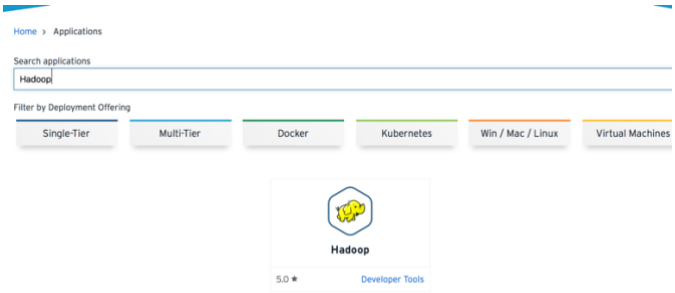
WEEK 5 ASSIGNMENT

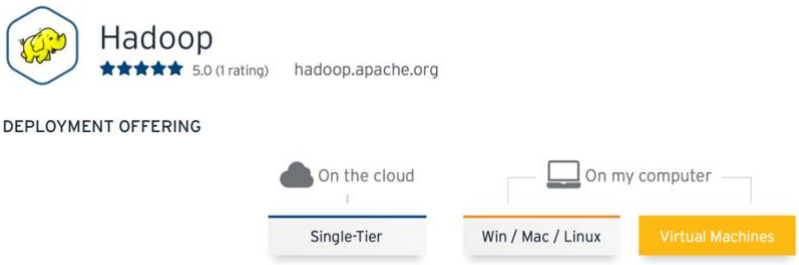
Large-Scale Data Storage Systems – DATA-5400 | Spring 2020

Christina Morgenstern

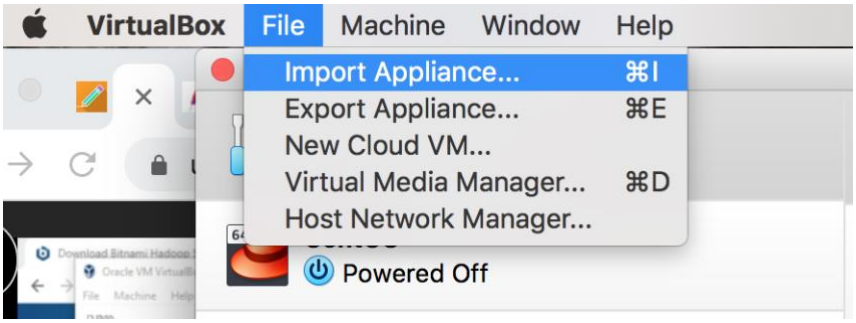
For this week´s assignment, I have chosen the hands-on option because I want to get experience with Hadoop. As for the choice of option, I started with Option 1 of using Hadoop on my VirtualBox.
To install Hadoop on the VM, I downloaded the Hadoop VirtualBox image from Bitnami. Bitnami offers easy to configure software apps for download, mostly open source and free (https://bitnami.com). On the bitnami website, go to Community and search for Hadoop.



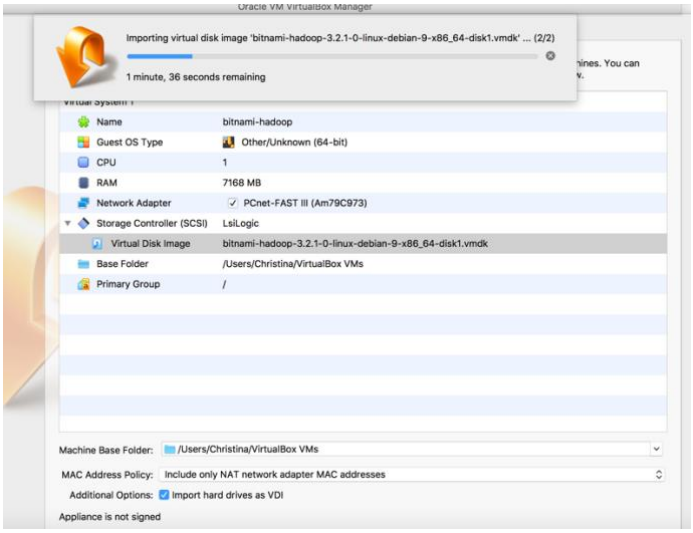
Choose to download the version for Virtual Machines.



On my Mac, I opened VirtualBox and imported the Hadoop VM image.

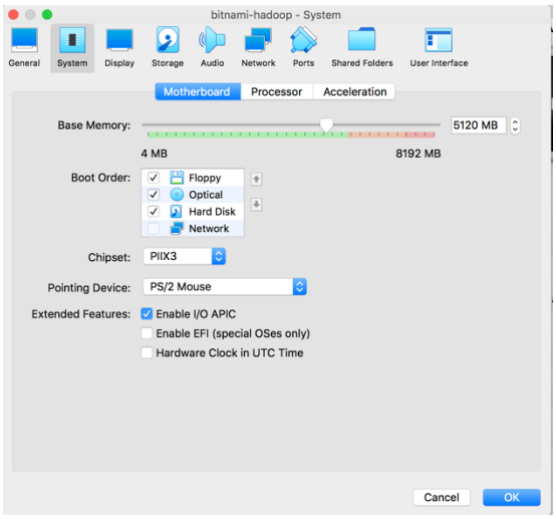


Choose the file previously downloaded for import.

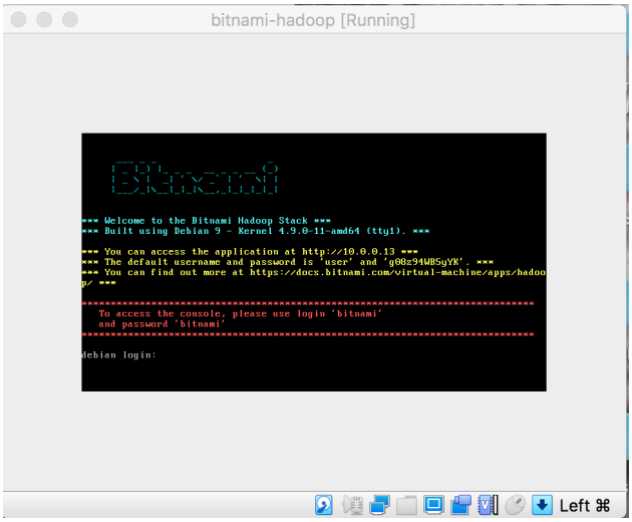


Bitnami Hadoop was successfully installed.

In the settings options, amend the memory. I couldn´t choose 8GB of memory because of already too little memory. So, I chose 5120 MB and 1 CPU as suggested in the lecture.



Start bitnami-hadoop and log on using bitnami as login and password.



Disable the firewall.

```
bitnami@debian:~$ sudo -i -u root
root@debian:~# systemctl disable ufw
Synchronizing state of ufw.service with SysV service script with /lib/systemd/sy
stemd-sysv-install.
Executing: /lib/systemd/systemd-sysv-install disable ufw
root@debian:~# exit
logout
bitnami@debian:~$ _
```

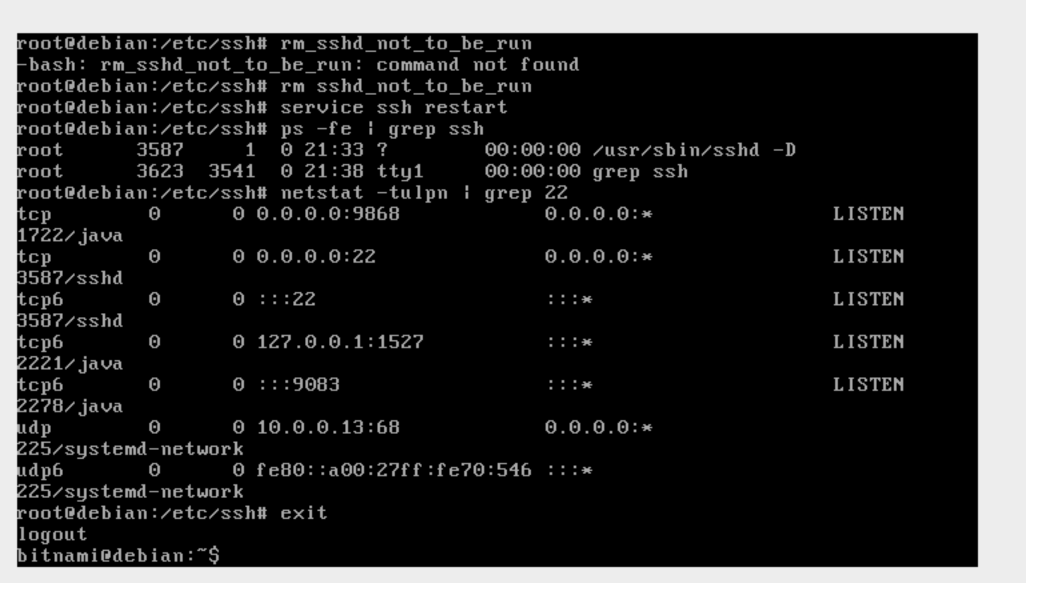
SSH via Putty, MobaXterm or Mac is disabled by default in Hadoop. The following changes were made in order to allow SSH.

```
bitnami@debian:~$ sudo -i -u root
root@debian:~# cd/etc/ssh
-bash: cd/etc/ssh: No such file or directory
root@debian:~# cd/etc/ssh
-bash: cd/etc/ssh: No such file or directory
root@debian:~# cd /etc
root@debian:/etc# cd ssh
root@debian:/etc/ssh# ls
moduli              ssh_host_ecdsa_key        ssh_host_rsa_key
ssh_config          ssh_host_ecdsa_key.pub    ssh_host_rsa_key.pub
sshd_config         ssh_host_ed25519_key
sshd_not_to_be_run  ssh_host_ed25519_key.pub
root@debian:/etc/ssh#
```

Remove the file “sshd_not_to_be_run”

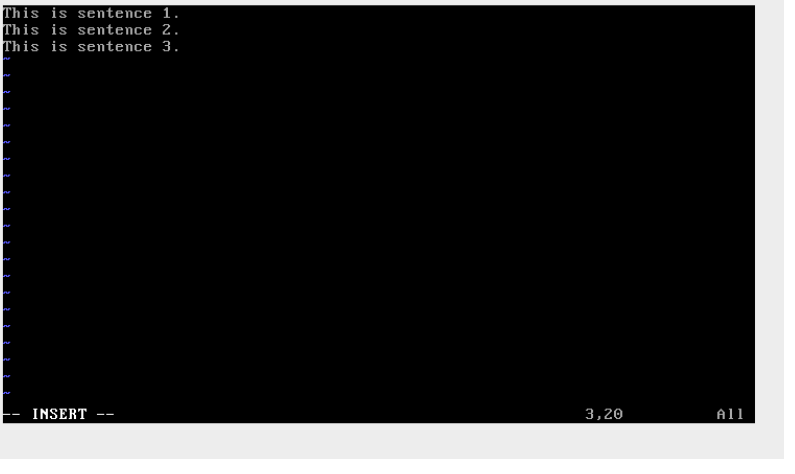
```
-bash: rm_sshd_not_to_be_run: command not found
root@debian:/etc/ssh# rm sshd_not_to_be_run
root@debian:/etc/ssh# _
```

Restart the SSH service and check if it is working.



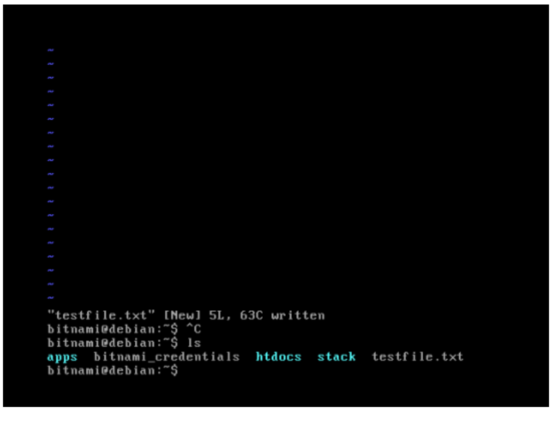
```
root@debian:/etc/ssh# rm_sshd_not_to_be_run
-bash: rm_sshd_not_to_be_run: command not found
root@debian:/etc/ssh# rm sshd_not_to_be_run
root@debian:/etc/ssh# service ssh restart
root@debian:/etc/ssh# ps -fe | grep ssh
root      3587     1  0 21:33 ?        00:00:00 /usr/sbin/sshd -D
root      3623  3541  0 21:38 tty1     00:00:00 grep ssh
root@debian:/etc/ssh# netstat -tulpn | grep 22
tcp        0      0 0.0.0.0:9868            0.0.0.0:*               LISTEN
1722/java
tcp        0      0 0.0.0.0:22              0.0.0.0:*               LISTEN
3587/sshd
tcp6       0      0 :::22                   :::*                    LISTEN
3587/sshd
tcp6       0      0 127.0.0.1:1527          :::*                    LISTEN
2221/java
tcp6       0      0 :::9083                 :::*                    LISTEN
2278/java
udp        0      0 10.0.0.13:68            0.0.0.0:*
225/systemd-network
udp6       0      0 fe80::a00:27ff:fe70:546 :::*
225/systemd-network
root@debian:/etc/ssh# exit
logout
bitnami@debian:~$
```

Create a text file (testfile.txt) using the vim Editor.



```
This is sentence 1.
This is sentence 2.
This is sentence 3.
-- INSERT --                                          3,20          All
```

Confirm that test file is there using the `ls` command.

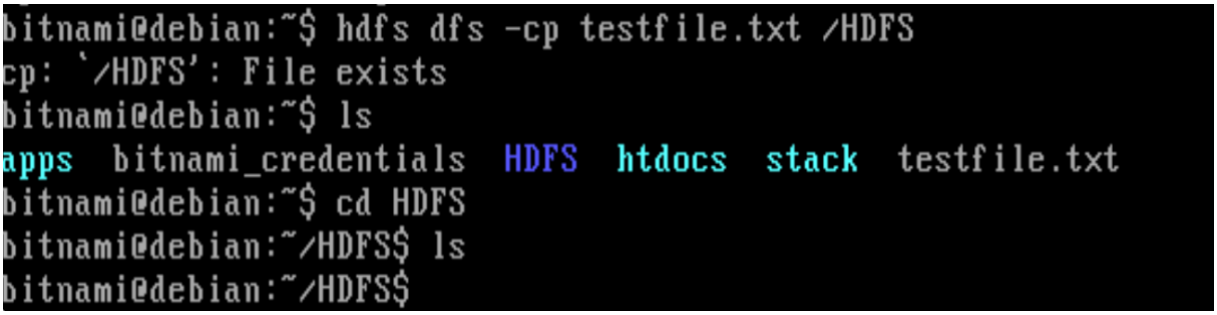


```
"testfile.txt" [New] 5L, 63C written
bitnami@debian:~$ ^C
bitnami@debian:~$ ls
apps  bitnami_credentials  htdocs  stack  testfile.txt
bitnami@debian:~$
```

Use the `cat` command to display the contents of the testfile.txt

Create a folder named HDFS using the `hdfs dfs -mkdir` command and verify using the `hdfs dfs -ls` command. (I previously created a folder named HDFS by using only the `mkdir` command without the hdfs dfs in front).

While I could create a folder and a text file, I was not able to copy the testfile.txt file to HDFS using the `cp` or `put` commands. It tells me that the file exists, but the folder remains empty.



Since, I am using less than 8 GB, I need to do the following changes in the yarn-site.xml file, otherwise the wordcount sample will be stuck forever.
Locate the yarn-site.xml file through going to the Hadoop folder using the `cd` command.

```
        link/ether 08:00:27:70:31:c8 brd ff:ff:ff:ff:ff:ff
        inet 10.0.0.13/24 brd 10.0.0.255 scope global dynamic enp0s3
           valid_lft 83087sec preferred_lft 83087sec
        inet6 fe80::a00:27ff:fe70:31c8/64 scope link
           valid_lft forever preferred_lft forever
bitnami@debian:~$ cd /opt/bitnami/hadoop/
bitnami@debian:/opt/bitnami/hadoop$ cd etc/hadoop/
bitnami@debian:/opt/bitnami/hadoop/etc/hadoop$ ls
capacity-scheduler.xml           kms-log4j.properties
configuration.xsl                kms-site.xml
container-executor.cfg           log4j.properties
core-site.xml                    mapred-env.cmd
hadoop-env.cmd                   mapred-env.sh
hadoop-env.sh                    mapred-queues.xml.template
hadoop-metrics2.properties       mapred-site.xml
hadoop-policy.xml                shellprofile.d
hadoop-user-functions.sh.example ssl-client.xml.example
hdfs-site.xml                    ssl-server.xml.example
httpfs-env.sh                    user_ec_policies.xml.template
httpfs-log4j.properties          workers
httpfs-signature.secret          yarn-env.cmd
httpfs-site.xml                  yarn-env.sh
kms-acls.xml                     yarnservice-log4j.properties
kms-env.sh                       yarn-site.xml
bitnami@debian:/opt/bitnami/hadoop/etc/hadoop$ vi yarn
```

Edit the yarn-site.xml file: Replace the two values of 2048 with 8192. Save the file and restart the VM.

```
<?xml version="1.0"?>
<configuration>
<!-- BITNAMI DEFAULT CONFIGURATION -->
<!-- Note: This section will be overwritten on server size changes -->

<!-- The minimum allocation for every container request at the RM (default: 1024
) -->
<property>
    <name>yarn.scheduler.minimum-allocation-mb</name>
    <value>1024</value>
</property>

<!-- The maximum allocation for every container request at the RM (default: 8192
) -->
<property>
    <name>yarn.scheduler.maximum-allocation-mb</name>
    <value>2048</value>
</property>

<!--  Amount of physical memory, in MB, that can be allocated for containers (de
fault: 8192) -->
<property>
    <name>yarn.nodemanager.resource.memory-mb</name>
    <value>2048</value>
"yarn-site.xml" [readonly] 95L, 3366C                         1,1           Top
```

```
<!-- The minimum allocation for every container request at the RM (default: 1024
) -->
<property>
    <name>yarn.scheduler.minimum-allocation-mb</name>
    <value>1024</value>
</property>

<!-- The maximum allocation for every container request at the RM (default: 8192
) -->
<property>
    <name>yarn.scheduler.maximum-allocation-mb</name>
    <value>8192</value>
</property>

<!--  Amount of physical memory, in MB, that can be allocated for containers (de
fault: 8192) -->
<property>
    <name>yarn.nodemanager.resource.memory-mb</name>
    <value>8192</value>
</property>
<!-- END BITNAMI DEFAULT CONFIGURATION -->

<!--
  Licensed under the Apache License, Version 2.0 (the "License");
-- INSERT --                                                  21,16         6%
```

The VM was restarted and the yarn file checked to confirm the changes.

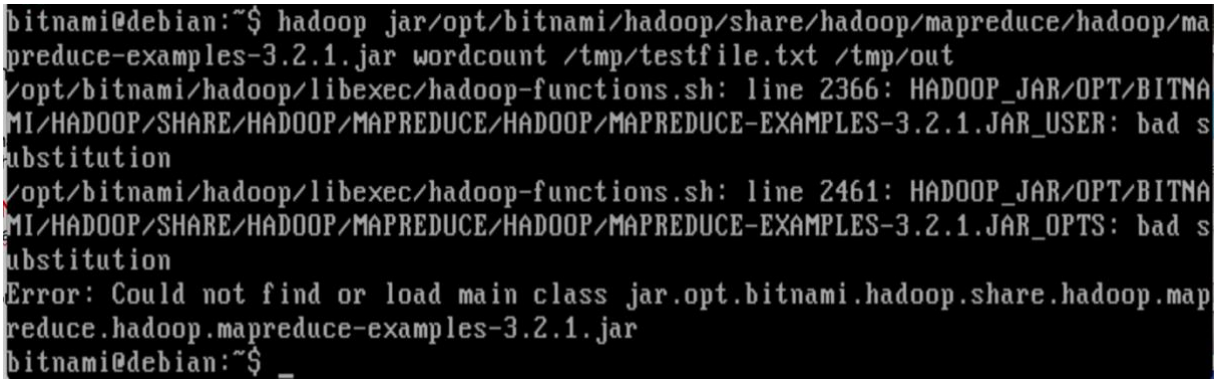
```
<!-- The minimum allocation for every container request at the RM (default: 10
) -->
<property>
    <name>yarn.scheduler.minimum-allocation-mb</name>
    <value>1024</value>
</property>

<!-- The maximum allocation for every container request at the RM (default: 81
) -->
<property>
    <name>yarn.scheduler.maximum-allocation-mb</name>
    <value>8192</value>
</property>

<!--  Amount of physical memory, in MB, that can be allocated for containers (
fault: 8192) -->
<property>
    <name>yarn.nodemanager.resource.memory-mb</name>
    <value>8192</value>
</property>
<!-- END BITNAMI DEFAULT CONFIGURATION -->

<!--
  Licensed under the Apache License, Version 2.0 (the "License");
"yarn-site.xml" 95L, 3366C                                    21,23         6
```

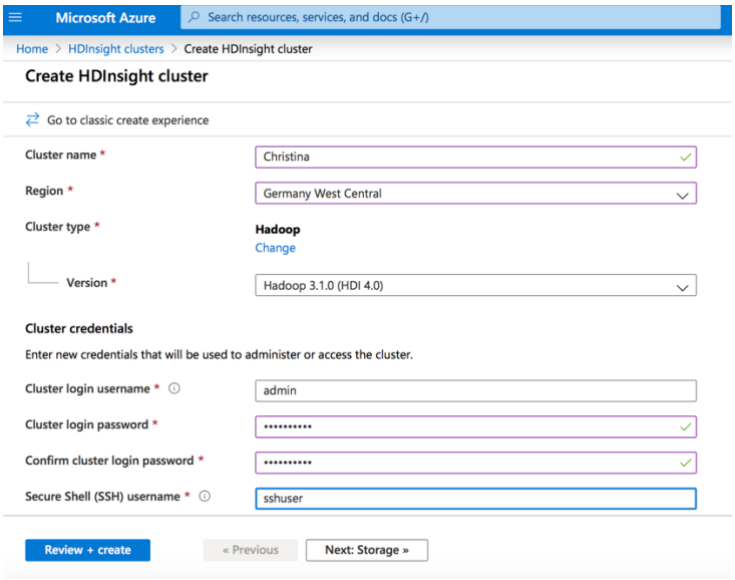
Unfortunately, the mapreduce is not working.



Option 2: Azure HDInsight

I signed up for a 30-day free trial within Azure.

To create an HDInsight Azure Cluster, I searched for the resource HDInsight cluster and filled in the forms as follows. I specified a cluster name, a region (Germany West Central), selected Hadoop for cluster type and specified login username and password.



Unfortunately, I failed to create the cluster because of subscription limits. I tried to change regions, the number of nodes as well as choice of cores available. But in any case, the cluster creation was aborted due to subscription limits.