

CentOS Stream 8 – Hadoop 3

Version:	1.0.0
Created by:	cloudimg

Table of Contents

1.) Overview.....	1
2.) Access & Security.....	1
3.) System Requirements.....	2
4.) Connecting to the Instance.....	2
5.) On Startup.....	2
6.) Filesystem Configuration.....	3
7.) Server Components.....	3
8.) Scripts and Log Files.....	4
9.) Using System Components.....	4

1.) Overview

This document is provided as a user guide for the CentOS Stream 8 – Hadoop 3 product offering on the Azure Marketplace. Please reach out to support@cloudimg.co.uk if any issues are encountered following this user guide for the chosen product offering.

2.) Access & Security

Please update the security group of the target instance to allow the below ports and protocols for access and connectivity.



Registered
Technology
Partner

cloudimg
(+44) 02045382725
3rd Floor 86-90 Paul Street London EC2A 4NE
support@cloudimg.co.uk
<https://cloudimg.co.uk>

Protocol	Type	Port	Description
SSH	TCP	22	SSH connectivity
TCP	TCP	8088	Hadoop Resource Manager UI
TCP	TCP	8042	Hadoop Node Manager UI

3.) System Requirements

The minimum system requirements for the chosen product offering can be found below

Minimum CPU	Minimum RAM	Required Disk Space
1	1 GB	20 GB

4.) Connecting to the Instance

Once launched in the Azure Virtual Machines Service, please connect to the instance via an SSH client using the **azureuser** with the key pair associated at launch. Once connected as the **azureuser**, you will be able to **sudo** to the **root** user by issuing the below command.

Switch to the root user

```
sudo su -
```

5.) On Startup

An OS package update script has been configured to run on boot to ensure the image is fully up to date at first use. You can disable this feature by removing the script from /stage/scripts/ and deleting the entry in crontab for the root user.

Disable the OS update script from running on reboot

```
rm -f /stage/scripts/initial_boot_update.sh
crontab -e
#DELETE THE BELOW LINE. SAVE AND EXIT THE FILE.
@reboot /stage/scripts/initial_boot_update.sh
```



Registered
Technology
Partner

cloudimg
(+44) 02045382725
3rd Floor 86-90 Paul Street London EC2A 4NE
support@cloudimg.co.uk
<https://cloudimg.co.uk>

6.) Filesystem Configuration

Please see below for a screenshot of the server disk configuration and specific mount point mappings for software locations.

```

Filesystem      Size  Used Avail Use% Mounted on
devtmpfs        1.9G   0  1.9G   0% /dev
tmpfs           2.0G   0  2.0G   0% /dev/shm
tmpfs           2.0G  8.6M  1.9G   1% /run
tmpfs           2.0G   0  2.0G   0% /sys/fs/cgroup
/dev/nvme0n1p2  38G  2.5G  33G   7% /
/dev/nvme0n1p1  2.0G 121M  1.7G   7% /boot
tmpfs           391M   0  391M   0% /run/user/1002
/dev/nvme1n1    9.8G  3.2G  6.1G  35% /apps
tmpfs           391M   0  391M   0% /run/user/1004
  
```

Mount Point	Description
/boot	Operating System Kernel files
/apps	Big Data components installation directory

7.) Server Components

Please see below for a list of installed server components and their respective installation paths. The below versions are subject to change on initial boot based on the initial_boot_update.sh script finding new versions of the software in the systems package repositories.

Component	Version	Software Home
Cloud-Init	22.1-1	/etc/cloud
Java	1.8	/apps/java
Hadoop	3.3.4	/apps/hadoop
Apache Hive	3.1.3	/apps/apache-hive
Apache HBase	2.4.15	/apps/apache-hbase
Apache Pig	0.17	/apps/apache-pig
Apache Spark	3.3.2	/apps/apache-spark
Apache Zookeeper	3.7.1	/apps/apache-zookeeper
Azure CLI	2.53.1	/lib64/az



Registered
Technology
Partner

cloudimg
(+44) 02045382725
3rd Floor 86-90 Paul Street London EC2A 4NE
support@cloudimg.co.uk
<https://cloudimg.co.uk>

8.) Scripts and Log Files

The below table provides a breakdown of any scripts & log files created to enhance the useability of the chosen offering.

Script/Log	Path	Description
Initial_boot_update.sh	/stage/scripts	Update the Operating System with the latest updates available.
Initial_boot_update.log	/stage/scripts	Provides output for initial_boot_update.sh
start_all.sh	/home/hadoop	Start all Hadoop services (Single Node)
stop_all.sh	/home/hadoop	Stop all Hadoop services (Single Node)
setup_ssh_hadoop_user.sh	/stage/scripts	Configures Hadoop user SSH key

9.) Using System Components

Instructions can be found below for using each component of the server build mentioned in section 7 of this user guide document.

Azure CLI

Using Azure CLI - as any OS user.

```
az
```

Cloud-Init

Edit the /etc/cloud/cloud.cfg file to reflect your desired configuration. A link to the cloud-init official documentation can be found below for referencing best practise for your use case.

<https://cloudinit.readthedocs.io/en/latest/>

```
vi /etc/cloud/cloud.cfg
```

Java



Registered
Technology
Partner

cloudimg
(+44) 02045382725
3rd Floor 86-90 Paul Street London EC2A 4NE
support@cloudimg.co.uk
<https://cloudimg.co.uk>

Java has been preinstalled on the instance and the below command can be used to verify the version currently installed.

```
java -version
```

Hadoop

Hadoop 3x has been preinstalled on the system. The installation location of hadoop can be found under the directory `/apps/hadoop`

Before starting the Hadoop service, please run the below script as the **root** user.

```
./stage/scripts/setup_ssh_hadoop_user.sh
```

Hadoop requires an SSH key for secure communication between the different nodes in a Hadoop cluster. When you run a Hadoop job, the Hadoop components (e.g., NameNode, DataNode, TaskTracker, and JobTracker) need to communicate with each other to distribute the data and execute the job. This communication can involve sensitive information, such as file locations, job status, and authentication credentials.

You can stop or start all hadoop services as the hadoop OS user by running the below commands.

```
#START ALL HADOOP SERVICES
sudo su - hadoop
cd $HOME
. ./start-all.sh

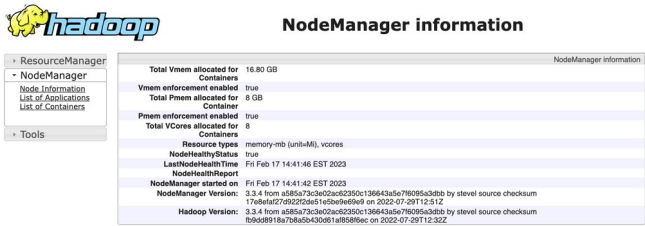
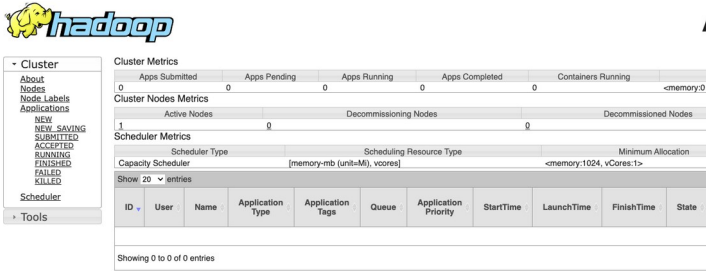
#STOP ALL HADOOP SERVICES
sudo su - hadoop
cd $HOME
. ./stop-all.sh
```



Registered
Technology
Partner

cloudimg
(+44) 02045382725
3rd Floor 86-90 Paul Street London EC2A 4NE
support@cloudimg.co.uk
<https://cloudimg.co.uk>

Once all of the services have been started, you can access the below URLs exchanging the values between <> with that of your own instance.

URL Purpose	Description
NodeManager UI	<p>This is the web interface for each NodeManager and can be accessed using the URL <a href="http://<PUBLIC/PRIVATEIP>:8042">http://<PUBLIC/PRIVATEIP>:8042. This page provides information about the node, including the resource usage and logs.</p> 
ResourceManager UI	<p>This is the web interface for the ResourceManager and can be accessed using the URL <a href="http://<PUBLIC/PRIVATEIP>:8088">http://<PUBLIC/PRIVATEIP>:8088. This page provides information about the YARN cluster, including the status of nodes, application statistics, and more.</p> 

Apache Hive - Please type the below command without copy and paste.

Apache Hive has been preinstalled on the system. You can run the below command to verify the version installed. To configure the system based on your requirements you can edit the config files located under the directory /apps/apache-hive/conf

```
hive --version
```

Apache HBase - Please type the below command without copy and paste.

Apache HBase has been preinstalled on the system. You can run the below commands to verify the version installed. To configure the system based on your requirements you can edit the config files located under the directory /apps/apache-hbase/conf

```
hbase --version
```



Registered
Technology
Partner

cloudimg
(+44) 02045382725
3rd Floor 86-90 Paul Street London EC2A 4NE
support@cloudimg.co.uk
<https://cloudimg.co.uk>

```
hbase shell
```

Apache Pig - Please type the below command without copy and paste.

Apache Pig has been preinstalled on the system. You can run the below command to verify the version installed. To configure the system based on your requirements you can edit the config files located under the directory /apps/apache-pig/conf

```
pig -version
```

Apache Spark - Please type the below command without copy and paste.

Apache Spark has been preinstalled on the system. You can run the below command to verify the version installed. To configure the system based on your requirements you can edit the config files located under the directory /apps/apache-spark/conf

```
spark-submit --version
```

Apache Zookeeper - Please type the below command without copy and paste.

Apache Zookeeper has been preinstalled on the system. You can run the below command to verify the version installed. To configure the system based on your requirements you can edit the config files located under the directory /apps/apache-zookeeper/conf

```
zkServer.sh version
```



Registered
Technology
Partner

cloudimg
(+44) 02045382725
3rd Floor 86-90 Paul Street London EC2A 4NE
support@cloudimg.co.uk
<https://cloudimg.co.uk>