2016

Creating a 3 Node Hadoop Cluster

Hortonworks Data Platform

NAME :- ARNAB MUKHERJEE

dell
Microsoft
24-Sep-16

Acknowledgement

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations.

I would like to extend my sincere thanks to all of them.I am highly indebted to Acadgild for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my mentor Mr. Manas Puri & my parents for their kind co-operation and encouragement which help me in completion of this project.I would like to express my special gratitude and thanks to my mentor again for giving me such attention and time.

Contents

Hortonworks Data Platform : Automated Install with Ambari

The Hortonworks Data Platform, powered by Apache Hadoop, is a massively scalable and 100% open source platform for storing, processing and analyzing large volumes of data. It is designed to deal with data from many sources and formats in a very quick, easy and cost-effective manner. The Hortonworks Data Platform consists of the essential set of Apache Hadoop projects including MapReduce, Hadoop Distributed File System (HDFS), HCatalog, Pig, Hive, HBase, Zookeeper and Ambari. Hortonworks is the major contributor of code and patches to many of these projects. These projects have been integrated and tested as part of the Hortonworks Data Platform release process and installation and configuration tools have also been included.

1. Getting Ready

This section describes the information and materials we should get ready to install a HDP cluster using Ambari. Ambari provides an end-to-end management and monitoring solution for our HDP cluster. Using the Ambari Web UI and REST APIs, we can deploy, operate, manage configuration changes, and monitor services for all nodes in our cluster from a central point. Here we are going to do this in our local machine with six nodes.

1) VMware workstation
2) Operating Systems Requirements - CentOS v6.8
3) Browser Requirements –
   Linux (CentOS)
   • Firefox 18
   • Google Chrome 26
4) Software Requirements –
   - yum and rpm (CentOS)
   - scp, curl, unzip, tar, and wget
   - For CentOS 6: Python 2.6.*

5) JDK Requirements –
   • Oracle JDK 1.8 64-bit (minimum JDK 1.8_60) (default)
   • Oracle JDK 1.7 64-bit (minimum JDK 1.7_67)
6) Database Requirements
   - PostgreSQL 8
   - PostgreSQL 9.1.13+,9.3
   - MySQL 5.6
   - Oracle 11gr2
   - Oracle 12c*

Note:-By default, Ambari will install an instance of PostgreSQL on the Ambari Server host. Optionally, to use an existing instance of PostgreSQL, MySQL or Oracle

7) Memory Requirements - The Ambari host should have at least 1 GB RAM, with 500 MB free.

2. Collect Information

Before deploying an HDP cluster, we should collect the following information
The fully qualified domain name (FQDN) of each host in our system. The Ambari install wizard supports using IP addresses. We can use hostname -f to check or verify the FQDN of a host.
A list of components we want to set up on each host.

3. Set Up Password-Less SSH

To have Ambari Server automatically install Ambari Agents on all our cluster hosts, we must set up password-less SSH connections between the Ambari Server host and all other hosts in the cluster. The Ambari Server host uses SSH public key authentication to remotely access and install the Ambari Agent.

1. Generate public and private SSH keys on the Ambari Server host.

```
ssh-keygen
```

2. Copy the SSH Public Key (id_rsa.pub) to the root account on our target hosts

```
ssh-copy-id –i /root/.ssh/ id_rsa.pub root@FQDN
```

3. From the Ambari Server, make sure we can connect to each host in the cluster using SSH, without having to enter a password.

ssh root@<remote.target.host> where <remote.target.host> has the value of each host name in our cluster

4. Edit ssh_config file and add below section in the end of the file

```
vi /etc/ssh/ssh_config
StrictHostKeyChecking no
```

5. Depending on our version of SSH, we may need to set permissions on the .ssh directory (to 700) and the authorized_keys file in that directory (to 600) on the target hosts.

```
chmod 700 ~/.ssh
chmod 600 ~/.ssh/authorized_keys
```

6. Retain a copy of the SSH Private Key on the machine from which we will run the webbased Ambari Install Wizard.

4. Enable NTP On The Cluster And On The Browser Host

The clocks of all the nodes in our cluster and the machine that runs the browser through which we access the Ambari Web interface must be able to synchronize with each other. To check that the NTP service will be automatically started upon boot, run the following command on each host:

CentOS

```
chkconfig --list ntpd
systemctl is-enabled ntpd
```

To set the NTP service to auto-start on boot, run the following command on each host:

```
chkconfig ntpd on
systemctl enable ntpd
```

To start the NTP service, run the following command on each host:

```
service ntpd start
systemctl start ntpd
```

5. Edit The Host File

1. Using a text editor, open the hosts file on every host in our cluster. For example:

```
vi /etc/hosts
```

2. Add a line for each host in our cluster. The line should consist of the IP address and the FQDN. For example:

```
1.2.3.4   <fully.qualified.domain.name>
```

6. Set The Hostname

1. Confirm that the hostname is set by running the following command:

```
hostname -f
```

This should return the <fully.qualified.domain.name> we just set.

2. Using a text editor, open the network configuration file on every host and set the desired network configuration for each host. For example

```
vi /etc/sysconfig/network
```

3. Modify the HOSTNAME property to set the fully qualified domain name

```
NETWORKING=yes

HOSTNAME=<fully.qualified.domain.name>
```

7. Configuring Iptables

For Ambari to communicate during setup with the hosts it deploys to and manages, certain ports must be open and available. The easiest way to do this is to temporarily disable iptables, as follows:

chkconfig iptables off

chkconfig ip6tables off

chkconfig NetworkManager off

chkconfig network on

reboot the system

Ambari checks whether iptables is running during the Ambari Server setup process. If iptables is running, a warning displays, reminding us to check that required ports are open and available. The Host Confirm step in the Cluster Install Wizard also issues a warning for each host that has iptables running.

8. Disable Ipv6 And Selinux

Edit /boot/grub/grub.conf and append below parameters in the end of the kernel line of default kernel & reboot the system.

ipv6.disable=1 selinux=0

9. Download Packages

http://public-repo-1.hortonworks.com/HDP/centos6/2.x/updates/2.4.2.0/HDP-2.4.2.0-centos6-rpm.tar.gz
http://public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.20/repos/centos6/HDP-UTILS-1.1.0.20-centos6.tar.gz
http://public-repo-1.hortonworks.com/ambari/centos6/2.x/updates/2.2.2.0/ambari-2.2.2.0-centos6.tar.gz

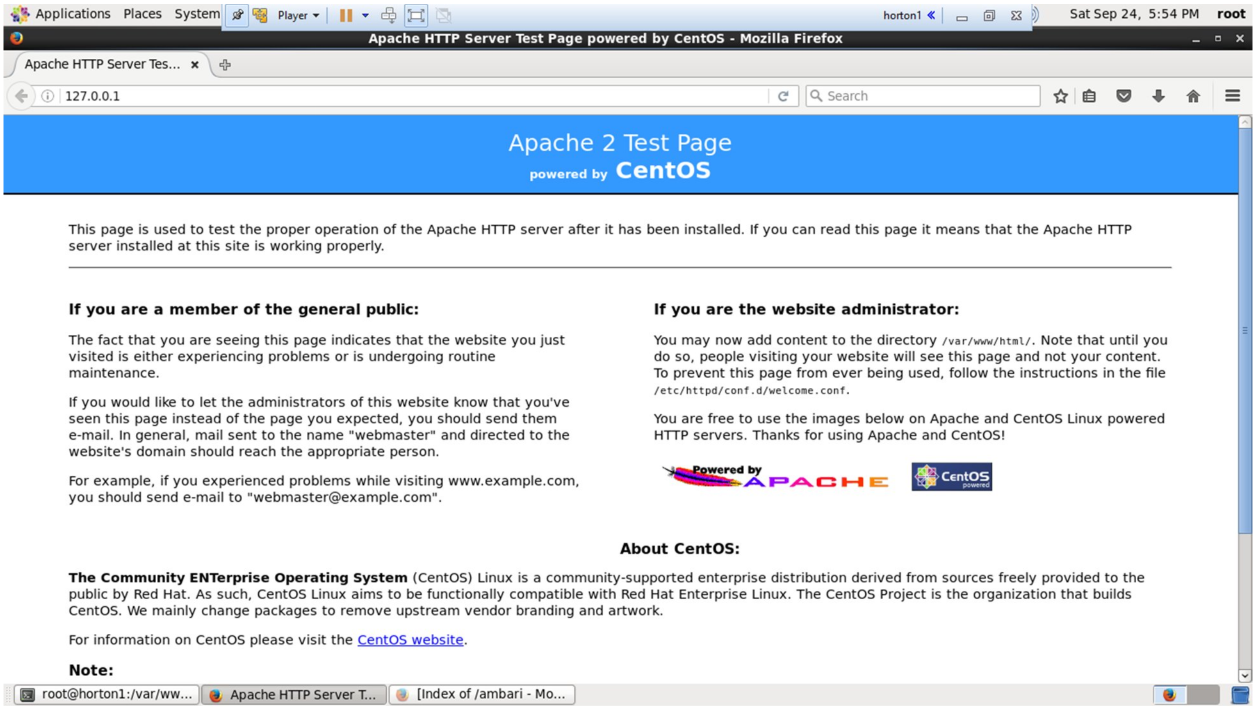
10. Getting Started Setting Up A Local Repository

1. Create an HTTP server.
a. On the mirror server, install an HTTP server (such as Apache httpd) .
b. Activate this web server.
c. Ensure that any firewall settings allow inbound HTTP access from our cluster nodes to our mirror server

```
yum install httpd
service http status
service httpd start
chkconfig http on
```

open web browser and check by typing these in url bar

```
127.0.0.1
```



Apache 2 Test Page
powered by CentOS

This page is used to test the proper operation of the Apache HTTP server after it has been installed. If you can read this page it means that the Apache HTTP server installed at this site is working properly.

If you are a member of the general public:

The fact that you are seeing this page indicates that the website you just visited is either experiencing problems or is undergoing routine maintenance.

If you would like to let the administrators of this website know that you've seen this page instead of the page you expected, you should send them e-mail. In general, mail sent to the name "webmaster" and directed to the website's domain should reach the appropriate person.

For example, if you experienced problems while visiting www.example.com, you should send e-mail to "webmaster@example.com".

If you are the website administrator:

You may now add content to the directory `/var/www/html/`. Note that until you do so, people visiting your website will see this page and not your content. To prevent this page from ever being used, follow the instructions in the file `/etc/httpd/conf.d/welcome.conf`.

You are free to use the images below on Apache and CentOS Linux powered HTTP servers. Thanks for using Apache and CentOS!

Powered by APACHE CentOS powered

About CentOS:

The Community ENTerprise Operating System (CentOS) Linux is a community-supported enterprise distribution derived from sources freely provided to the public by Red Hat. As such, CentOS Linux aims to be functionally compatible with Red Hat Enterprise Linux. The CentOS Project is the organization that builds CentOS. We mainly change packages to remove upstream vendor branding and artwork.

For information on CentOS please visit the CentOS website.

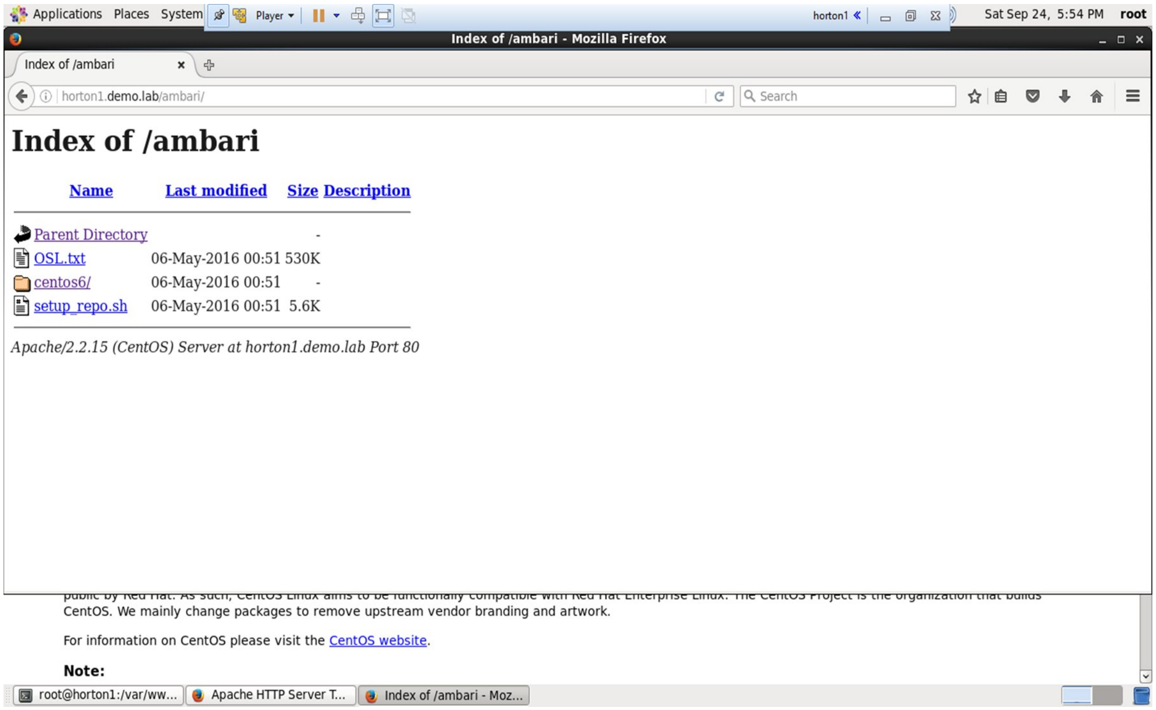
Note:

2. On our mirror server, create a directory for our web server.:

```
mkdir -p /var/www/html/
```

move the extracted tar files in this directori and create soft link

```
tar xzf ambari-2.2.2.0-centos6.tar.gz
tar xzf HDP-UTILS-1.1.0.20-centos6.tar.gz
tar xzf HDP-2.4.2.0-centos6-rpm.tar.gz
mv -v AMBARI-2.2.2.0   /var/www/html
mv -v HDP   /var/www/html
mv -v HDP-UTILS-1.1.0.20   /var/www/html
ln –s /var/www/html/ AMBARI-2.2.2.0   ambari
ln –s /var/www/html/ HDP  hdp
ln –s /var/www/html/ HDP-UTILS-1.1.0.20  hdp-utils
```



Index of /ambari

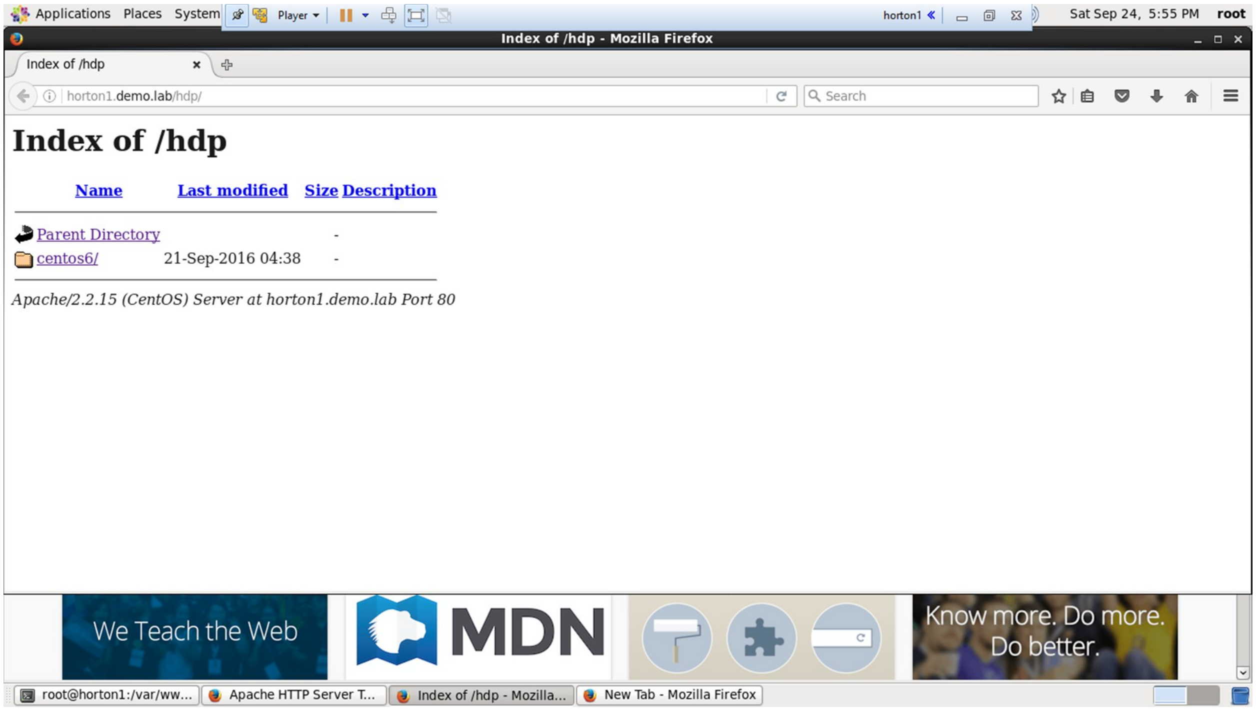
Name	Last modified	Size	Description
Parent Directory		-	
OSL.txt	06-May-2016 00:51	530K	
centos6/	06-May-2016 00:51	-	
setup_repo.sh	06-May-2016 00:51	5.6K	

Apache/2.2.15 (CentOS) Server at horton1.demo.lab Port 80

public by Red Hat. As such, CentOS Linux aims to be functionally compatible with Red Hat Enterprise Linux. The CentOS Project is the organization that builds CentOS. We mainly change packages to remove upstream vendor branding and artwork.

For information on CentOS please visit the CentOS website.

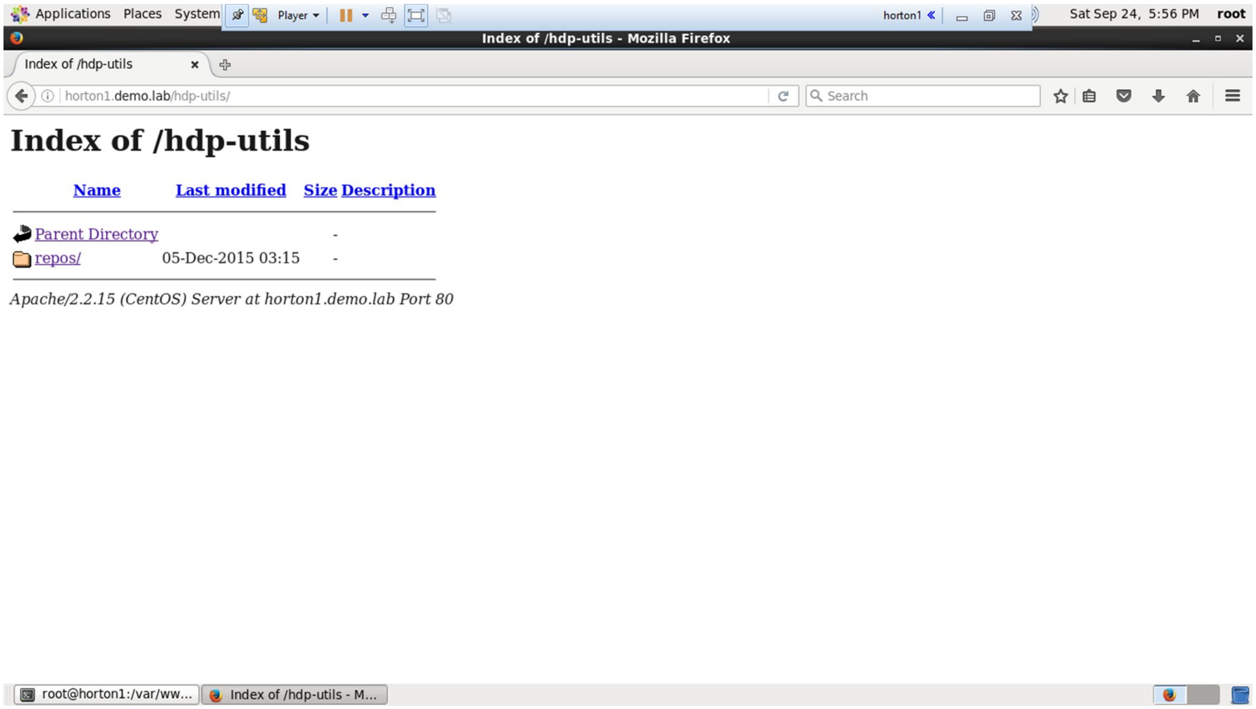
Note:



Index of /hdp

Name	Last modified	Size	Description
Parent Directory		-	
centos6/	21-Sep-2016 04:38	-	

Apache/2.2.15 (CentOS) Server at horton1.demo.lab Port 80


Applications Places System
Player
horton1
Sat Sep 24, 5:56 PM root
Index of /hdp-utils - Mozilla Firefox
Index of /hdp-utils
horton1.demo.lab/hdp-utils/
Search
root@horton1:/var/ww...
Index of /hdp-utils - M...

Index of /hdp-utils

Name	Last modified	Size	Description
Parent Directory		-	
repos/	05-Dec-2015 03:15	-	

Apache/2.2.15 (CentOS) Server at horton1.demo.lab Port 80

3.Now open the file /etc/yum.repos.d/ambari.repo and enter the following information

[AMBARI-2.2.2.X]

name=ambary

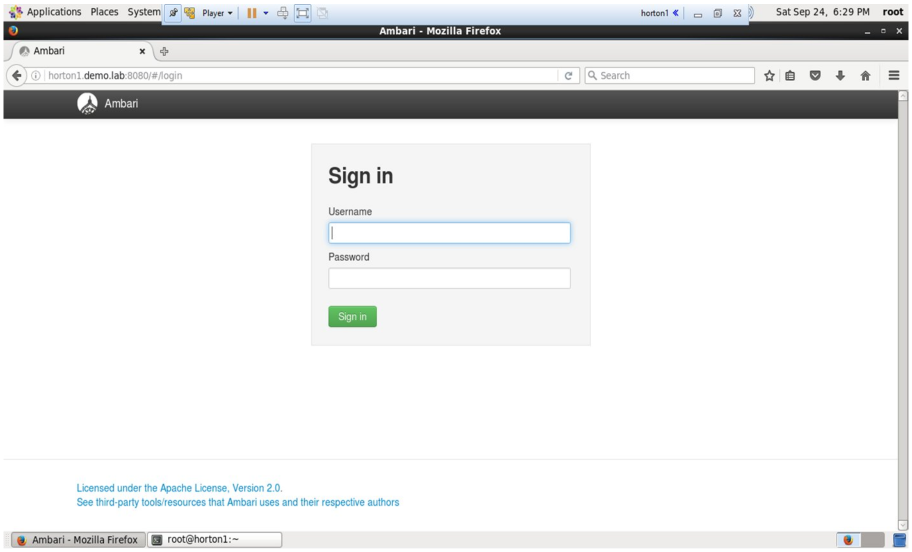
```
baseurl=http://horton1.demo.lab/ambari/centos6/2.2.2.0-460

gpgkey=http://horton1.demo.lab/ambari/centos6/2.2.2.0-460/RPM-GPG-KEY/RPM-GPG-KEY-Jenkins

enabled=1
```

4. Confirm availability of the repositories

```
yum repolist
```

11. Installing Ambari

1. Log in to our host as root.
2. Clean the repository
   ```
   yum clean all
   ```
3. Confirm that the repository is configured by checking the repo list.
   ```
   yum repolist
   ```
   We should see values similar to the following for Ambari repositories in the list.
4. Find ambary packages
   ```
   Yum list all | grep –I ambari
   ```
   ```
   ambari-agent.x86_64                  2.2.2.0-460          @AMBARI-2.2.2.X
   ambari-metrics-grafana.x86_64          2.2.2.0-460          @AMBARI-2.2.2.X
   ambari-metrics-hadoop-sink.x86_64        2.2.2.0-460          @AMBARI-2.2.2.X
   ambari-metrics-monitor.x86_64          2.2.2.0-460          @AMBARI-2.2.2.X
   ambari-server.x86_64                 2.2.2.0-460          @AMBARI-2.2.2.X
   ambari-metrics-collector.x86_64         2.2.2.0-460          AMBARI-2.2.2.X
   ambari-metrics-common.noarch            2.2.2.0-460          AMBARI-2.2.2.X
   smartsense-hst.x86_64                1.2.2.0-460          AMBARI-2.2.2.X
   ```
5. Install the Ambari bits. This also installs the default PostgreSQL Ambari database.
   ```
   yum install ambari-server.x86_64
   ```
6. Choose the default options.
7. To check the ambari status
   ```
   ambari-server status
   ```
8. To start ambary server
   ```
   ambari-server start
   ```
9. To stop ambari server
   ```
   ambary-server stop
   ```



Licensed under the Apache License, Version 2.0.
See third-party tools/resources that Ambari uses and their respective authors

12. Installing, Configuring, And Deploying A HDP Cluster

1. Log In to Apache Ambari

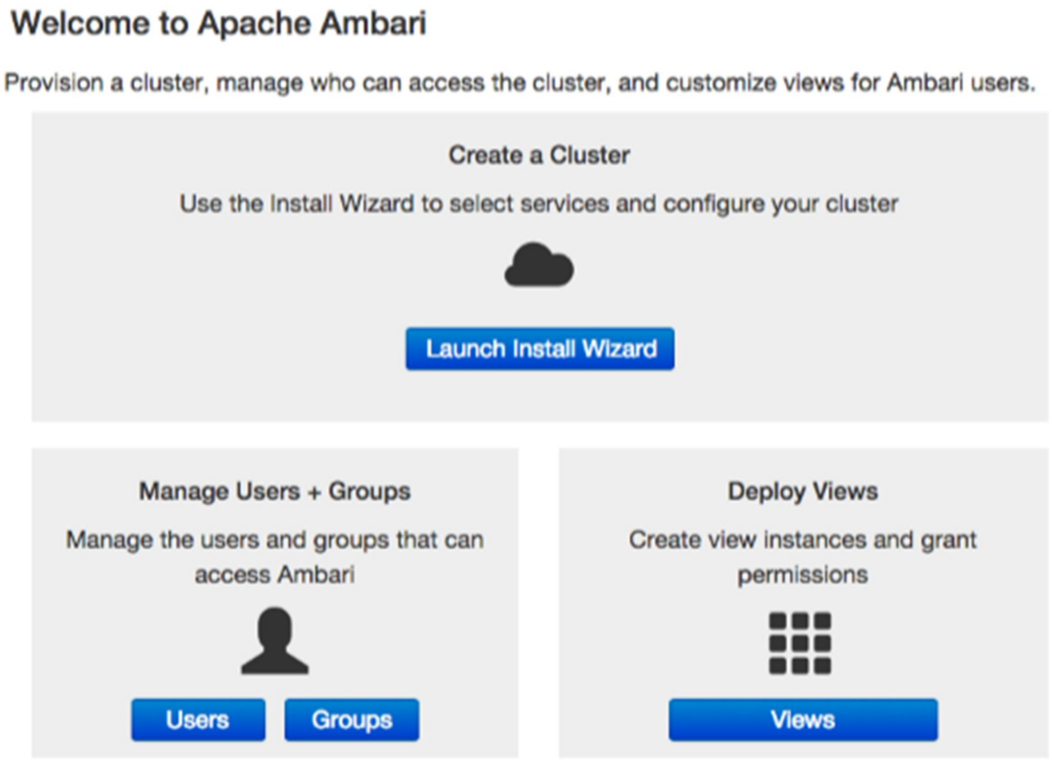
After starting the Ambari service, open Ambari Web using a web browser.

Point our browser to http://<our.ambari.server>:8080,where <our.ambari.server> is the name of our ambari server host. For example, a default Ambari server host is located at http://c6401.ambari.apache.org:8080.

Log in to the Ambari Server using the default user name/password: admin/admin. We can change these credentials later.

.2. Launching the Ambari Install Wizard

From the Ambari Welcome page, choose Launch Install Wizard



Welcome to Apache Ambari

Provision a cluster, manage who can access the cluster, and customize views for Ambari users.

Create a Cluster

Use the Install Wizard to select services and configure your cluster

Manage Users + Groups

Manage the users and groups that can access Ambari

Deploy Views

Create view instances and grant permissions

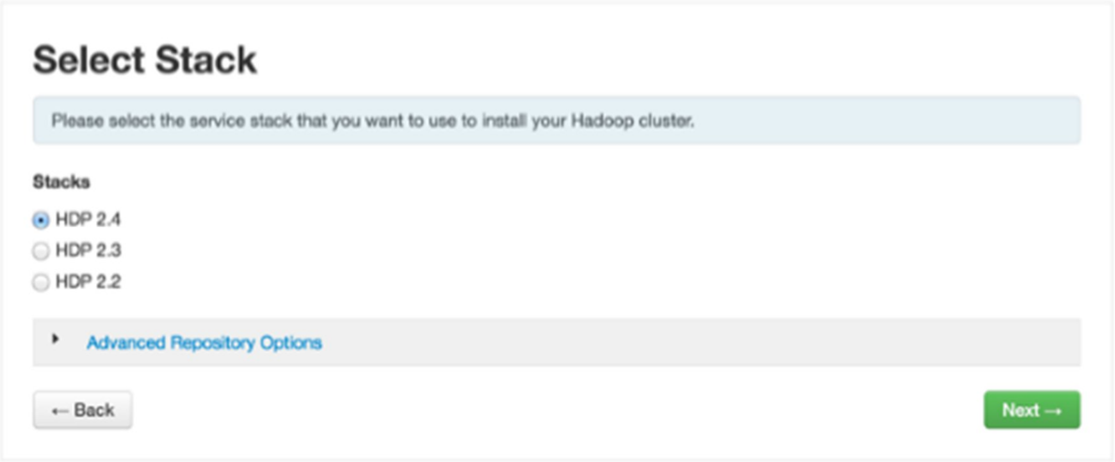
3. Name our Cluster

In Name our cluster, type a name for the cluster we want to create. Use no white spaces or special characters in the name.

Choose Next

4. Select Stack

The Service Stack (the Stack) is a coordinated and tested set of HDP components. Use a radio button to select the Stack version we want to install.



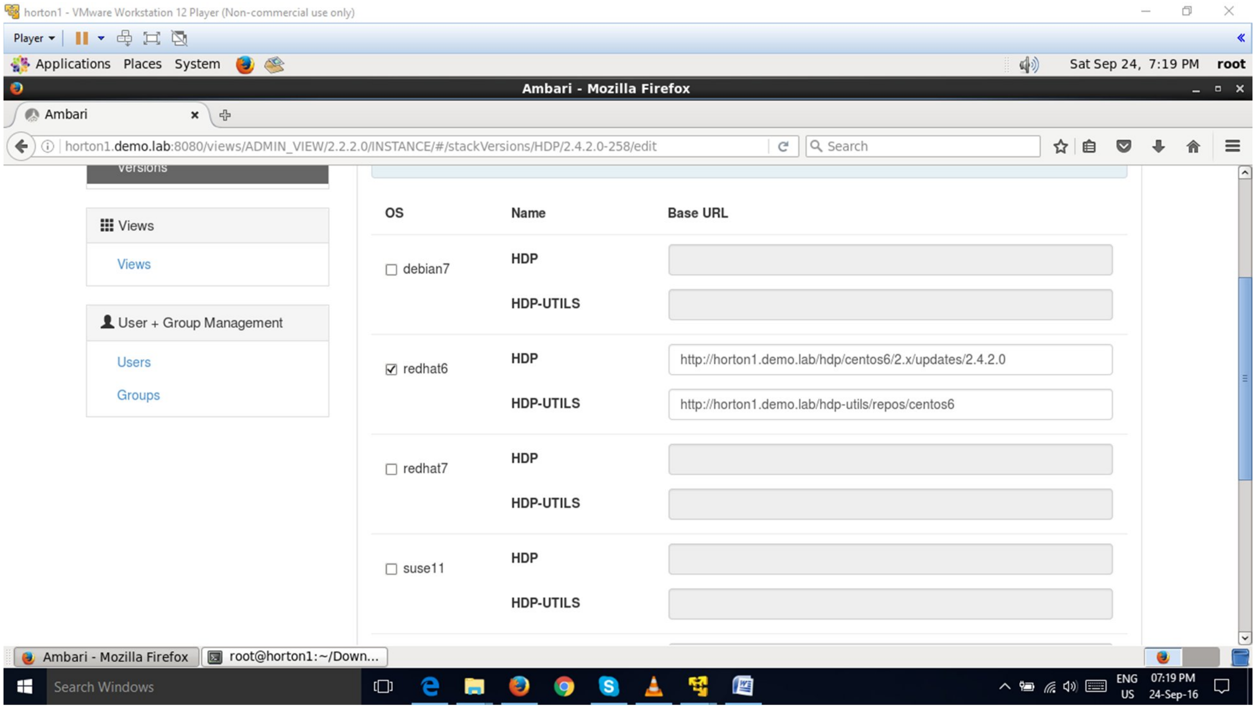
Select Stack

Please select the service stack that you want to use to install your Hadoop cluster.

Stacks

● HDP 2.4
○ HDP 2.3
○ HDP 2.2

5. Expand Advanced Repository Options to select the Base URL of a repository from which Stack software packages download.



OS	Name	Base URL
□ debian7	HDP	
	HDP-UTILS	
☑ redhat6	HDP	http://horton1.demo.lab/hdp/centos6/2.x/updates/2.4.2.0
	HDP-UTILS	http://horton1.demo.lab/hdp-utils/repos/centos6
□ redhat7	HDP	
	HDP-UTILS	
□ suse11	HDP	
	HDP-UTILS	

6. In order to build up the cluster, the install wizard prompts us for general information about how we want to set it up. We need to supply the FQDN of each of our hosts. The wizard also needs to access the private key file we created in Set Up Password-less SSH. Using the host names and key file information, the wizard can locate, access, and interact securely with all hosts in the cluster.

7. Confirm Hosts prompts us to confirm that Ambari has located the correct hosts for our cluster and to check those hosts to make sure they have the correct directories, packages, and processes required to continue the install.

If any hosts were selected in error, we can remove them by selecting the appropriate checkboxes and clicking the grey Remove Selected button. To remove a single host, click the small white Remove button in the Action column.

At the bottom of the screen, we may notice a yellow box that indicates some warnings were encountered during the check process. For example, our host may have already had a copy of wget or curl. Choose Click here to see the warnings to see a list of what was checked and what caused the warning. The warnings page also provides access to a python script that can help we clear any issues we may encounter and let we run Rerun Checks.

When we are satisfied with the list of hosts, choose Next

8. Based on the Stack chosen during Select Stack, we are presented with the choice of Services to install into the cluster. HDP Stack comprises many services. We may choose to install any other available services now, or to add services later. The install wizard selects all available services for installation by default.

- Choose none to clear all selections, or choose all to select all listed services.
- Choose or clear individual checkboxes to define a set of services to install now.
- After selecting the services to install now, choose Next

9. The Ambari install wizard assigns the master components for selected services to appropriate hosts in our cluster and displays the assignments in Assign Masters. The left column shows services and current hosts. The right column shows current master component assignments by host, indicating the number of CPU cores and amount of RAM installed on each host.

- To change the host assignment for a service, select a host name from the drop-down menu for that service.
- To remove a ZooKeeper instance, click the green minus icon next to the host address we want to remove.
- When we are satisfied with the assignments, choose Next

10. Assign Slaves and Clients

The Ambari installation wizard assigns the slave components (DataNodes, NodeManagers, and RegionServers) to appropriate hosts in our cluster. It also attempts to select hosts for installing the appropriate set of clients.

- Use all or none to select all of the hosts in the column or none of the hosts, respectively.
  If a host has an asterisk next to it, that host is also running one or more master components. Hover our mouse over the asterisk to see which master components are on that host.
- Fine-tune our selections by using the checkboxes next to specific hosts.
- When we are satisfied with our assignments, choose Next.

11. The Customize Services step presents we with a set of tabs that let we review and modify our HDP cluster setup. The wizard attempts to set reasonable defaults for each of the options. We are strongly encouraged to review these settings as our requirements might be slightly different.
    Browse through each service tab and by hovering our cursor over each of the properties, we can see a brief description of what the property does. The number of service tabs shown depends on the services we decided to install in our cluster. Any tab that requires input shows a red badge with the number of properties that need attention. Select each service tab that displays a red badge number and enter the appropriate information.
    Directories
    The choice of directories where HDP will store information is critical. Ambari will attempt to choose reasonable defaults based on the mount points available in our environment but we are strongly encouraged to review the default directory settings recommended by Ambari. In particular, confirm directories such as /tmp and /var are not being used for HDFS NameNode directories and DataNode directories under the HDFS tab.
    Passwords
    We must provide database passwords for the Hive and Oozie services and the Master Secret for Knox. Using Hive as an example, choose the Hive tab and expand the Advanced section. In Database Password field marked in red, provide a password, then retype to confirm it.

12. The assignments we have made are displayed. Check to make sure everything is correct. If we need to make changes, use the left navigation bar to return to the appropriate screen.

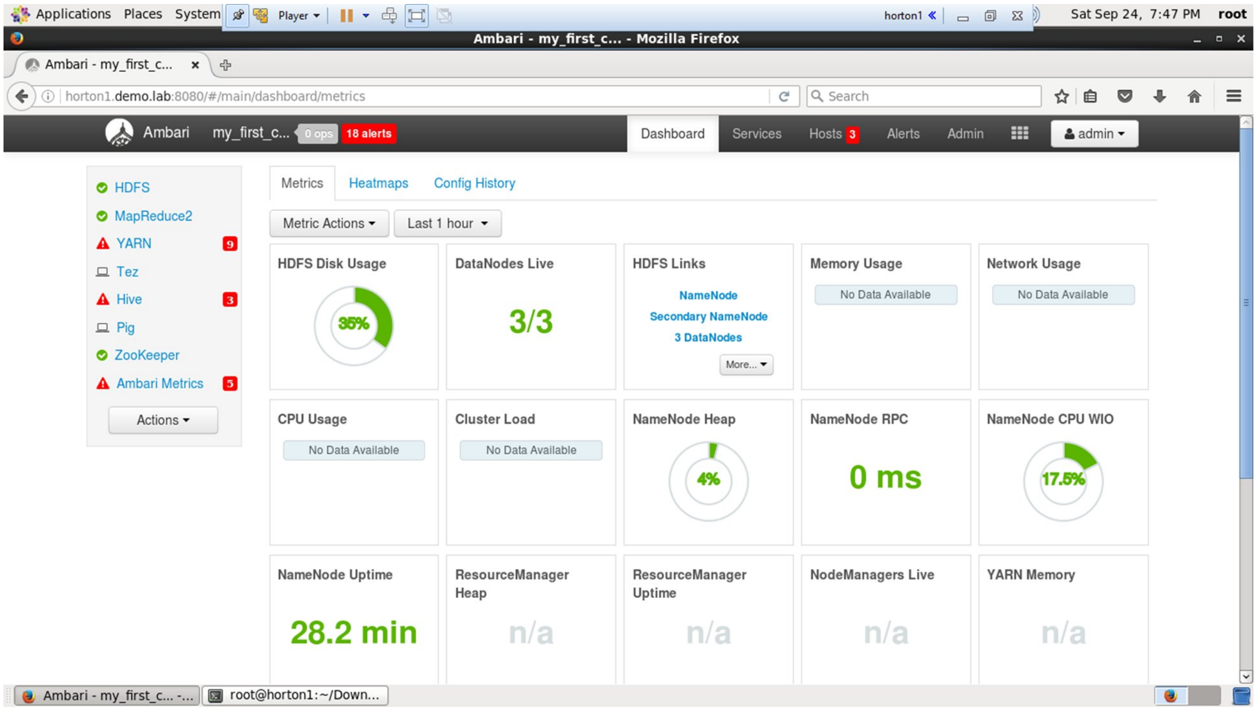
    To print our information for later reference, choose Print.

    When we are satisfied with our choices, choose Deploy.

13. The progress of the install displays on the screen. Ambari installs, starts, and runs a simple test on each component. Overall status of the process displays in progress bar at the top of the screen and host-by-host status displays in the main section. Do not refresh our browser during this process. Refreshing the browser may interrupt the progress indicators.
    To see specific information on what tasks have been completed per host, click the link in the Message column for the appropriate host. In the Tasks pop-up, click the individual

task to see the related log files. We can select filter conditions by using the Show dropdown list. To see a larger version of the log contents, click the Open icon or to copy the contents to the clipboard, use the Copy icon.
When Successfully installed and started the services appears, choose Next.

14. The Summary page provides us a summary list of the accomplished tasks. Choose Complete. Ambari Web GUI displays



Applications Places System
Player
horton1
Sat Sep 24, 7:47 PM root
Ambari - my_first_c... - Mozilla Firefox
Ambari - my_first_c...
horton1.demo.lab:8080/#/main/dashboard/metrics
Search
Ambari my_first_c... 0 ops 18 alerts
Dashboard
Services
Hosts 3
Alerts
Admin
admin
HDFS
MapReduce2
YARN 9
Tez
Hive 3
Pig
ZooKeeper
Ambari Metrics 5
Actions
Metrics
Heatmaps
Config History
Metric Actions
Last 1 hour
HDFS Disk Usage
35%
DataNodes Live
3/3
HDFS Links
NameNode
Secondary NameNode
3 DataNodes
More...
Memory Usage
No Data Available
Network Usage
No Data Available
CPU Usage
No Data Available
Cluster Load
No Data Available
NameNode Heap
4%
NameNode RPC
0 ms
NameNode CPU WIO
17.5%
NameNode Uptime
28.2 min
ResourceManager Heap
n/a
ResourceManager Uptime
n/a
NodeManagers Live
n/a
YARN Memory
n/a
Ambari - my_first_c... -...
root@horton1:~/Down...