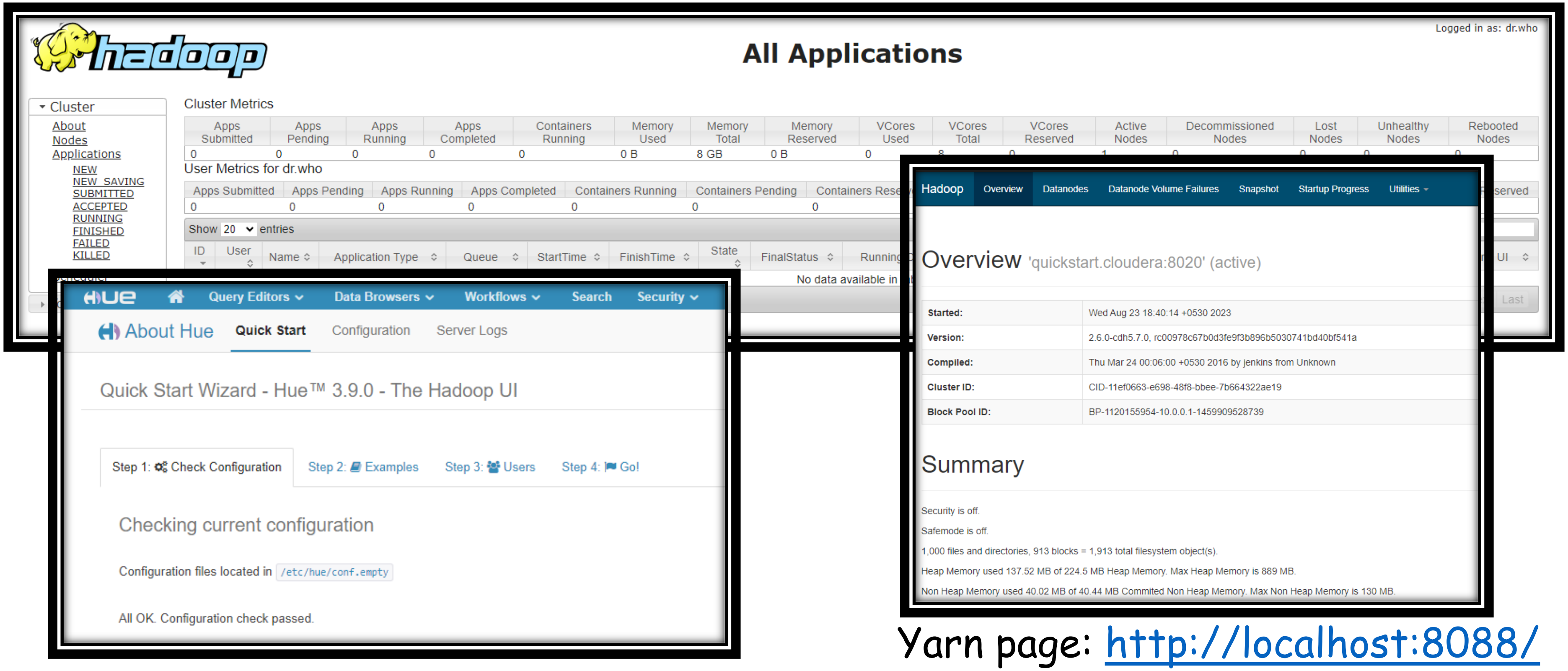
- docker inspect [CONTAINER ID]



```
C:\Users\vvtes>docker inspect 8fdeb8977ff2
[
    {
        "Id": "8fdeb8977ff2696322b5e48fa1f1fa619687359d8e482f0f7b11bf5465e3803c",
        "Created": "2023-08-23T13:09:51.5592891Z",
        "Path": "/usr/bin/docker-quickstart",
        "Args": [],
        "State": {
            "Status": "running",
            "Running": true,
            "Paused": false,
            "Restarting": false,
            "OOMKilled": false,
            "Dead": false,
            "Pid": 2321,
            "ExitCode": 0,
            "Error": "",
            "StartedAt": "2023-08-23T13:09:51.9852072Z",
            "FinishedAt": "0001-01-01T00:00:00Z"
        },
        "Image": "sha256:4239cd2958c6f0a6bb1d11b134337d76bad05412445cb82014f137ecbe06cc20",
        "ResolvConfPath": "/var/lib/docker/containers/8fdeb8977ff2696322b5e48fa1f1fa619687359d8e482f0f7b11bf5465e3803c/r
esolv.conf",
```

We are all set!

NameNode: http://localhost:50070/



All Applications

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	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	0	0	0	0 B	8 GB	0 B	0	8	0	1	0	0	0	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Rese
0	0	0	0	0	0	0

Show 20 entries

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running
								No data available in	

Quick Start Wizard - Hue™ 3.9.0 - The Hadoop UI

Checking current configuration

Configuration files located in `/etc/hue/conf.empty`

All OK. Configuration check passed.

Overview 'quickstart.cloudera:8020' (active)

Started:	Wed Aug 23 18:40:14 +0530 2023
Version:	2.6.0-cdh5.7.0, rc00978c67b0d3fe9f3b896b5030741bd40bf541a
Compiled:	Thu Mar 24 00:06:00 +0530 2016 by jenkins from Unknown
Cluster ID:	CID-11ef0663-e698-48f8-bbee-7b664322ae19
Block Pool ID:	BP-1120155954-10.0.0.1-1459909528739

Summary

Security is off.

Safemode is off.

1,000 files and directories, 913 blocks = 1,913 total filesystem object(s).

Heap Memory used 137.52 MB of 224.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 40.02 MB of 40.44 MB Commited Non Heap Memory. Max Non Heap Memory is 130 MB.

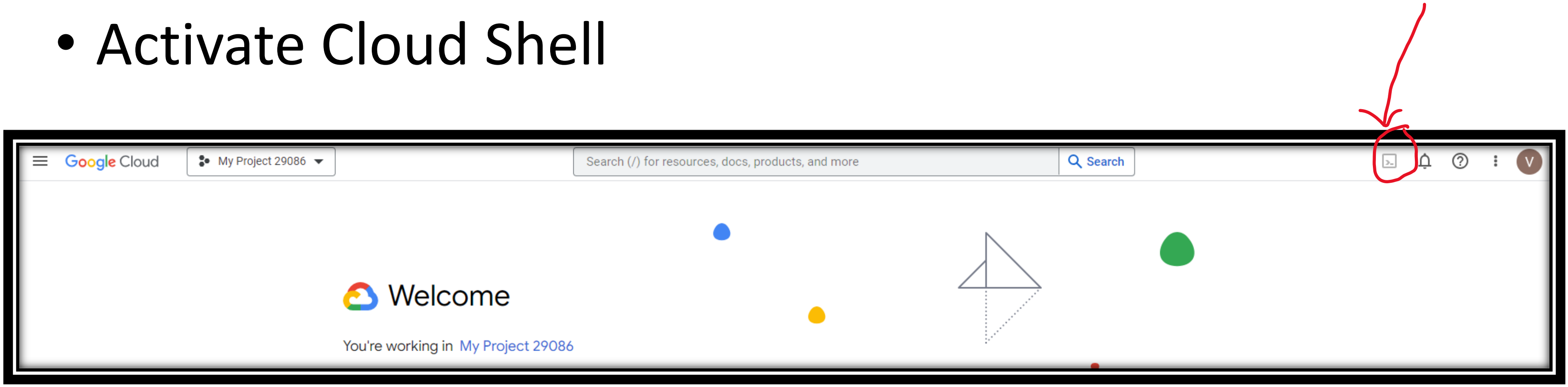
Yarn page: http://localhost:8088/

HUE- http://localhost:8888/

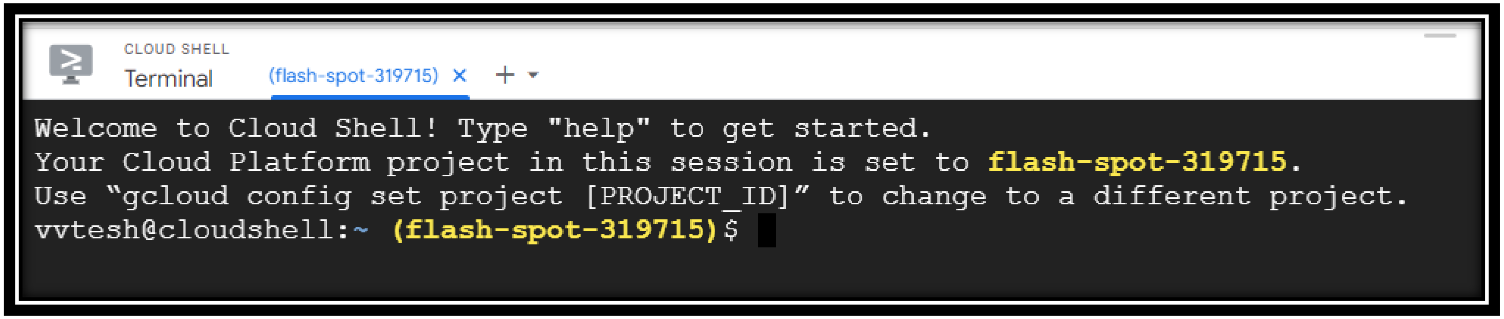
Default username / password : cloudera / cloudera

We could do all that on the cloud!

- Go to Google Cloud
  - https://cloud.google.com/
- Activate Cloud Shell


Google Cloud
My Project 29086
Search (/) for resources, docs, products, and more
Search
V
Welcome
You're working in My Project 29086

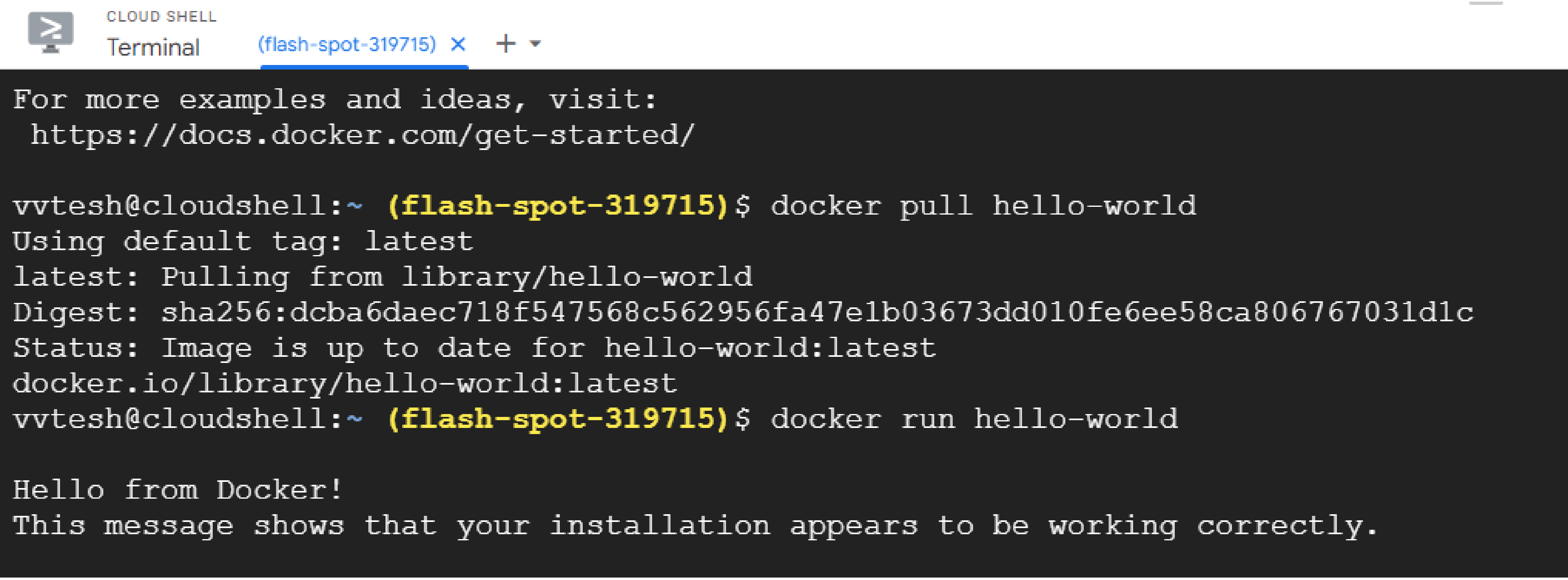
- You should see this



```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to flash-spot-319715.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
vvtesh@cloudshell:~ (flash-spot-319715)$ 
```

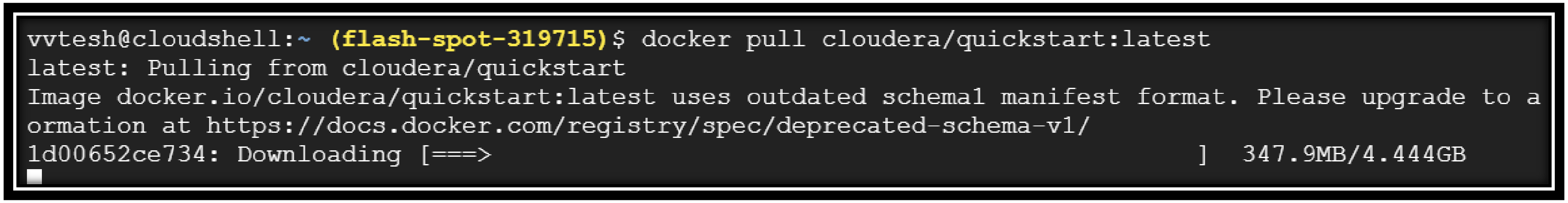
Let us install cloudera here!

- Docker is pre-installed.

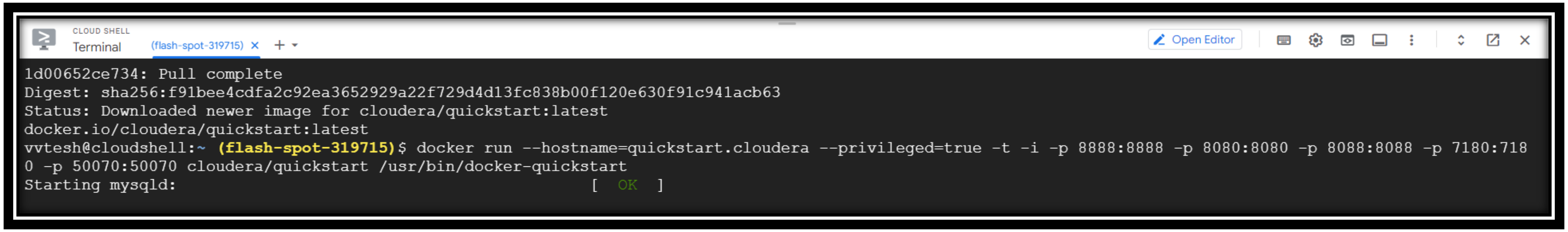


```
For more examples and ideas, visit:
 https://docs.docker.com/get-started/

vvtesh@cloudshell:~ (flash-spot-319715)$ docker pull hello-world
Using default tag: latest
latest: Pulling from library/hello-world
Digest: sha256:dcba6daec718f547568c562956fa47e1b03673dd010fe6ee58ca806767031d1c
Status: Image is up to date for hello-world:latest
docker.io/library/hello-world:latest
vvtesh@cloudshell:~ (flash-spot-319715)$ docker run hello-world

Hello from Docker!
This message shows that your installation appears to be working correctly.
```

Install Cloudera



```
vvtesh@cloudshell:~ (flash-spot-319715)$ 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
ormation at https://docs.docker.com/registry/spec/deprecated-schema-v1/
1d00652ce734: Downloading [===>                                              ]  347.9MB/4.444GB
```

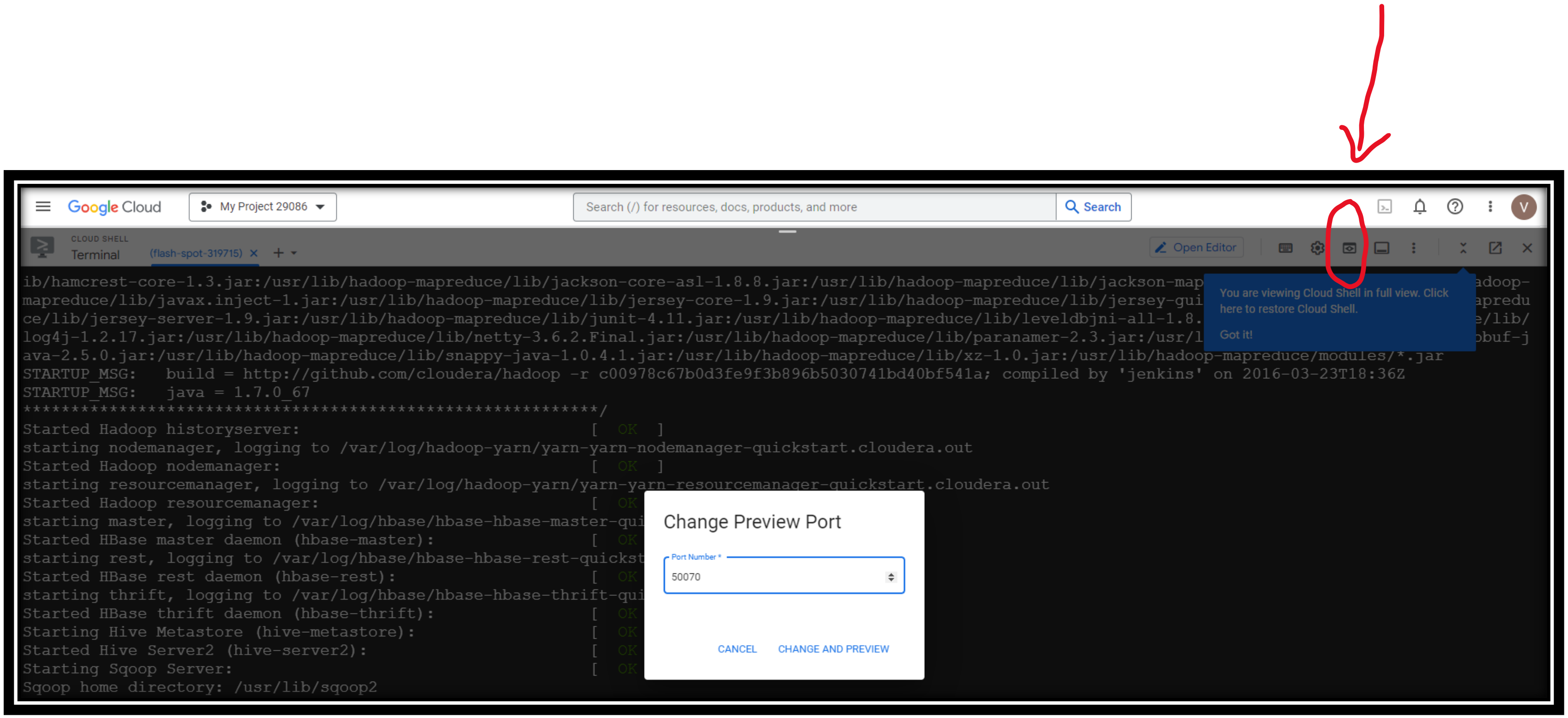


```
1d00652ce734: Pull complete
Digest: sha256:f91bee4cdfa2c92ea3652929a22f729d4d13fc838b00f120e630f91c941acb63
Status: Downloaded newer image for cloudera/quickstart:latest
docker.io/cloudera/quickstart:latest
vvtesh@cloudshell:~ (flash-spot-319715)$ docker run --hostname=quickstart.cloudera --privileged=true -t -i -p 8888:8888 -p 8080:8080 -p 8088:8088 -p 7180:718
0 -p 50070:50070 cloudera/quickstart /usr/bin/docker-quickstart
Starting mysqld:                                           [  OK  ]
```

Try the Preview Port

Click here

Change port to 50070



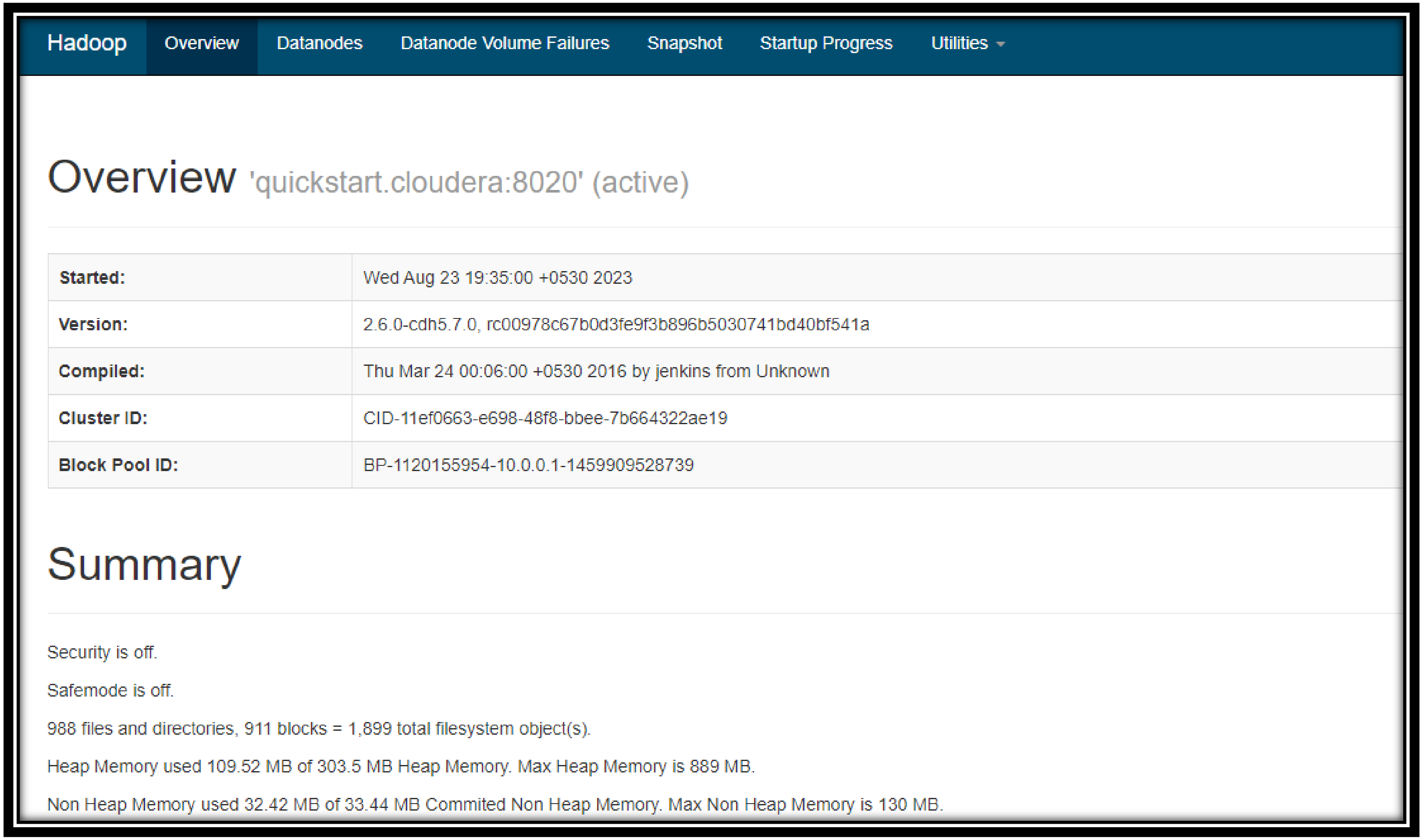
You are viewing Cloud Shell in full view. Click here to restore Cloud Shell.

```
ib/hamcrest-core-1.3.jar:/usr/lib/hadoop-mapreduce/lib/jackson-core-asl-1.8.8.jar:/usr/lib/hadoop-mapreduce/lib/jackson-map                               adoop-
mapreduce/lib/javax.inject-1.jar:/usr/lib/hadoop-mapreduce/lib/jersey-core-1.9.jar:/usr/lib/hadoop-mapreduce/lib/jersey-gui                               apredu
ce/lib/jersey-server-1.9.jar:/usr/lib/hadoop-mapreduce/lib/junit-4.11.jar:/usr/lib/hadoop-mapreduce/lib/leveldbjni-all-1.8.                              e/lib/
log4j-1.2.17.jar:/usr/lib/hadoop-mapreduce/lib/netty-3.6.2.Final.jar:/usr/lib/hadoop-mapreduce/lib/paranamer-2.3.jar:/usr/l                              obuf-j
ava-2.5.0.jar:/usr/lib/hadoop-mapreduce/lib/snappy-java-1.0.4.1.jar:/usr/lib/hadoop-mapreduce/lib/xz-1.0.jar:/usr/lib/hadoop-mapreduce/modules/*.jar
STARTUP_MSG:   build = http://github.com/cloudera/hadoop -r c00978c67b0d3fe9f3b896b5030741bd40bf541a; compiled by 'jenkins' on 2016-03-23T18:36Z
STARTUP_MSG:   java = 1.7.0_67
************************************************************/
Started Hadoop historyserver:                              [  OK  ]
starting nodemanager, logging to /var/log/hadoop-yarn/yarn-yarn-nodemanager-quickstart.cloudera.out
Started Hadoop nodemanager:                                [  OK  ]
starting resourcemanager, logging to /var/log/hadoop-yarn/yarn-yarn-resourcemanager-quickstart.cloudera.out
Started Hadoop resourcemanager:                            [  OK
starting master, logging to /var/log/hbase/hbase-hbase-master-qui
Started HBase master daemon (hbase-master):                [  OK
starting rest, logging to /var/log/hbase/hbase-hbase-rest-quickst
Started HBase rest daemon (hbase-rest):                    [  OK
starting thrift, logging to /var/log/hbase/hbase-hbase-thrift-qui
Started HBase thrift daemon (hbase-thrift):                [  OK
Starting Hive Metastore (hive-metastore):                  [  OK
Started Hive Server2 (hive-server2):                       [  OK
Starting Sqoop Server:                                     [  OK
Sqoop home directory: /usr/lib/sqoop2
```

Change Preview Port

Port Number *: 50070

Namenode Information Opens in a Browser



Overview 'quickstart.cloudera:8020' (active)

Started:	Wed Aug 23 19:35:00 +0530 2023
Version:	2.6.0-cdh5.7.0, rc00978c67b0d3fe9f3b896b5030741bd40bf541a
Compiled:	Thu Mar 24 00:06:00 +0530 2016 by jenkins from Unknown
Cluster ID:	CID-11ef0663-e698-48f8-bbee-7b664322ae19
Block Pool ID:	BP-1120155954-10.0.0.1-1459909528739

Summary

Security is off.

Safemode is off.

988 files and directories, 911 blocks = 1,899 total filesystem object(s).

Heap Memory used 109.52 MB of 303.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 32.42 MB of 33.44 MB Commited Non Heap Memory. Max Non Heap Memory is 130 MB.

Thank You