

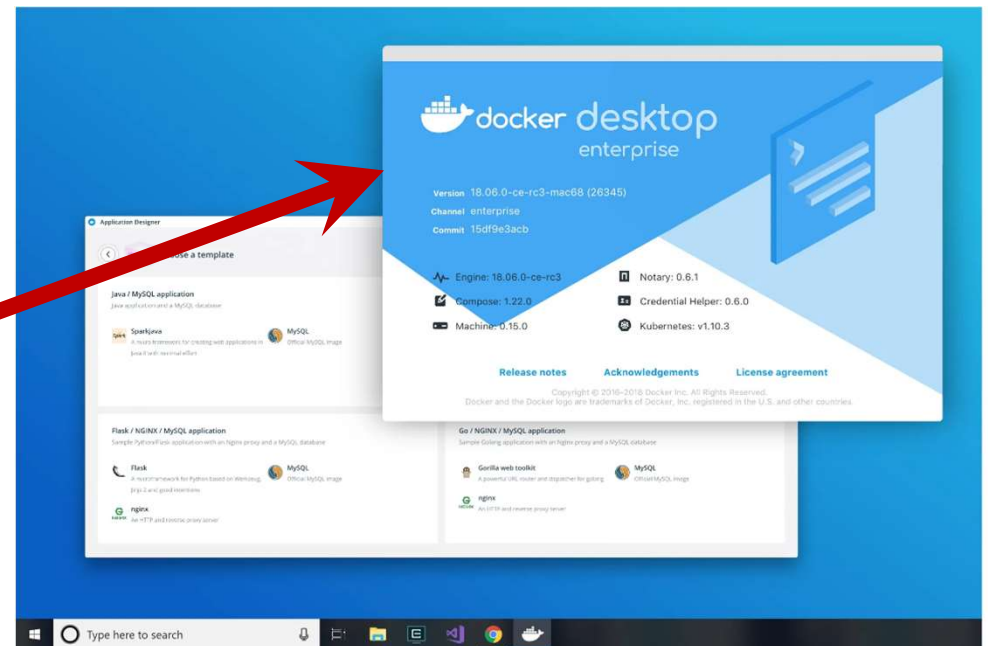


Running Hadoop with Docker

**Pseudo Distributed mode using a
single Docker container**

Environment Used

- Host environment
 - Windows
- Container environment
 - Docker Desktop
- SSH to interact with container





Steps

1 – Configuration

- Generate SSH keys
- Build Docker image
- Run container from image
- Install Hadoop

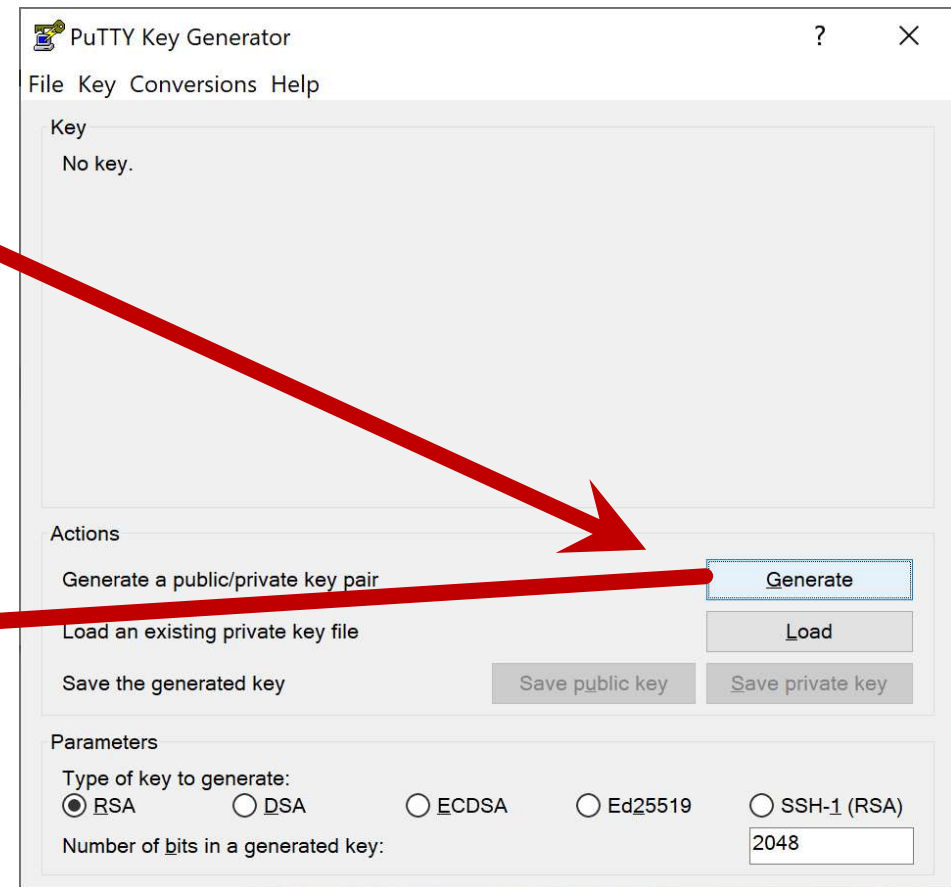
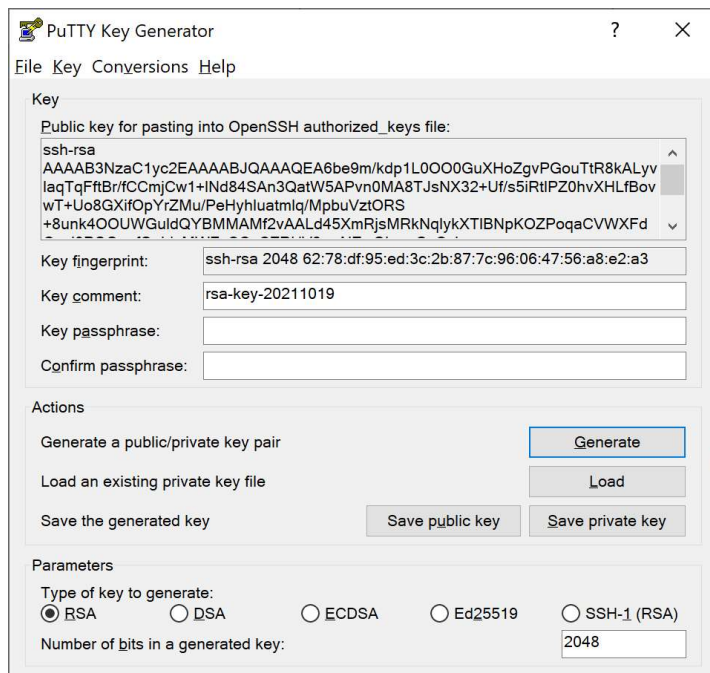
2 – Testing

- Install examples
- Configure examples
- Build examples
- Run Word Count example

Configuration

Generate SSH keys

- Generate a public/private key pair using PuTTYgen



Configuration

Generate SSH keys

- Save both keys:

- Public

- Private

Password to protect access to the private key

Private key will be managed by the PuTTY agent

Public key will be later placed in the **authorized_keys** file

The same key can be used in different computers

The screenshot shows the PuTTY Key Generator window. The 'Key' section displays the public key for pasting into the OpenSSH authorized_keys file. The 'Key fingerprint' is shown as 'ssh-rsa 2048 62:78:df:95:ed:3c:2b:87:7c:96:06:47:56:a8:e2:a3'. The 'Key comment' is 'rsa-key-20211019'. The 'Key passphrase' and 'Confirm passphrase' fields are filled with dots. The 'Actions' section includes buttons for 'Generate', 'Load', 'Save public key', and 'Save private key'. The 'Parameters' section shows 'Type of key to generate' set to 'RSA' and 'Number of bits in a generated key' set to '2048'.

Configuration

Files need to build Docker image

Definition for user aliases

Build Docker image

Run Docker container

Scripts for the user bin directory

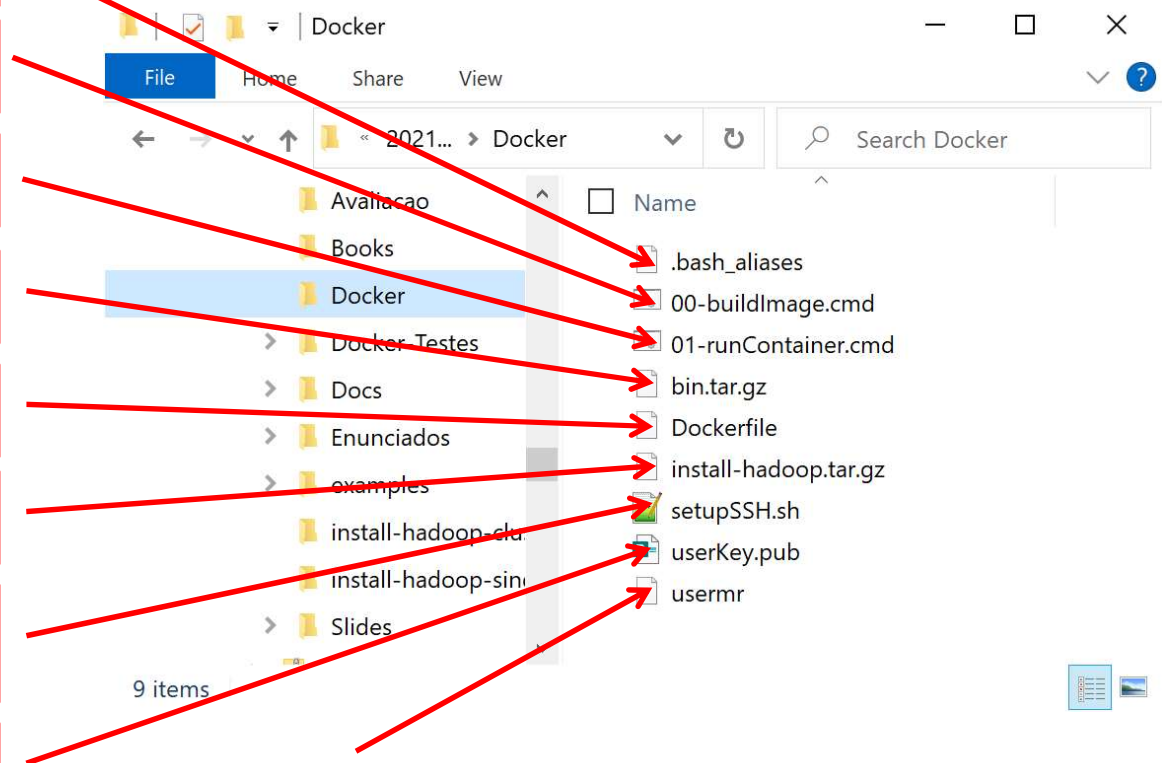
Docker file

Scripts used to install Hadoop

Scripts used to setup SSH in
password less mode

SSH public key

Remove sudo password



Configuration

Build Docker image

Use the definitions
contained in the file
"Dockerfile"

- Execute script "goBuildDocker.cmd"

```
docker build -t cdle.ubuntu.2023.2024 .
```

```
C:\WINDOWS\system32\cmd.exe
docker build -t hadoop.ubuntu .
[+] Building 1.1s (2/3)
=> [internal] load build definition from Dockerfile                                0.0s
=> => transferring dockerfile: 32B                                                0.0s
=> [internal] load .dockerignore                                                  0.0s
=> => transferring context: 2B                                                    0.0s
=> [internal] load metadata for docker.io/library/ubuntu:latest                 1.0s

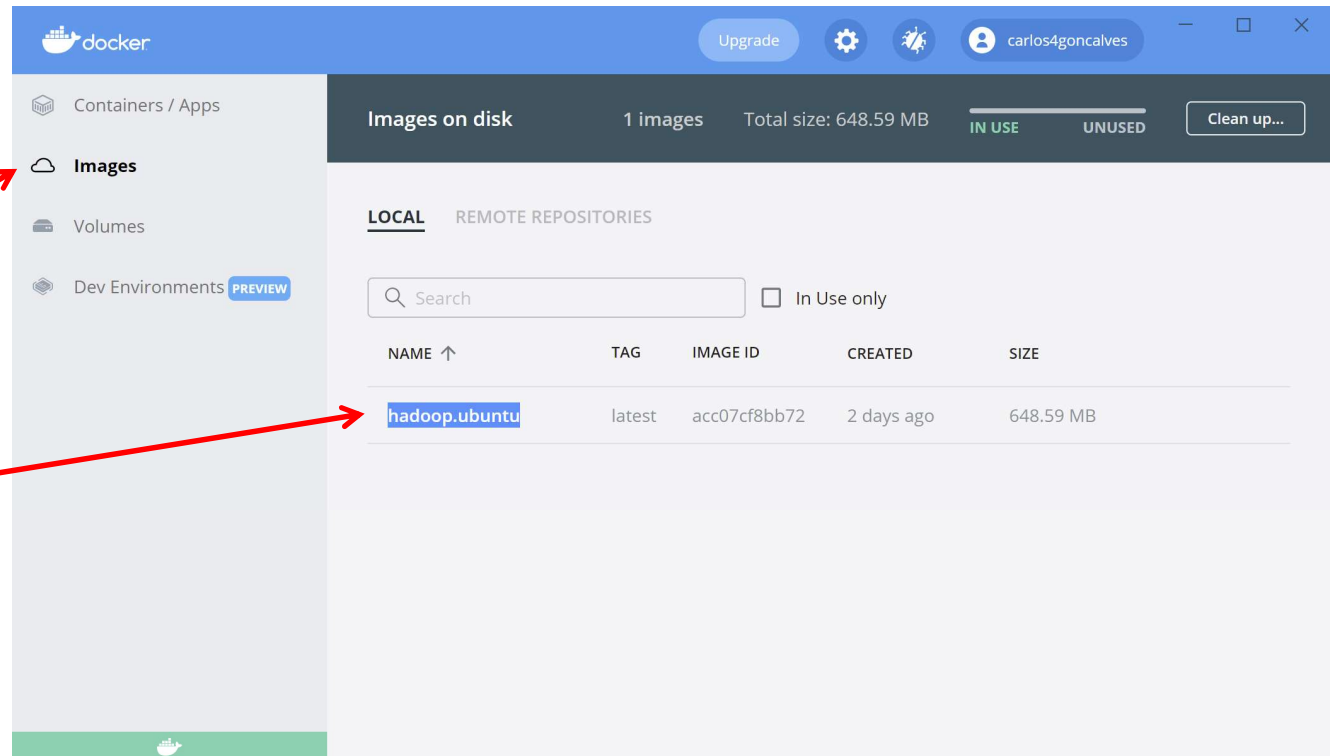
=> exporting to image                                                            0.1s
=> => exporting layers                                                            0.0s
=> => writing image sha256:acc07cf8bb72730f75d5eba4dbe5f84c26d6ba5c1c6f841ae7355a5c719000c7 0.0s
=> => naming to docker.io/library/hadoop.ubuntu                                0.0s

Use 'docker scan' to run Snyk tests against images to find vulnerabilities and learn how to fix them
Press any key to continue . . .
```

Configuration

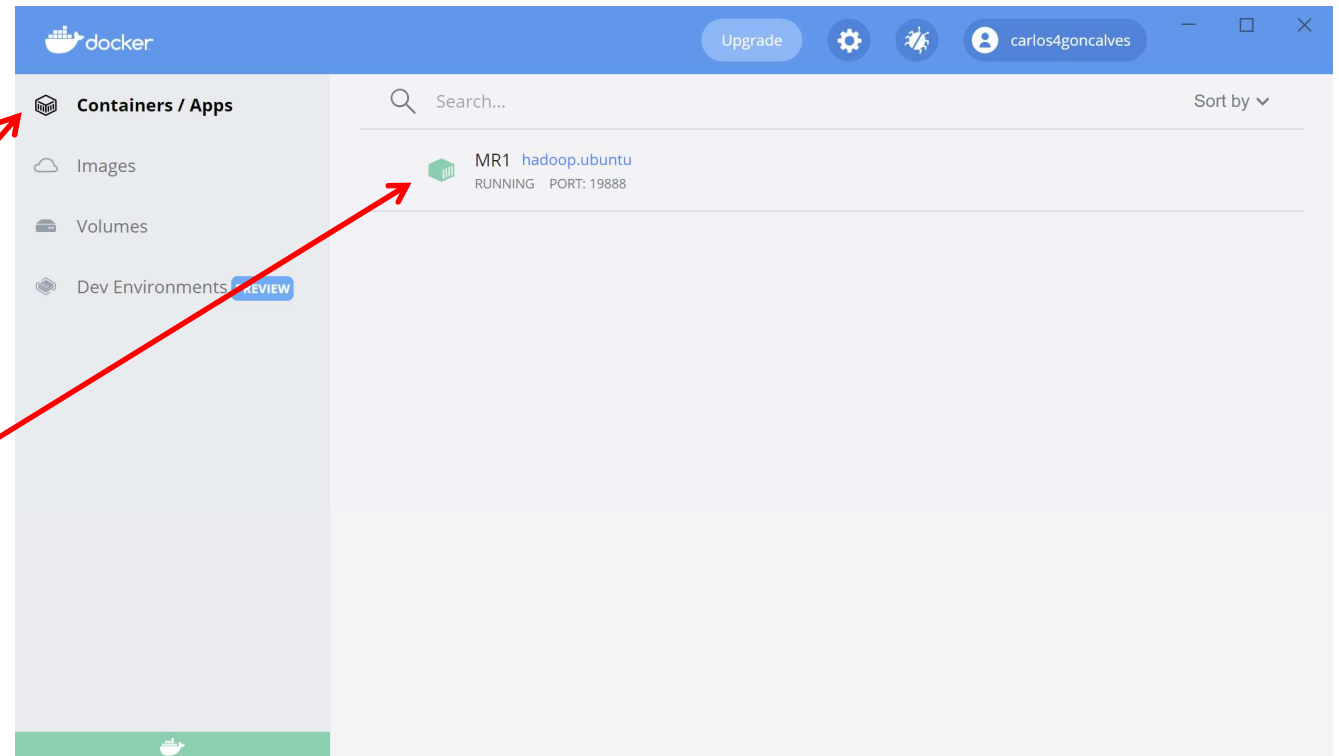
Build Docker image

```
docker compose -f docker-compose-23-24.yml -p cdle-23-24 up -d
```



Configuration

Run container from image



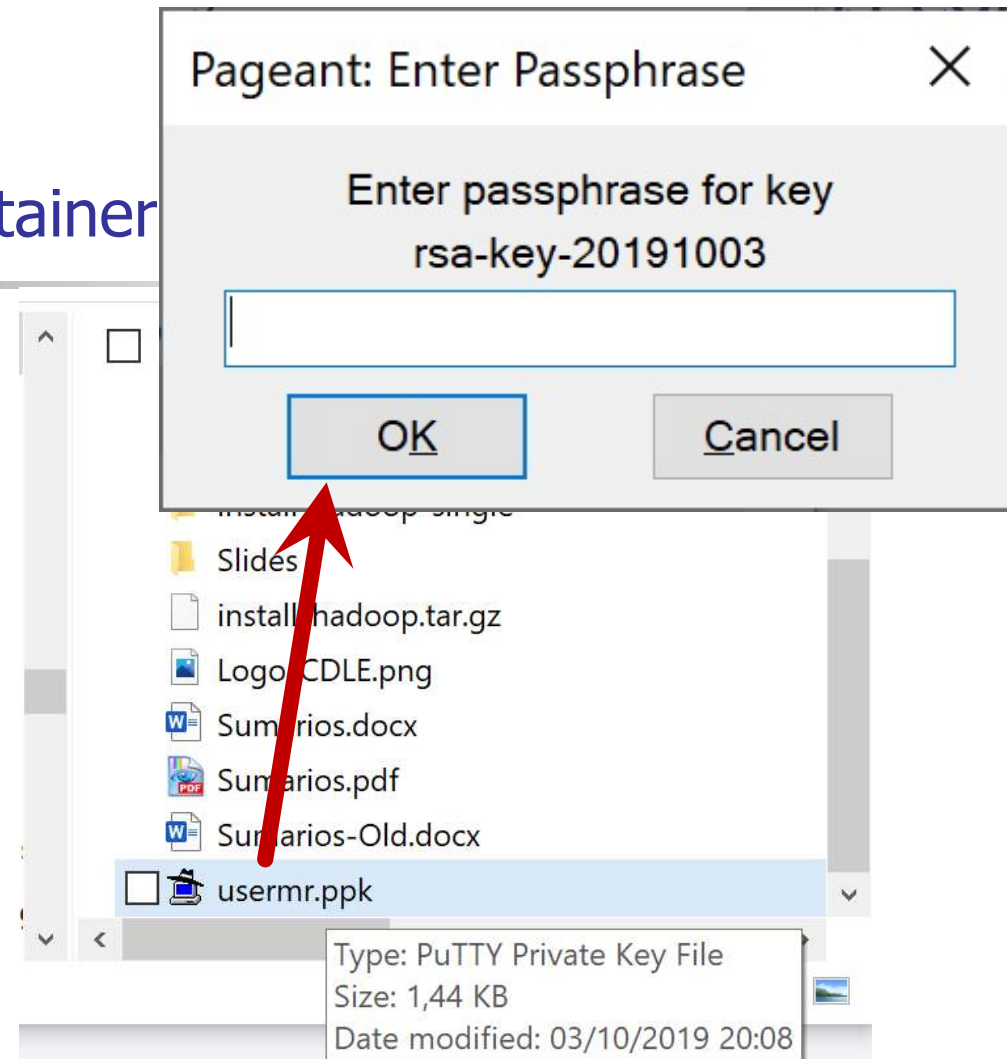
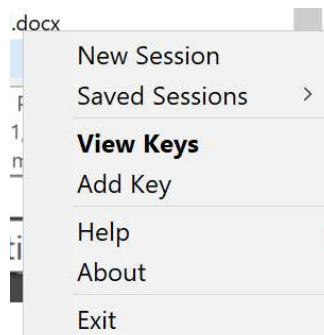
Existing containers

Created container

Configuration

Install Hadoop – Access container

- Start the PuTTY agent to handle the private keys
- Login in the container using PuTTY



Configuration

Install Hadoop – Access container

- The SSH server inside the container is available in port 22
- In the host this port is mapped onto port 222

```
C:\WINDOWS\system32\cmd.exe
docker run --hostname hadoop --name MR1 --detach -p 222:22 -p
9888 hadoop.ubuntu
```

Configuration

Install Hadoop – Access container

Server address

Server port

Session name

PuTTY Configuration

Category:

- Session
 - Logging
- Terminal
 - Keyboard
 - Bell
 - Features
- Window
 - Appearance
 - Behaviour
 - Translation
- + Selection
 - Colours
- Connection
 - Data
 - Proxy
 - Telnet
 - Rlogin
 - SSH
 - Kex
 - Host keys
 - Cipher
- + Auth
 - TTY
 - X11

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address) Port

Connection type:
☐ Raw ☐ Telnet ☐ Rlogin ☒ SSH ☐ Serial

Load, save or delete a stored session

Saved Sessions

INCD
M5
MQTT
MQTT-Nat
POLITECID-ECONET
hadoop
hadoop-docker

Load Save Delete

Close window on exit:
☐ Always ☐ Never ☒ Only on clean exit

About Help Open Cancel

Run

Configuration

Install Hadoop – Access container

Auto login user
name

PuTTY Configuration

Category:

- Session
 - Logging
- Terminal
 - Keyboard
 - Bell
 - Features
- Window
 - Appearance
 - Behaviour
 - Translation
 - + Selection
 - Colours
- Connection
 - Data
 - Proxy
 - Telnet
 - Rlogin
 - SSH
 - Kex
 - Host keys
 - Cipher
 - + Auth
 - TTY
 - X11

Data to send to the server

Login details

Auto-login username

When username is not specified:
☒ Prompt ☐ Use system username (cgonc)

Terminal details

Terminal-type string

Terminal speeds

Environment variables

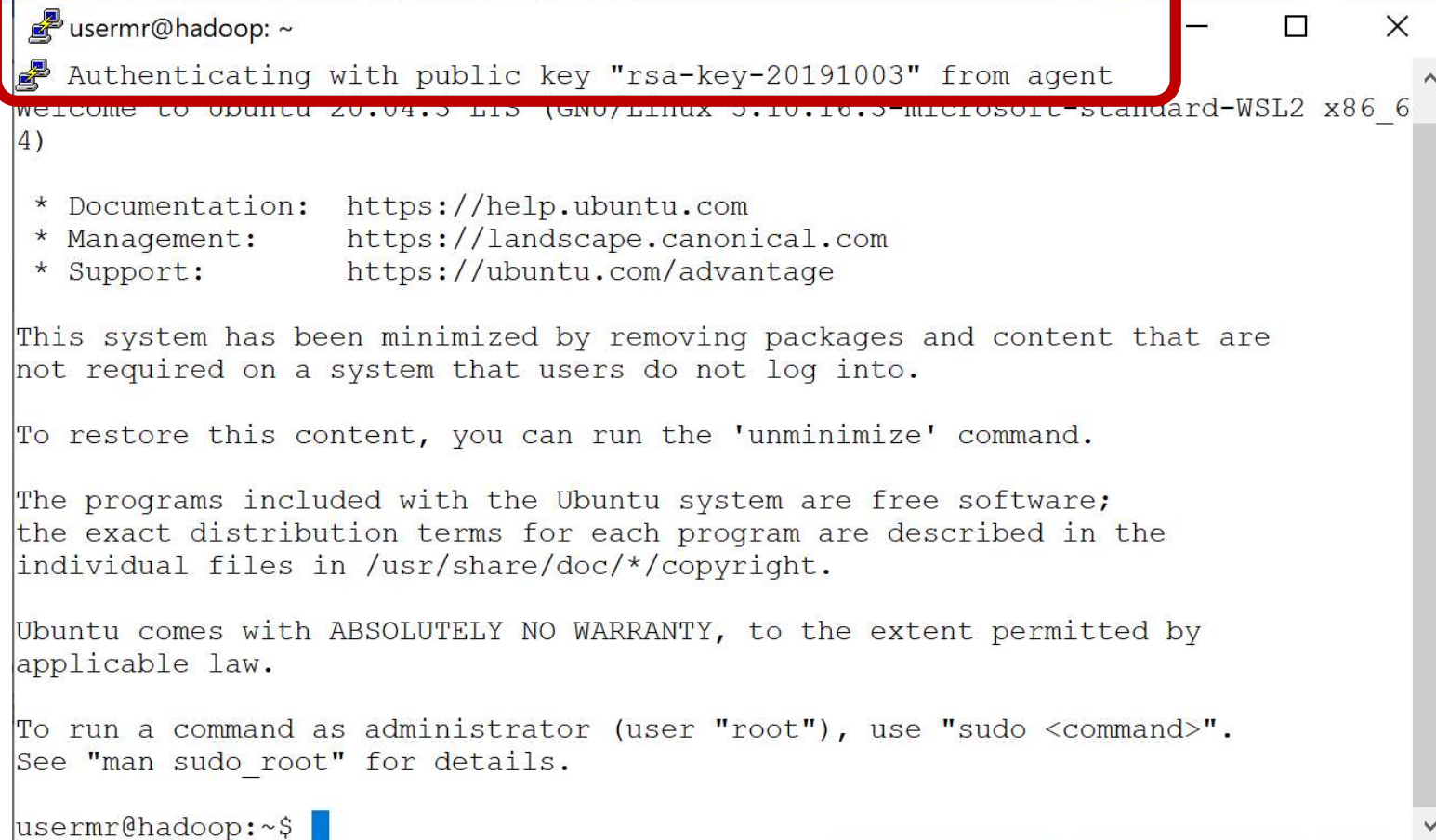
Variable		Add
Value	<input type="text"/>	Remove

Run

About Help Open Cancel

Configuration

Install Hadoop – Access container

A terminal window titled 'usermr@hadoop: ~' is shown. The window has a red border. The terminal output shows the user logging in, followed by a welcome message and system information. The text is as follows:

```
usermr@hadoop: ~  
Authenticating with public key "rsa-key-20191003" from agent  
Welcome to Ubuntu 20.04.3 LTS (GNU/Linux 5.10.16.3-microsoft-standard-WSL2 x86_64)  
  
 * Documentation:  https://help.ubuntu.com  
 * Management:    https://landscape.canonical.com  
 * Support:       https://ubuntu.com/advantage  
  
This system has been minimized by removing packages and content that are  
not required on a system that users do not log into.  
  
To restore this content, you can run the 'unminimize' command.  
  
The programs included with the Ubuntu system are free software;  
the exact distribution terms for each program are described in the  
individual files in /usr/share/doc/*/copyright.  
  
Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by  
applicable law.  
  
To run a command as administrator (user "root"), use "sudo <command>".  
See "man sudo_root" for details.  
  
usermr@hadoop:~$
```

Configuration

Install Hadoop

- Steps sequence

1. Start a SSH session
 - Execute script `"00-a-java-install.sh"`
2. Start a SSH session
 - Execute script `"00-b-ant-install.sh"`
3. Start a new session
 - Execute script `"00-c-maven-install.sh"`
4. Start a new session
 - Execute script `"00-d-ssh-env.sh"`
5. Start a new session
 - Execute script `"installHadoop.sh"`
6. Start a new session
 - Execute script `"11-hadoop-InitUser.sh usermr"`

Configuration

Install Hadoop – Verify installation

```
usermr@hadoop: ~  
doop/bin/hadoop fs -chmod -R 0777 /user/history""  
sshpass -p hdfs ssh -o StrictHostKeyChecking=no hdfs@localhost ""/work/hadoop/ha  
doop/bin/hadoop fs -chown hadoop:hadoop /user/history""
```

Web UI is available at:

For the HDFS service (Name node):

<http://localhost:9870/>

For the HDFS service (Secondary name node):

<http://localhost:9868/>

For the HDFS service (Data nodes):

<http://localhost:9864/>

For the MapReduce service:

<http://localhost:19888/>

For the YARN service (Resource Manager):

<http://localhost:8088/>

For the YARN service (Nodes Manager):

<http://localhost:8042/>

usermr@hadoop:~\$

Configuration

Install Hadoop – Verify installation

http://localhost:9870/

Namenode information

localhost:9870/dfshealth.html#tab-overview

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Overview 'localhost:8020' (✓active)

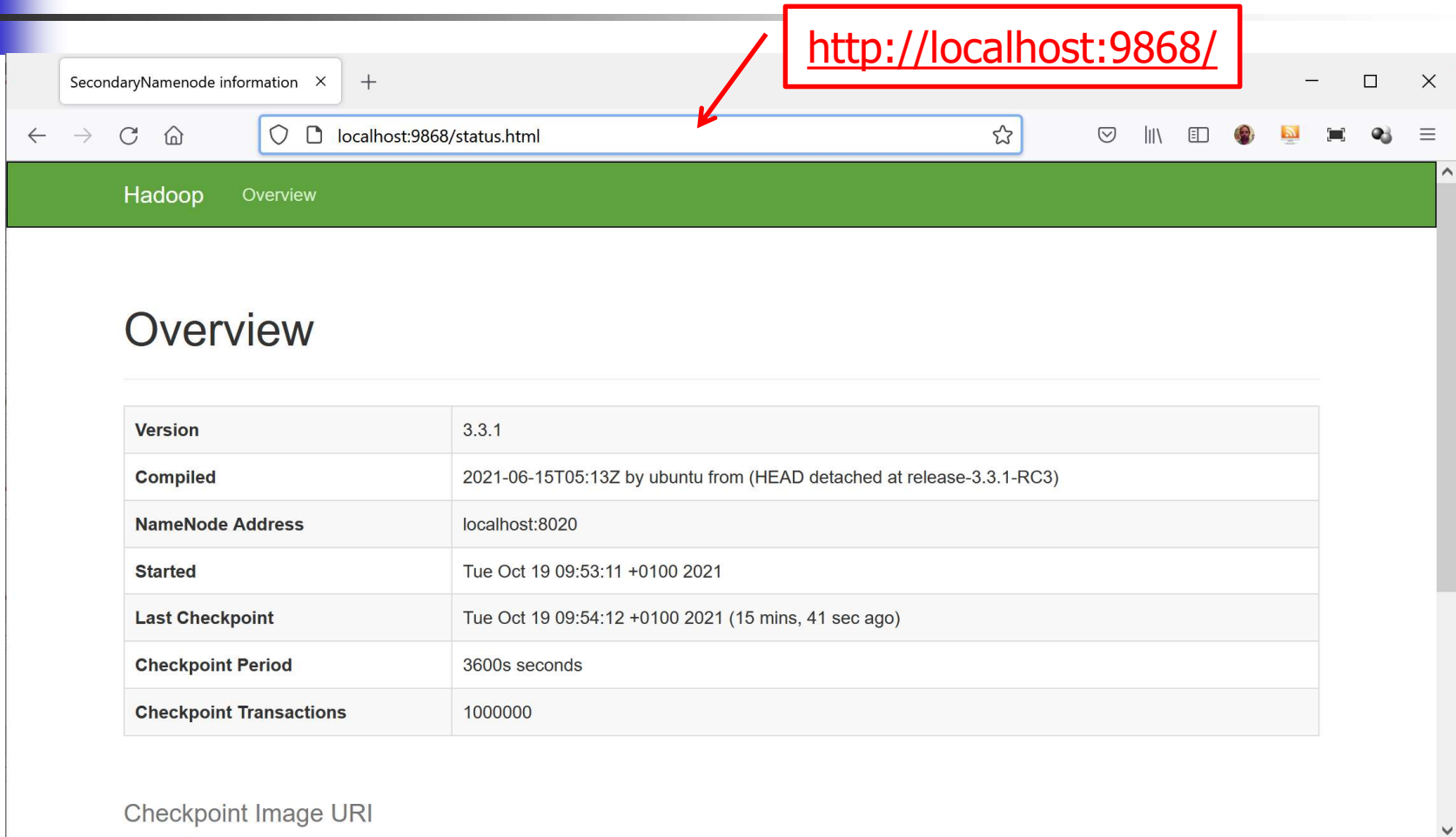
Started:	Tue Oct 19 09:53:07 +0100 2021
Version:	3.3.1, ra3b9c37a397ad4188041dd80621bdeefc46885f2
Compiled:	Tue Jun 15 06:13:00 +0100 2021 by ubuntu from (HEAD detached at release-3.3.1-RC3)
Cluster ID:	CID-c361f099-9c84-4af6-a55f-730f703c79e1
Block Pool ID:	BP-2048766310-172.17.0.2-1634633581359

Summary

Security is off.

Configuration

Install Hadoop – Verify installation



SecondaryNameNode information x +

localhost:9868/status.html

Hadoop Overview

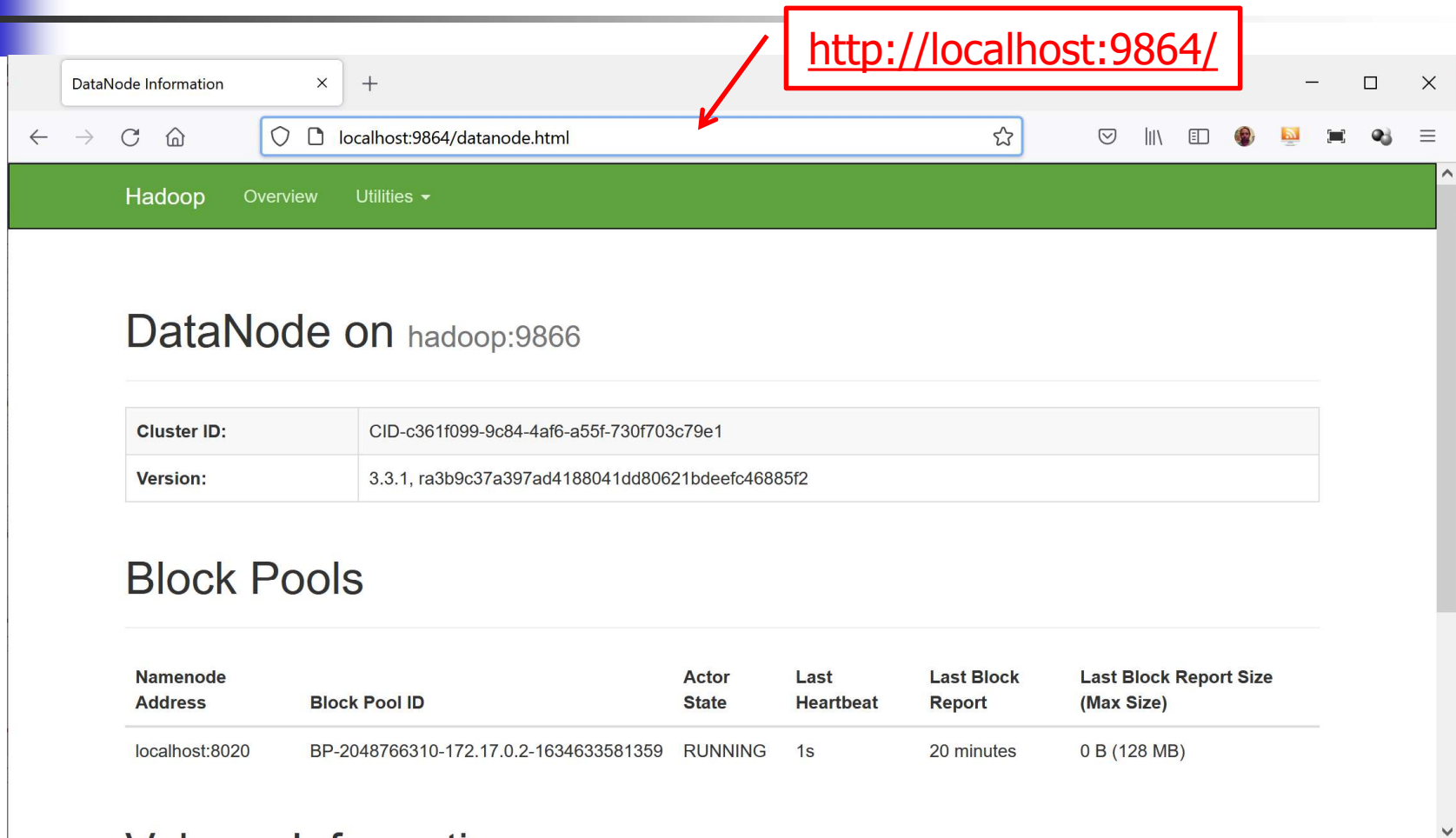
Overview

Version	3.3.1
Compiled	2021-06-15T05:13Z by ubuntu from (HEAD detached at release-3.3.1-RC3)
NameNode Address	localhost:8020
Started	Tue Oct 19 09:53:11 +0100 2021
Last Checkpoint	Tue Oct 19 09:54:12 +0100 2021 (15 mins, 41 sec ago)
Checkpoint Period	3600s seconds
Checkpoint Transactions	1000000

Checkpoint Image URI

Configuration

Install Hadoop – Verify installation



<http://localhost:9864/>

localhost:9864/datanode.html

Hadoop Overview Utilities

DataNode on hadoop:9866

Cluster ID:	CID-c361f099-9c84-4af6-a55f-730f703c79e1
Version:	3.3.1, ra3b9c37a397ad4188041dd80621bdeefc46885f2

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
localhost:8020	BP-2048766310-172.17.0.2-1634633581359	RUNNING	1s	20 minutes	0 B (128 MB)


Configuration

Install Hadoop – Verify installation

<http://localhost:19888/>

JobHistory

localhost:19888/jobhistory

 **JobHistory**

Application

About
Jobs

Tools

Retired Jobs

Show 20 entries

Search:

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed
No data available in table											

Showing 0 to 0 of 0 entries

First Previous

Configuration

Install Hadoop – Verify installation

<http://localhost:8088/>

The screenshot displays the Hadoop web interface. The browser address bar shows <http://localhost:8088/>. The interface includes a sidebar with navigation links and a main content area with various metrics and tables.

Cluster Metrics

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

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimize Resource Usage
Capacity Scheduler	[memory-mb (unit=Mb), vcores]	<memory:1024, vCores:1>

Showing 0 to 0 of 0 entries

Testing

Install examples

- Start a SSH session
- Upload the examples and maintain the directory structure suggested in Moodle

```
usermr@hadoop: ~  
usermr@hadoop:~$ dir  
total 53M  
drwxr-xr-x 1 usermr usermr 4.0K Oct 19 09:30 .  
drwxr-xr-x 1 root   root   4.0K Oct 19 08:52 ..  
-rw-r--r-- 1 usermr usermr  33 Oct 16 12:02 .bash_aliases  
-rw----- 1 usermr usermr 136 Oct 19 09:16 .bash_history  
-rw-r--r-- 1 usermr usermr 220 Feb 25 2020 .bash_logout  
-rw-r--r-- 1 usermr usermr 3.7K Feb 25 2020 .bashrc  
drwx----- 2 usermr usermr 4.0K Oct 19 08:31 .cache  
-rw-r--r-- 1 usermr usermr 844 Oct 17 09:55 .profile  
drwxr-xr-x 1 usermr usermr 4.0K Oct 19 08:52 .ssh  
drwxrwxr-x 2 usermr usermr 4.0K Mar 16 2021 bin  
-rw-rw-r-- 1 usermr usermr 49M Oct 19 09:28 examples.zip  
-rw-rw-r-- 1 usermr usermr 221K Oct 19 09:28 gutenber.zip  
drwxrwxr-x 2 usermr usermr 4.0K Oct 14 18:40 install-hadoop  
-rwxr-xr-x 1 usermr usermr 179 Oct 16 11:52 setupSSH.sh  
-rw-rw-r-- 1 usermr usermr 1.5M Oct 19 09:28 temperatures.zip  
-rw-rw-r-- 1 usermr usermr 1.8M Oct 19 09:28 wikipedia.zip  
usermr@hadoop:~$
```

Testing

Install examples

```
usermr@hadoop: ~/examples
usermr@hadoop:~/examples$ tree -d -L 2
.
|-- Demos
|-- Projects
|   |-- 01-Temperatures
|   |-- 02-WordCount
|   |-- 03-FileSystem
|   |-- 04-Streams
|   |-- 05-Configuration
|   |-- 06-MapReduce
|   `-- 07-OpenCV
-- conf
-- input
|   |-- gutenber
|   |-- temperatures
|   `-- wikipedia
-- output
|   |-- imagens
|   `-- videos

17 directories
usermr@hadoop:~/examples$
```

Examples

Word count examples

Directory with input data files

Directory for output data files

Testing Install examples

- On a SSH session execute the commands:
 - `cd`
 - `sudo chown -R usermr:hadoop examples/`
 - `sudo chmod -R o-w examples/`
 - `cd /home`
 - `sudo chown usermr:hadoop /home/usermr`

Testing

Build examples

- `cd ~/examples/Projects/`
- `./build.sh`

```
usermr@hadoop: ~/examples/Projects
[INFO] Ex20-FileDecompressor-02 ..... SUCCESS [ 0.882 s]
[INFO] Ex21-PooledStreamCompressor ..... SUCCESS [ 0.936 s]
[INFO] 05-Configuration ..... SUCCESS [ 0.007 s]
[INFO] Ex22-ReadConfiguration-01 ..... SUCCESS [ 1.143 s]
[INFO] Ex23-ReadConfiguration-02 ..... SUCCESS [ 0.950 s]
[INFO] Ex24-ReadConfiguration-03 ..... SUCCESS [ 0.856 s]
[INFO] Ex25-ConfigurationPrinter ..... SUCCESS [ 0.844 s]
[INFO] 06-MapReduce ..... SUCCESS [ 0.008 s]
[INFO] Ex26-MapReduce-01 ..... SUCCESS [ 0.805 s]
[INFO] Ex27-MapReduce-02 ..... SUCCESS [ 0.915 s]
[INFO] 07-OpenCV ..... SUCCESS [ 0.010 s]
[INFO] Utils-OpenCV ..... SUCCESS [ 0.583 s]
[INFO] Demo01-OpenCV-ExtractFramesFromVideo ..... SUCCESS [ 14.916 s]
[INFO] Demo02-OpenCV-IdentifyObjectsInPictures ..... SUCCESS [ 0.600 s]
[INFO] -----
[INFO] BUILD SUCCESS
[INFO] -----
[INFO] Total time: 01:24 min
[INFO] Finished at: 2021-10-19T09:46:34Z
[INFO] -----
usermr@hadoop:~/examples/Projects$
```

Testing

Run Word Count example

- `cd ~/examples/Projects/02-WordCount/Ex10-WordCount-01/`

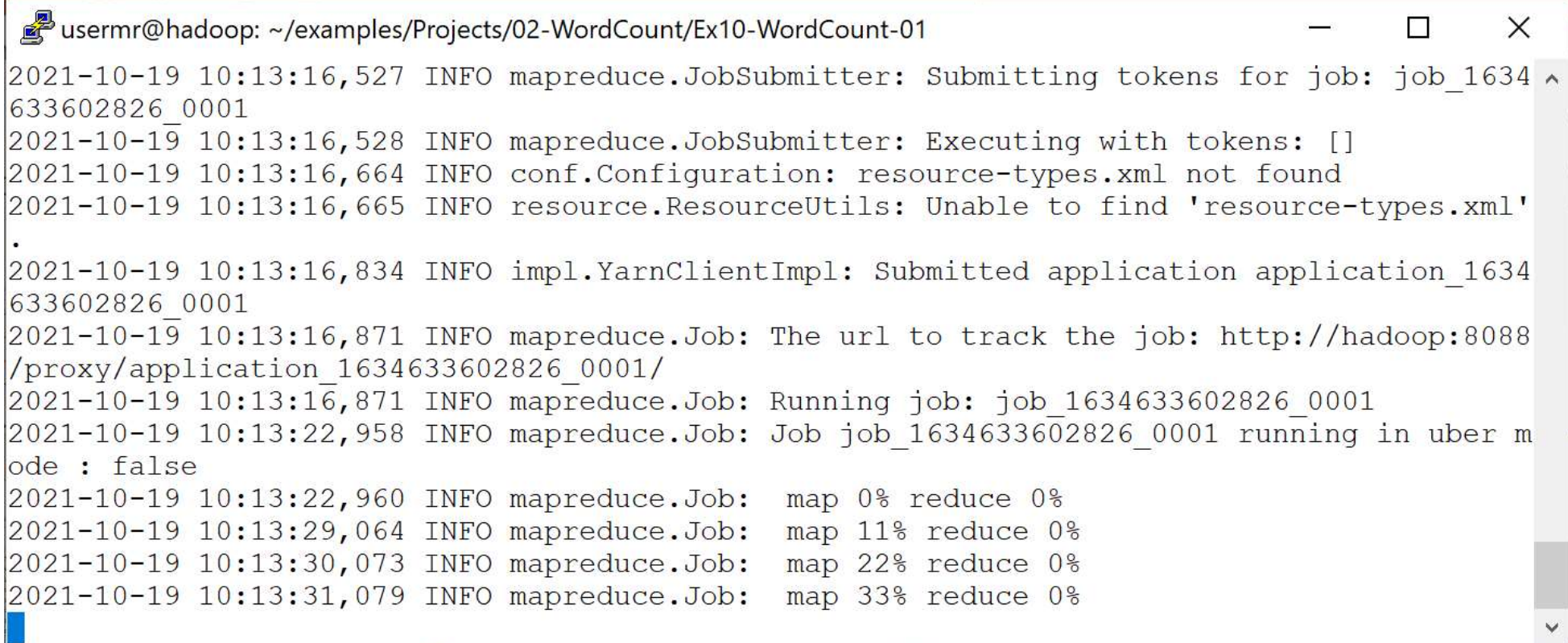
```
usermr@hadoop: ~/examples/Projects/02-WordCount/Ex10-WordCount-01
usermr@hadoop:~/examples/Projects/02-WordCount/Ex10-WordCount-01$ ./run.sh
Invalid arguments!
Usage:
./run.sh <File System type>

Where <File System type> can be:

local - local file system (file://)
HDFS - HDFS file system (hdfs://)
usermr@hadoop:~/examples/Projects/02-WordCount/Ex10-WordCount-01$
```

Testing

Run Word Count example – Local file system



```
usermr@hadoop: ~/examples/Projects/02-WordCount/Ex10-WordCount-01
2021-10-19 10:13:16,527 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634633602826_0001
2021-10-19 10:13:16,528 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-19 10:13:16,664 INFO conf.Configuration: resource-types.xml not found
2021-10-19 10:13:16,665 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'
.
2021-10-19 10:13:16,834 INFO impl.YarnClientImpl: Submitted application application_1634633602826_0001
2021-10-19 10:13:16,871 INFO mapreduce.Job: The url to track the job: http://hadoop:8088/proxy/application_1634633602826_0001/
2021-10-19 10:13:16,871 INFO mapreduce.Job: Running job: job_1634633602826_0001
2021-10-19 10:13:22,958 INFO mapreduce.Job: Job job_1634633602826_0001 running in uber mode : false
2021-10-19 10:13:22,960 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-19 10:13:29,064 INFO mapreduce.Job:  map 11% reduce 0%
2021-10-19 10:13:30,073 INFO mapreduce.Job:  map 22% reduce 0%
2021-10-19 10:13:31,079 INFO mapreduce.Job:  map 33% reduce 0%
```

Testing

Run Word Count example – Local file system – Results

```
usermr@hadoop: ~/examples/Projects/02-WordCount/Ex10-WordCount-01
Result sorted by key - MapReduce defaults - (first 5 lines)
hadoop fs -text file:///home/usermr/examples/output/gutenberg/mixed/part-r-00001 2>/dev/
null | head -n 5
"Defects".      8
"Defects,"      2
"House" 2
"Information    2
"Plain 4

Result sorted (by value) using the linux sort command
hadoop fs -text file:///home/usermr/examples/output/gutenberg/mixed/part-r-00001 2>/dev/
null | sort -k 2,2 -n -r | head -n 5
shall 470
this 402
is 334
any 330
as 258
usermr@hadoop:~/examples/Projects/02-WordCount/Ex10-WordCount-01$
```


Testing

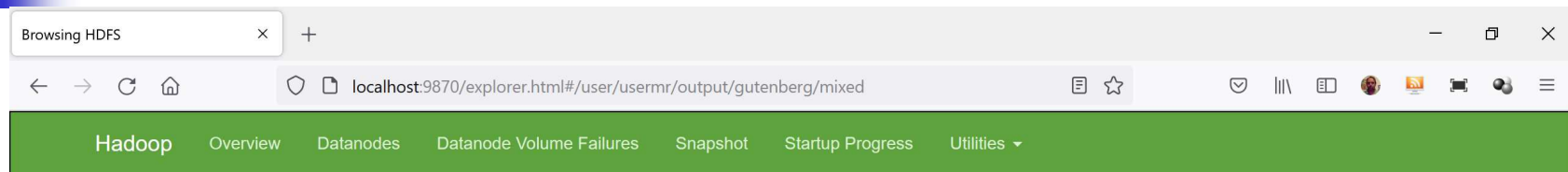
Run Word Count example – HDFS file system – Results

```
usermr@hadoop: ~/examples/Projects/02-WordCount/Ex10-WordCount-01
Result sorted by key - MapReduce defaults - (first 5 lines)
hadoop fs -text hdfs:///user/usermr/output/gutenberg/mixed/part-r-00001 2>/dev/null | head -n 5
"Defects".      8
"Defects,"      2
"House" 2
"Information"   2
"Plain" 4

Result sorted (by value) using the linux sort command
hadoop fs -text hdfs:///user/usermr/output/gutenberg/mixed/part-r-00001 2>/dev/null | sort -k 2,2 -n -r | head -n 5
shall 470
this 402
is 334
any 330
as 258
usermr@hadoop:~/examples/Projects/02-WordCount/Ex10-WordCount-01$
```

Testing

Run Word Count example – HDFS file system – Results



Browse Directory

/user/usermr/output/gutenberg/mixed

Show entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	usermr	hadoop	0 B	Oct 19 11:16	1	128 MB	_SUCCESS	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	usermr	hadoop	27.52 KB	Oct 19 11:16	1	128 MB	part-r-00000	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	usermr	hadoop	26.2 KB	Oct 19 11:16	1	128 MB	part-r-00001	<input type="checkbox"/>

Showing 1 to 3 of 3 entries

Hadoop, 2021.

Testing


Run Word Count example – Jobs history

JobHistory

localhost:19888/jobhistory

80%

Logged in as: dr.who

 **JobHistory**

Application

About Jobs

Tools

Retired Jobs

Show 20 entries

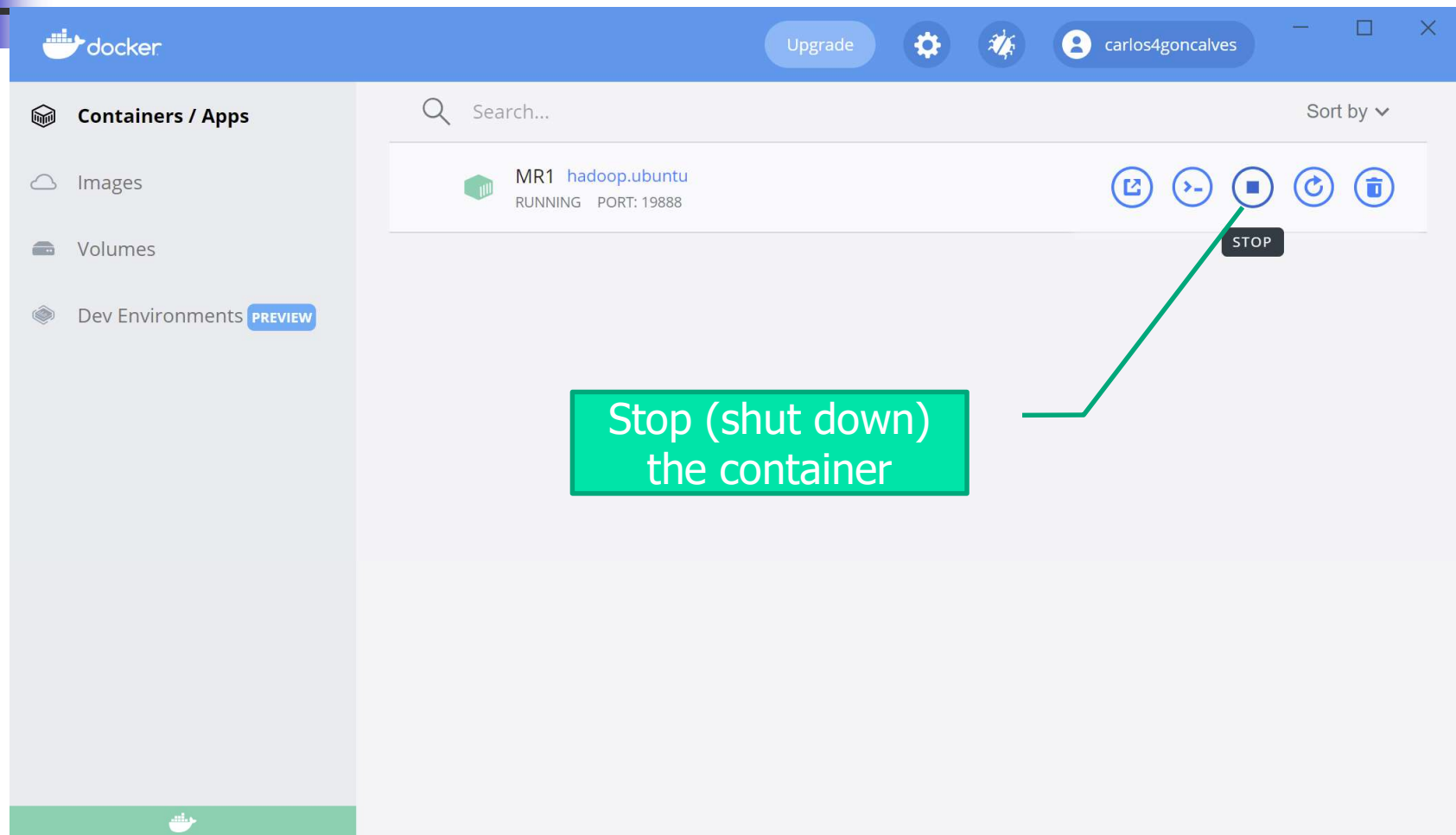
Search:

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed	Elapsed Time
2021.10.19 10:16:40 GMT	2021.10.19 10:16:44 GMT	2021.10.19 10:16:58 GMT	job_1634633602826_0002	Word Count Ver 1	usermr	default	SUCCEEDED	9	9	2	2	00hrs, 00mins, 13sec
2021.10.19 10:13:16 GMT	2021.10.19 10:13:21 GMT	2021.10.19 10:13:37 GMT	job_1634633602826_0001	Word Count Ver 1	usermr	default	SUCCEEDED	9	9	2	2	00hrs, 00mins, 15sec

Showing 1 to 2 of 2 entries

First Previous 1 Next Last

Shut down the container





Bring up the container

- On the next login Hadoop components aren't available!
- It is necessary to explicitly bring up the Hadoop components
- The script `"07-hadoopPseudoDistributed-Start.sh"` brings up all the Hadoop components

Bring up the container