Setting Up
Big Data Engineering Lab
using
Cloudera Data Platform
on
Google Cloud Platform
<u>Hands-on Guide</u>

Overview 4

Prerequisites 4

Create GCP Instance 4

Creating User 7

Configuring Selinux 8

Set Static Hostname and edit Network Configuration 9

Disabling the Firewall 11

Configuring Oracle Java 12

Installing Server 13

Configure Swappiness and Installing NTP 14

Configuring MySQL 16

Configure the External Database 17

Set Firewall rule on GCP 19

Cloudera Data Platform Installation 21

Working with Cloudera Manager 23

Set Up a Cluster 31

Testing Your Hadoop Installation 37

Hands-On-Exercise: Installing Hive 38

Hands-On-Exercise: Installing Tez 42

Hands-On-Exercise: Installing Hive on Tez 44

Hive Validation 46

Hands-On-Exercise: Deploying Spark 2.4 48

Running Job on Apache Spark2 52

Hands-On-Exercise: Installing Kafka 54

Kafka Validation 57

Hands-On-Exercise: Installing Impala 58

Troubleshooting 61

References 62

Trademark Information

The names and logos of Apache products mentioned in our courses, including those listed below, are trademarks of the Apache Software Foundation.

- Apache Hadoop
- Apache HBase
- Apache Hive
- Apache Impala (incubating)
- Apache Kafka
- Apache Spark
- Apache ZooKeeper

All other product names, logos, and brands cited herein are the property of their respective owners.

Overview

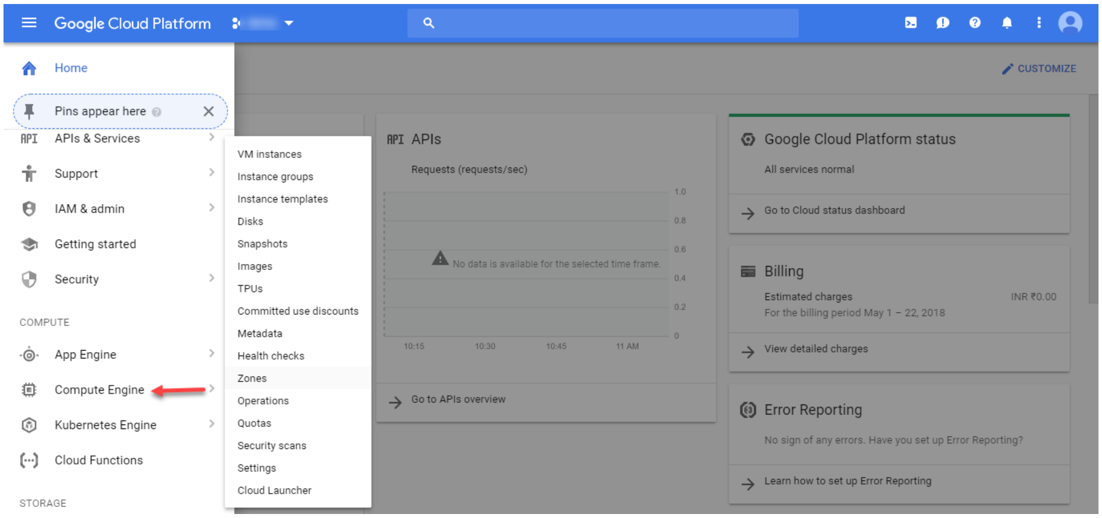
The purpose of this document is to provide hands-on exposure to participants to setup Big Data Engineering Lab in Pseudo-distributed mode (Single Machine Cluster). The document leverages new Cloudera platform i.e. Cloudera Data Platform for deploying HDFS, YARN, Hive, Spark etc. on Google Cloud Platform. This document leverages free credits of Google Cloud Platform so there is no need to pay anything to anyone for running labs till free credits are available.

Prerequisites

1. Must setup Google account as mentioned in steps as documented in **"sign-up-for-google-cloud-platform.pdf"**
2. Execute below steps in Incognito mode using Google Account which has been setup in Step 1
3. Knowledge of basic unix commands and familiarity with vi editor will be helpful

Create GCP Instance

1. Go to the VM instances



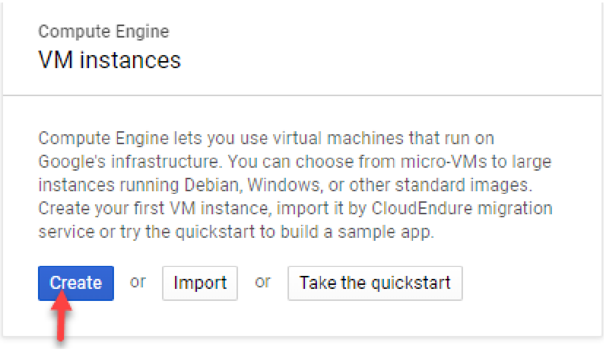
No data is available for the selected time frame.

All services normal

For the billing period May 1 – 22, 2018

No sign of any errors. Have you set up Error Reporting?

2. Select your project and click Continue.
3. Click the Create instance button.

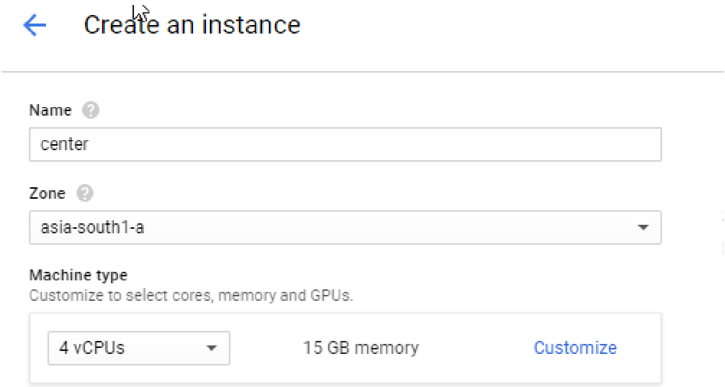


Compute Engine
VM instances

Compute Engine lets you use virtual machines that run on Google's infrastructure. You can choose from micro-VMs to large instances running Debian, Windows, or other standard images. Create your first VM instance, import it by CloudEndure migration service or try the quickstart to build a sample app.

4. Specify a Name for your first instance “center”.
5. Change the Zone for this instance “asia-south-1-a“ (best is to choose closest to your location).
6. Select a Machine type for your instance.

   4 vCPUs 15 GB memory.



Create an instance

Name: center

Zone: asia-south1-a

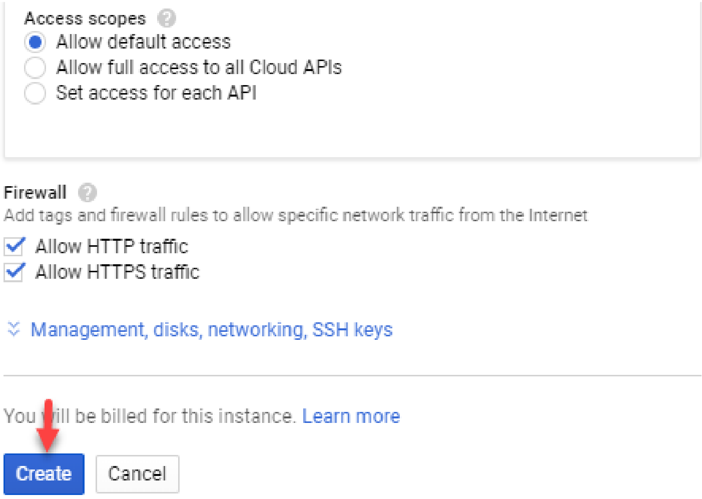
Machine type
Customize to select cores, memory and GPUs.

4 vCPUs — 15 GB memory

7. In the Boot disk section, click Change to configure your boot disk.
8. In the OS images tab, choose an “**CentOS 7**” image.
9. Select Boot disk type **“Standard persistent disk”** and size should be **“200GB”**
10. Click the **“Select”** Button.
11. In Access scopes Select **“Allow default access”**
12. In Firewall, choose below options

    Allow HTTP traffic and **Allow HTTPS traffic**.

13. Click the Create button to create and start the instance.



Access scopes

● Allow default access

○ Allow full access to all Cloud APIs

○ Set access for each API

Firewall

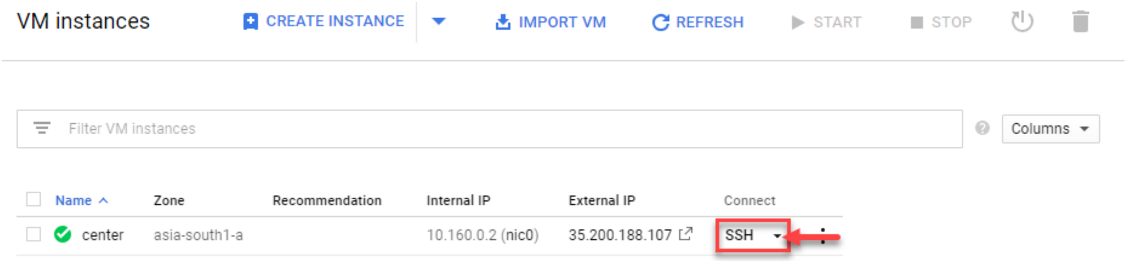
Add tags and firewall rules to allow specific network traffic from the Internet

☑ Allow HTTP traffic

☑ Allow HTTPS traffic

You will be billed for this instance. Learn more

Click on ssh to open the terminal



Name	Zone	Recommendation	Internal IP	External IP	Connect
center	asia-south1-a		10.160.0.2 (nic0)	35.200.188.107	SSH

Creating User

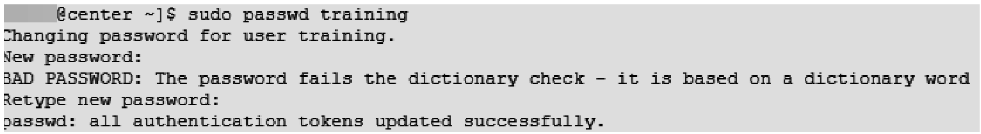
You have to create a user “training” and set its password “password”

Add user in CentOS, open a terminal

```
$ sudo useradd training
```

Assign a password using passwd command and set the password: “password”

```
$ sudo passwd training
```



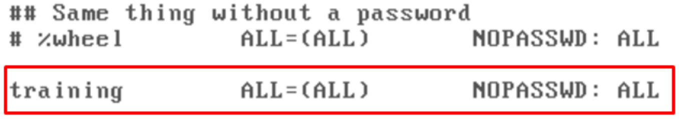
```
[redacted]@center ~]$ sudo passwd training
Changing password for user training.
New password:
BAD PASSWORD: The password fails the dictionary check - it is based on a dictionary word
Retype new password:
passwd: all authentication tokens updated successfully.
```

Edit the sudoers file through vi command

```
$ sudo vi /etc/sudoers
```

You have to follow either of the 2 options

1. Add a new highlighted line in this file
   training   ALL=(ALL)   NOPASSWD: ALL



```
## Same thing without a password
# %wheel          ALL=(ALL)         NOPASSWD: ALL

training          ALL=(ALL)         NOPASSWD: ALL
```

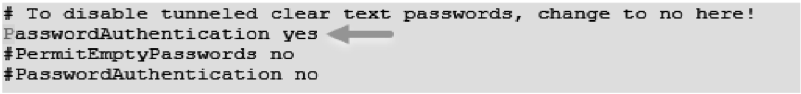
2. Replace # %wheel with training in above line.

Once this is done, save the file use Esc :wq!

Set PasswordAuthentication as yes and restart ssh daemon

```
$ sudo vi /etc/ssh/sshd_config
```

Uncomment **`PasswordAuthentication yes`** and comment the **`PasswordAuthentication no`**



```
# To disable tunneled clear text passwords, change to no here!
PasswordAuthentication yes
#PermitEmptyPasswords no
#PasswordAuthentication no
```

```
$ sudo systemctl restart sshd.service
```

Configuring Selinux

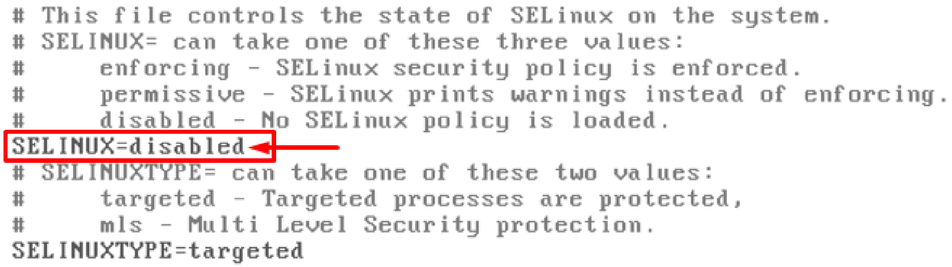
You'll have to set Selinux mode to **"disabled"** by configuring selinux file located in /etc/sysconfig directory.

Open the file and make changes:-

```
$ sudo vi /etc/sysconfig/selinux
```

Update the following line:-

SELINUX=disabled



```
# This file controls the state of SELinux on the system.
# SELINUX= can take one of these three values:
#     enforcing - SELinux security policy is enforced.
#     permissive - SELinux prints warnings instead of enforcing.
#     disabled - No SELinux policy is loaded.
SELINUX=disabled
# SELINUXTYPE= can take one of these two values:
#     targeted - Targeted processes are protected,
#     mls - Multi Level Security protection.
SELINUXTYPE=targeted
```

Save the file and exit.

Reboot the machine.

```
$ sudo reboot
```

Set Static Hostname and edit Network Configuration

```
sudo hostnamectl set-hostname training.io --static
```

Now verify…

```
hostname -f
```

Create and edit a file /usr/sbin/hosts.sh and set it executable

```
sudo touch /usr/sbin/hosts.sh
sudo chmod 755 /usr/sbin/hosts.sh
sudo vi /usr/sbin/hosts.sh
```

Paste the below script in hosts.sh

```
#!/bin/sh
#
# Script to determine the FQDN of a node in GCP and update hosts file
#
sudo hostnamectl set-hostname training.io --static
myhost=`hostname -f`
ipaddr=`ifconfig eth0 | grep "inet " | grep -oE
"\b([0-9]{1,3}\.){3}[0-9]{1,3}\b" | head -1`
echo '127.0.0.1     localhost localhost.localdomain localhost4
localhost4.localdomain4' > /etc/hosts.latest
echo $ipaddr  '  ' $myhost >> /etc/hosts.latest

mv /etc/hosts /etc/hosts.old
mv /etc/hosts.latest /etc/hosts
```

Edit rc.local to run the hosts.sh script at boot for CentOS 7+, rc.local is not set as executable by default!

```
sudo chmod +x /etc/rc.d/rc.local
sudo systemctl enable rc-local
sudo systemctl start rc-local
```

Add below command inside ***/etc/rc.d/rc.local***

```
sudo sh /usr/sbin/hosts.sh
```

Network Configuration

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

Check hostname using below command

```
hostname -f
```

Copy hostname and paste in a network configuration file

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

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

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

Disabling the Firewall

To disable the firewall on each host in your cluster, perform the following steps on each host.

For iptables, save the existing rule set:

```
sudo iptables-save > ~/firewall.rules
```

Disable the firewall:

RHEL 7 compatible:

```
sudo systemctl disable firewalld
sudo systemctl stop firewalld
```

Configuring Oracle Java

1. **Check the current java version:-**

```
$ java -version
```

We can see that there is no OpenJdk installed

2. **Install OpenJDK 8 JDK**

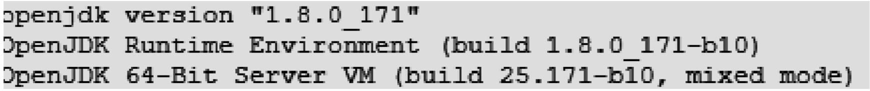
To install **OpenJDK** 8 JDK using yum, run this command:

```
$ sudo yum install java-1.8.0-openjdk-devel -y
```

At the confirmation prompt, enter y then RETURN to continue with the installation.

3. **Now re-check the java version.**

```
$ java -version
```



```
openjdk version "1.8.0_171"
OpenJDK Runtime Environment (build 1.8.0_171-b10)
OpenJDK 64-Bit Server VM (build 25.171-b10, mixed mode)
```

4. **Set JAVA_HOME variable and append to the PATH**

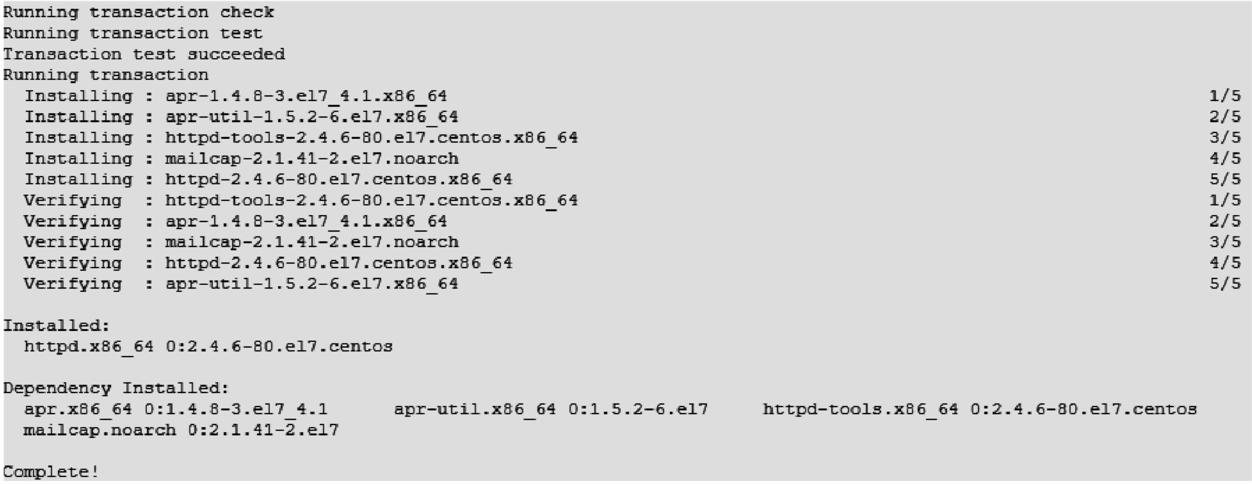
```
$ echo $JAVA_HOME
$ export
JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.171-8.b10.el7_5.x86_64
$ export PATH=$PATH:$JAVA_HOME
$ echo $JAVA_HOME
$ echo $PATH
```

Note: Please specify the path of JAVA_HOME as per your version.

Installing Server

1. Install the httpd daemon and restart

```
$ sudo yum install httpd -y
```

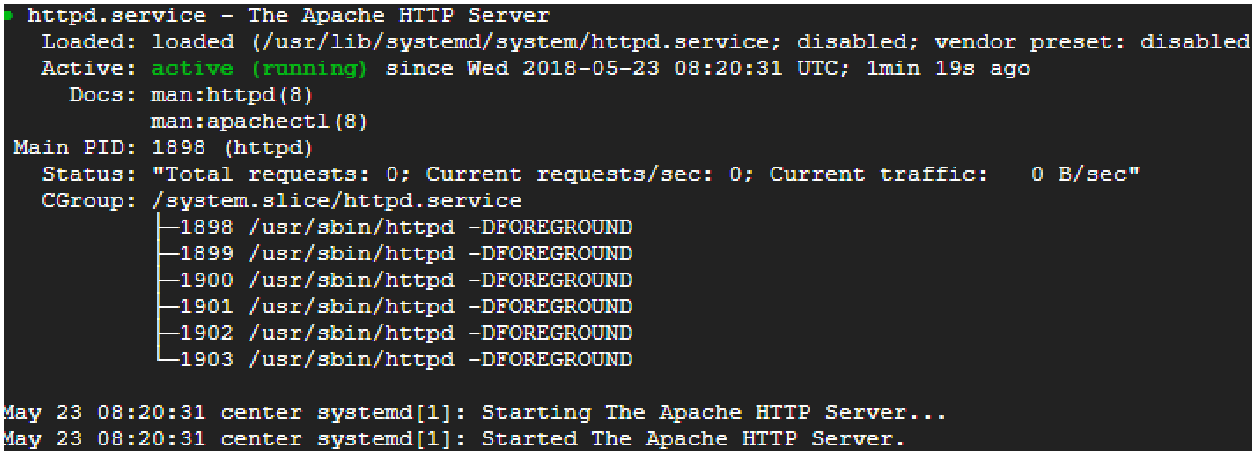


```
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Installing : apr-1.4.8-3.el7_4.1.x86_64                                                          1/5
  Installing : apr-util-1.5.2-6.el7.x86_64                                                         2/5
  Installing : httpd-tools-2.4.6-80.el7.centos.x86_64                                              3/5
  Installing : mailcap-2.1.41-2.el7.noarch                                                         4/5
  Installing : httpd-2.4.6-80.el7.centos.x86_64                                                    5/5
  Verifying  : httpd-tools-2.4.6-80.el7.centos.x86_64                                              1/5
  Verifying  : apr-1.4.8-3.el7_4.1.x86_64                                                          2/5
  Verifying  : mailcap-2.1.41-2.el7.noarch                                                         3/5
  Verifying  : httpd-2.4.6-80.el7.centos.x86_64                                                    4/5
  Verifying  : apr-util-1.5.2-6.el7.x86_64                                                         5/5

Installed:
  httpd.x86_64 0:2.4.6-80.el7.centos

Dependency Installed:
  apr.x86_64 0:1.4.8-3.el7_4.1      apr-util.x86_64 0:1.5.2-6.el7     httpd-tools.x86_64 0:2.4.6-80.el7.centos
  mailcap.noarch 0:2.1.41-2.el7

Complete!
```

```
$ sudo systemctl start httpd.service
```

Check Status

```
$ sudo systemctl status httpd.service
```



```
● httpd.service - The Apache HTTP Server
   Loaded: loaded (/usr/lib/systemd/system/httpd.service; disabled; vendor preset: disabled
   Active: active (running) since Wed 2018-05-23 08:20:31 UTC; 1min 19s ago
     Docs: man:httpd(8)
           man:apachectl(8)
 Main PID: 1898 (httpd)
   Status: "Total requests: 0; Current requests/sec: 0; Current traffic:   0 B/sec"
   CGroup: /system.slice/httpd.service
           ├─1898 /usr/sbin/httpd -DFOREGROUND
           ├─1899 /usr/sbin/httpd -DFOREGROUND
           ├─1900 /usr/sbin/httpd -DFOREGROUND
           ├─1901 /usr/sbin/httpd -DFOREGROUND
           ├─1902 /usr/sbin/httpd -DFOREGROUND
           └─1903 /usr/sbin/httpd -DFOREGROUND

May 23 08:20:31 center systemd[1]: Starting The Apache HTTP Server...
May 23 08:20:31 center systemd[1]: Started The Apache HTTP Server.
```

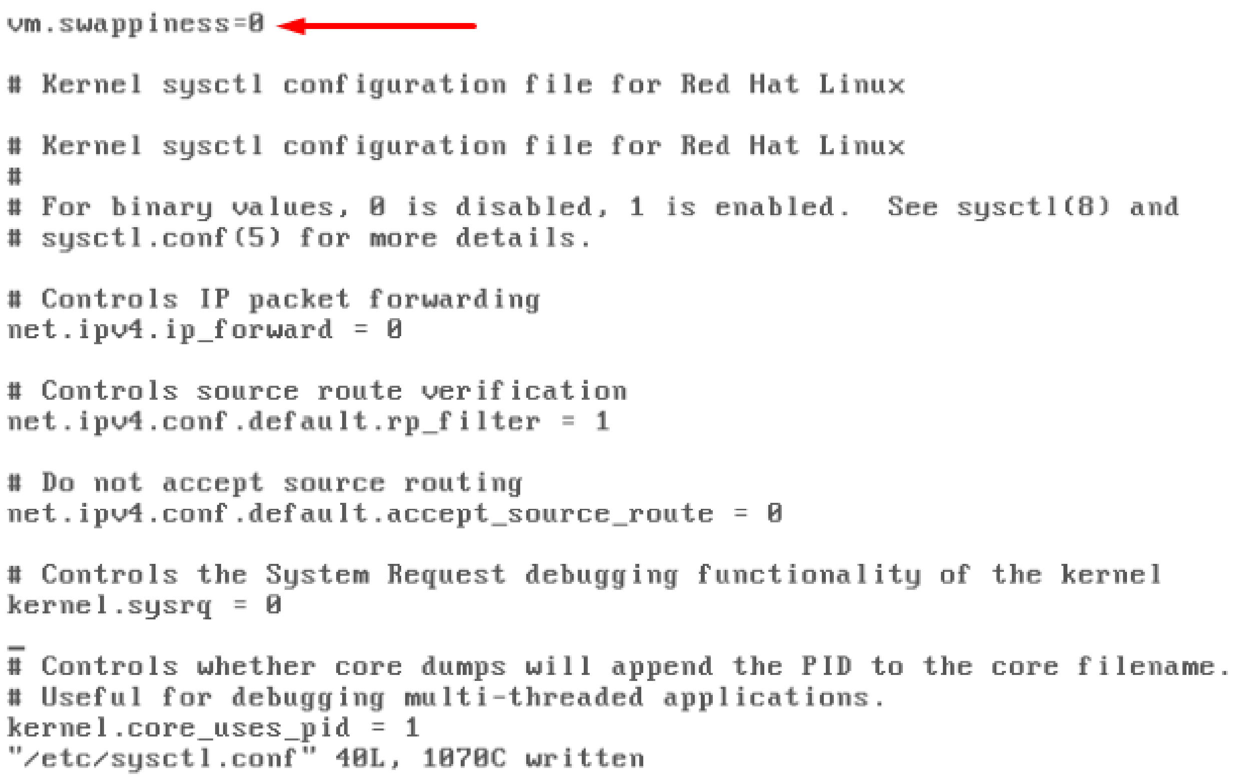
```
$ sudo systemctl enable httpd.service
```

Configure Swappiness and Installing NTP

Configure VM Swappiness

```
$ sudo vi /etc/sysctl.conf
```

Add this line: *vm.swappiness=0*



```
vm.swappiness=0

# Kernel sysctl configuration file for Red Hat Linux

# Kernel sysctl configuration file for Red Hat Linux
#
# For binary values, 0 is disabled, 1 is enabled.  See sysctl(8) and
# sysctl.conf(5) for more details.

# Controls IP packet forwarding
net.ipv4.ip_forward = 0

# Controls source route verification
net.ipv4.conf.default.rp_filter = 1

# Do not accept source routing
net.ipv4.conf.default.accept_source_route = 0

# Controls the System Request debugging functionality of the kernel
kernel.sysrq = 0

_
# Controls whether core dumps will append the PID to the core filename.
# Useful for debugging multi-threaded applications.
kernel.core_uses_pid = 1
"/etc/sysctl.conf" 40L, 1070C written
```

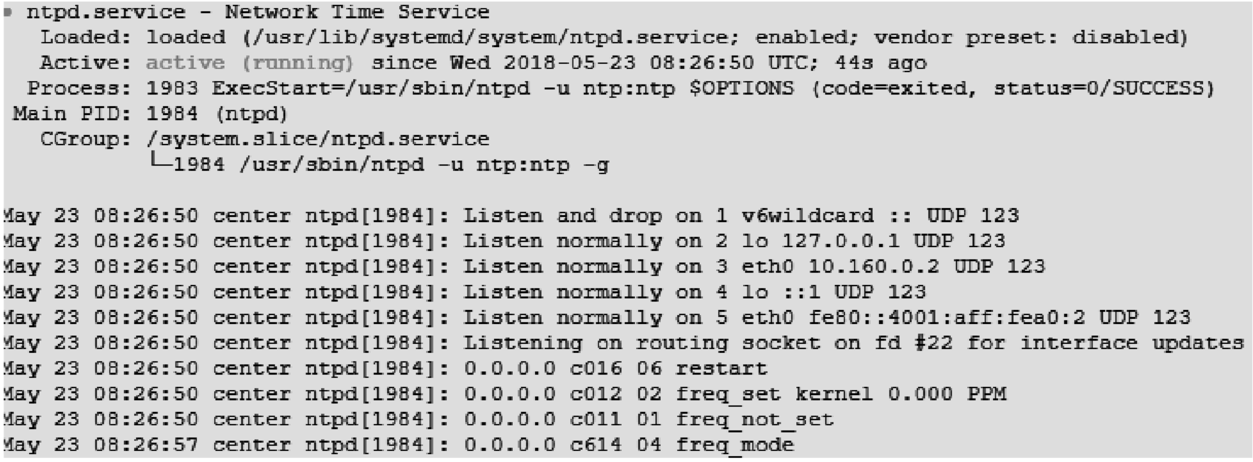
Reboot the machine to make the changes effective

Install the ntp package

```
$ sudo yum install ntp -y
```

Start the service and verify its status.

```
$ sudo systemctl start ntpd.service
$ sudo systemctl status ntpd.service
```



```
● ntpd.service - Network Time Service
   Loaded: loaded (/usr/lib/systemd/system/ntpd.service; enabled; vendor preset: disabled)
   Active: active (running) since Wed 2018-05-23 08:26:50 UTC; 44s ago
  Process: 1983 ExecStart=/usr/sbin/ntpd -u ntp:ntp $OPTIONS (code=exited, status=0/SUCCESS)
 Main PID: 1984 (ntpd)
   CGroup: /system.slice/ntpd.service
           └─1984 /usr/sbin/ntpd -u ntp:ntp -g

May 23 08:26:50 center ntpd[1984]: Listen and drop on 1 v6wildcard :: UDP 123
May 23 08:26:50 center ntpd[1984]: Listen normally on 2 lo 127.0.0.1 UDP 123
May 23 08:26:50 center ntpd[1984]: Listen normally on 3 eth0 10.160.0.2 UDP 123
May 23 08:26:50 center ntpd[1984]: Listen normally on 4 lo ::1 UDP 123
May 23 08:26:50 center ntpd[1984]: Listen normally on 5 eth0 fe80::4001:aff:fea0:2 UDP 123
May 23 08:26:50 center ntpd[1984]: Listening on routing socket on fd #22 for interface updates
May 23 08:26:50 center ntpd[1984]: 0.0.0.0 c016 06 restart
May 23 08:26:50 center ntpd[1984]: 0.0.0.0 c012 02 freq_set kernel 0.000 PPM
May 23 08:26:50 center ntpd[1984]: 0.0.0.0 c011 01 freq_not_set
May 23 08:26:57 center ntpd[1984]: 0.0.0.0 c614 04 freq_mode
```

Ensure that the service starts automatically on reboot.

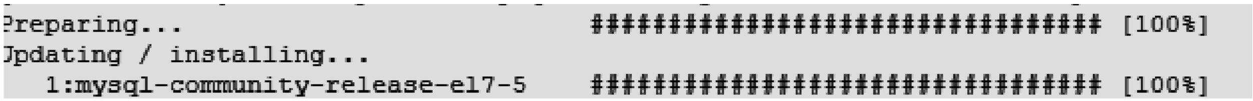
```
$ sudo systemctl enable ntpd.service
```

Configuring MySQL

Install MySQL

Download and add the repository, then update.

```
$ sudo yum install wget -y
$ wget http://repo.mysql.com/mysql-community-release-el7-5.noarch.rpm
$ sudo rpm -ivh mysql-community-release-el7-5.noarch.rpm
```



```
Preparing...                          ################################# [100%]
Updating / installing...
  1:mysql-community-release-el7-5     ################################# [100%]
```

```
$ sudo yum update -y
```

Install MySQL as usual and start the service. During installation, you will be asked if you want to accept the results from the rpm file's GPG verification. If no error or mismatch occurs, enter `y`.

```
$ sudo yum install mysql-server -y
```

Start the mysql Daemon

```
$ sudo systemctl start mysqld
```

Ensure that daemon stays active even after reboot

```
$ sudo systemctl enable mysqld
```

Installing mySQL JDBC driver

```
$ sudo yum install mysql-connector-java -y
```

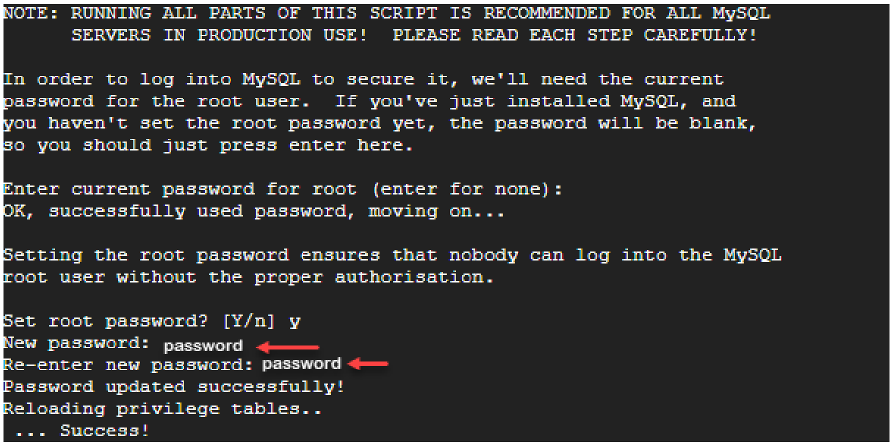
Configure the External Database

1. Run the Script

Service	Database	User
Cloudera Manager Server	scm	scmuser
Activity Monitor	amon	amonuser
Reports Manager	rman	rmanuser
Hue	hue	hueuser
Hive Metastore Server	metastore	hiveuser
Oozie	oozie	oozieuser
Data Analytics Studio	das	dasuser
Ranger	ranger	rangeradmin

```
$ vi mysql-setup.sql

CREATE DATABASE scm DEFAULT CHARACTER SET utf8;
GRANT ALL on scm.* TO 'scmuser'@'%' IDENTIFIED BY 'password';
CREATE DATABASE metastore DEFAULT CHARACTER SET utf8;
GRANT ALL on metastore.* TO 'hiveuser'@'%' IDENTIFIED BY 'password';
CREATE DATABASE amon DEFAULT CHARACTER SET utf8;
GRANT ALL on amon.* TO 'amonuser'@'%' IDENTIFIED BY 'password';
CREATE DATABASE rman DEFAULT CHARACTER SET utf8;
GRANT ALL on rman.* TO 'rmanuser'@'%' IDENTIFIED BY 'password';
CREATE DATABASE hue DEFAULT CHARACTER SET utf8;
GRANT ALL on hue.* TO 'hueuser'@'%' IDENTIFIED BY 'password';
CREATE DATABASE oozie DEFAULT CHARACTER SET utf8;
GRANT ALL on oozie.* TO 'oozieuser'@'%' IDENTIFIED BY 'password';
CREATE DATABASE das DEFAULT CHARACTER SET utf8;
GRANT ALL on das.* TO 'dasuser'@'%' IDENTIFIED BY 'password';
CREATE DATABASE ranger DEFAULT CHARACTER SET utf8;
GRANT ALL on ranger.* TO 'rangeradmin'@'%' IDENTIFIED BY 'password';
```

```
$ mysql -u root < mysql-setup.sql
$ sudo /usr/bin/mysql_secure_installation
```



```
NOTE: RUNNING ALL PARTS OF THIS SCRIPT IS RECOMMENDED FOR ALL MySQL
      SERVERS IN PRODUCTION USE!  PLEASE READ EACH STEP CAREFULLY!

In order to log into MySQL to secure it, we'll need the current
password for the root user.  If you've just installed MySQL, and
you haven't set the root password yet, the password will be blank,
so you should just press enter here.

Enter current password for root (enter for none):
OK, successfully used password, moving on...

Setting the root password ensures that nobody can log into the MySQL
root user without the proper authorisation.

Set root password? [Y/n] y
New password: password
Re-enter new password: password
Password updated successfully!
Reloading privilege tables..
 ... Success!
```

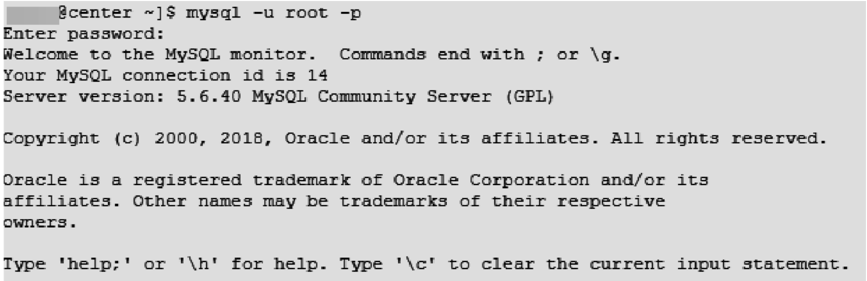
Set Password: password

Note: There is no current password for root in mysql so simply press Enter without entering password

2. Open the mysql Shell

```
$ mysql -u root -p
```

Enter the password: “password”



```
[redacted]@center ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 14
Server version: 5.6.40 MySQL Community Server (GPL)

Copyright (c) 2000, 2018, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
```

```
mysql > show databases;
```

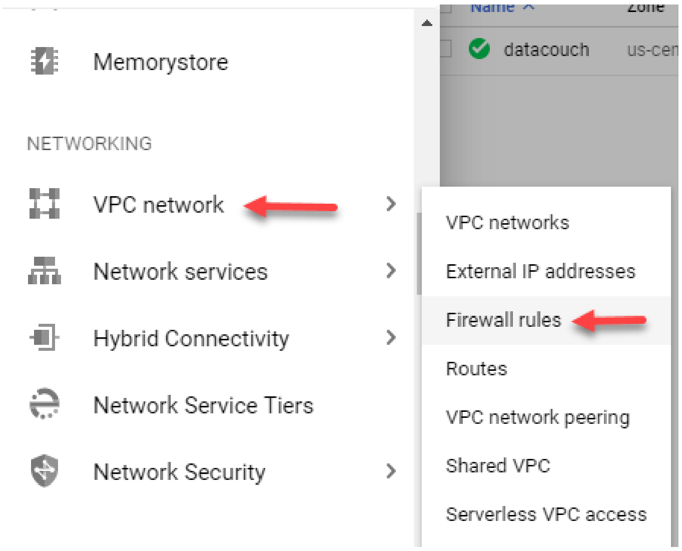
To exit the Shell

```
mysql > exit;
```

Set Firewall rule on GCP

Click on VPC network then click on Firewall rules

1. Go to the **FIREWALL RULES PAGE** page.



2. Click Create a firewall rule.

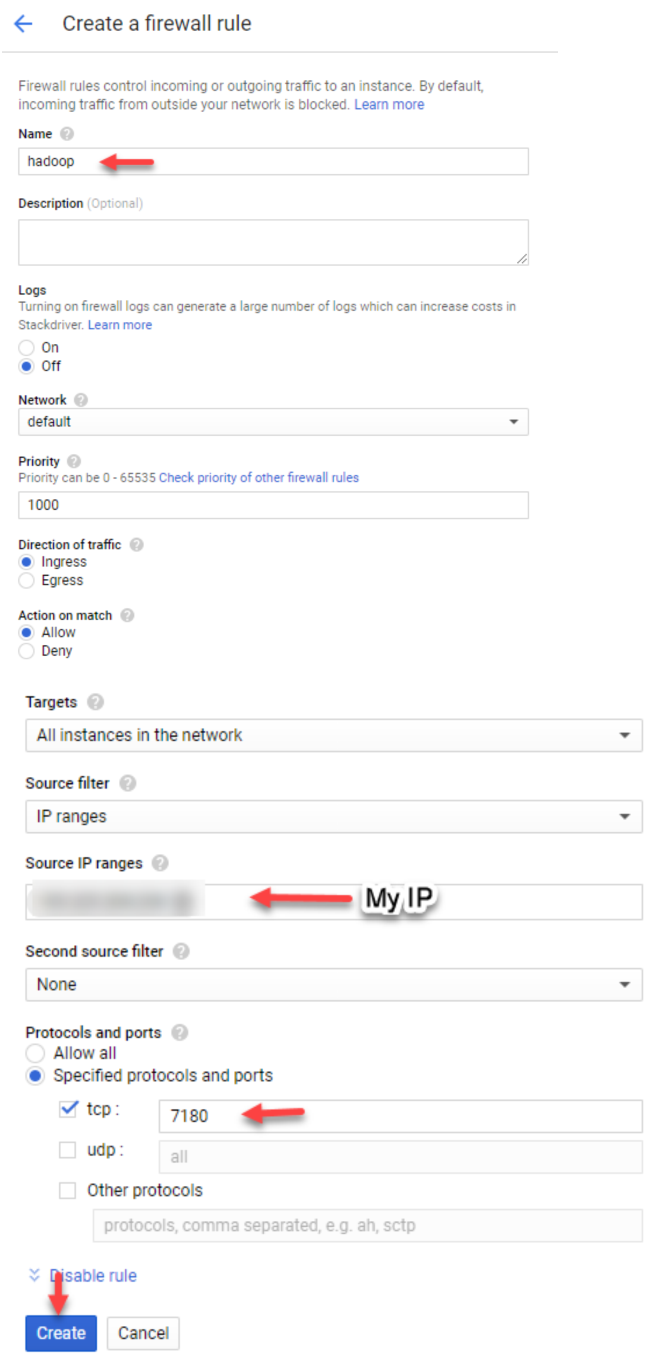
3. Populate the following fields:

   Name: hadoop

   Source filter: IP ranges.

   Source IP ranges: The peer network's IP address ranges to accept from the peer VPN gateway.

4. Assign you IP in Source IP ranges:



Create a firewall rule

Firewall rules control incoming or outgoing traffic to an instance. By default, incoming traffic from outside your network is blocked. Learn more

Name: hadoop

Description (Optional):

Logs
Turning on firewall logs can generate a large number of logs which can increase costs in Stackdriver. Learn more

○ On
● Off

Network: default

Priority
Priority can be 0 - 65535 Check priority of other firewall rules

1000

Direction of traffic
● Ingress
○ Egress

Action on match
● Allow
○ Deny

Targets: All instances in the network

Source filter: IP ranges

Source IP ranges: [redacted] My IP

Second source filter: None

Protocols and ports
○ Allow all
● Specified protocols and ports

☑ tcp : 7180
□ udp : all
□ Other protocols
protocols, comma separated, e.g. ah, sctp

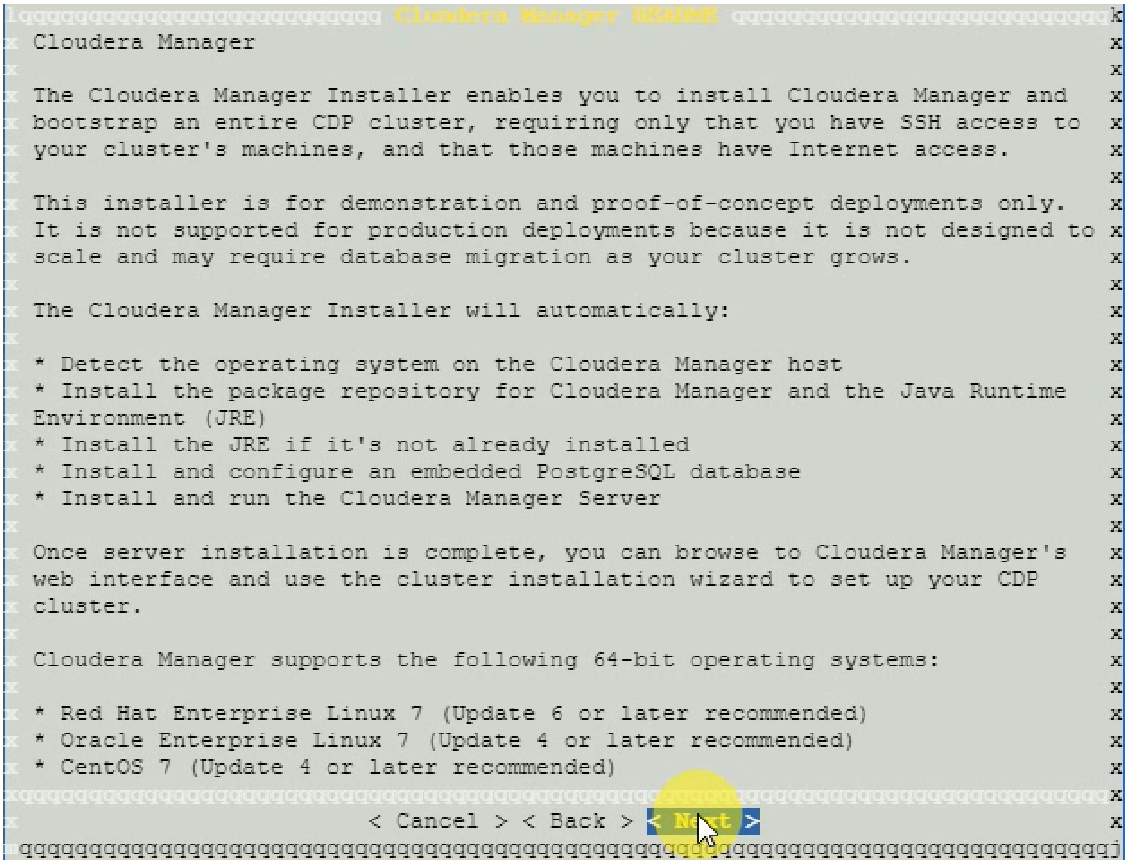
5. Specific port: tcp: 7180

Click Create.

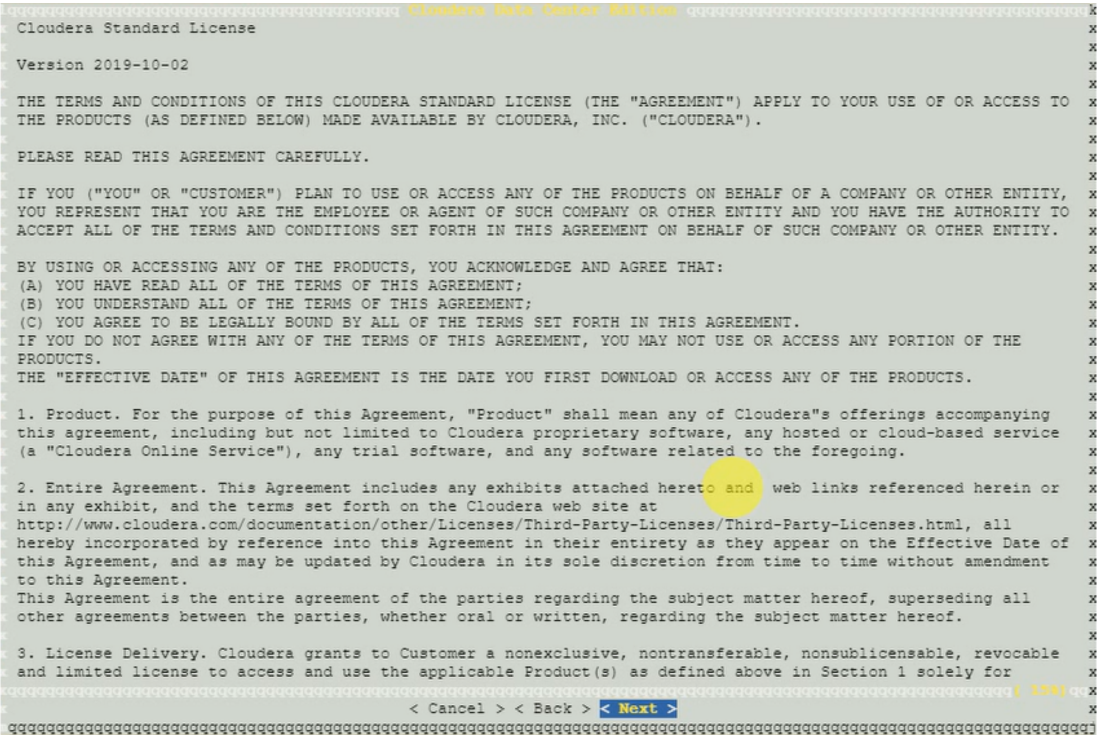
Cloudera Data Platform Installation

Installation instructions: Execute below instructions on Terminal to begin an automated CDP installation.

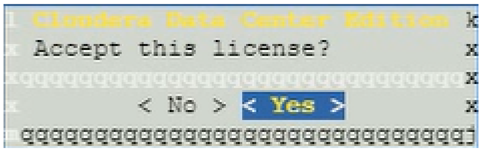
```
$ wget https://archive.cloudera.com/cm7/7.1.3/cloudera-manager-installer.bin

$ chmod u+x cloudera-manager-installer.bin

$ sudo ./cloudera-manager-installer.bin
```



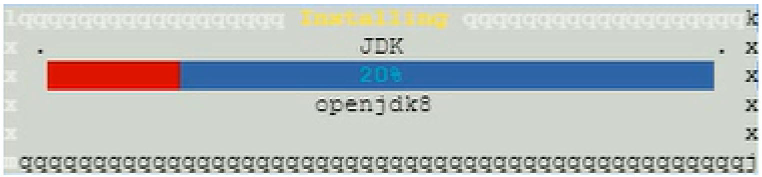
```
Cloudera Manager README

Cloudera Manager

The Cloudera Manager Installer enables you to install Cloudera Manager and
bootstrap an entire CDP cluster, requiring only that you have SSH access to
your cluster's machines, and that those machines have Internet access.

This installer is for demonstration and proof-of-concept deployments only.
It is not supported for production deployments because it is not designed to
scale and may require database migration as your cluster grows.

The Cloudera Manager Installer will automatically:

* Detect the operating system on the Cloudera Manager host
* Install the package repository for Cloudera Manager and the Java Runtime
Environment (JRE)
* Install the JRE if it's not already installed
* Install and configure an embedded PostgreSQL database
* Install and run the Cloudera Manager Server

Once server installation is complete, you can browse to Cloudera Manager's
web interface and use the cluster installation wizard to set up your CDP
cluster.

Cloudera Manager supports the following 64-bit operating systems:

* Red Hat Enterprise Linux 7 (Update 6 or later recommended)
* Oracle Enterprise Linux 7 (Update 4 or later recommended)
* CentOS 7 (Update 4 or later recommended)
```



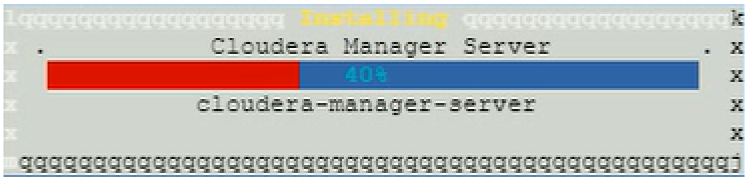
```
Cloudera Data Center Edition

Cloudera Standard License

Version 2019-10-02

THE TERMS AND CONDITIONS OF THIS CLOUDERA STANDARD LICENSE (THE "AGREEMENT") APPLY TO YOUR USE OF OR ACCESS TO
THE PRODUCTS (AS DEFINED BELOW) MADE AVAILABLE BY CLOUDERA, INC. ("CLOUDERA").

PLEASE READ THIS AGREEMENT CAREFULLY.

IF YOU ("YOU" OR "CUSTOMER") PLAN TO USE OR ACCESS ANY OF THE PRODUCTS ON BEHALF OF A COMPANY OR OTHER ENTITY,
YOU REPRESENT THAT YOU ARE THE EMPLOYEE OR AGENT OF SUCH COMPANY OR OTHER ENTITY AND YOU HAVE THE AUTHORITY TO
ACCEPT ALL OF THE TERMS AND CONDITIONS SET FORTH IN THIS AGREEMENT ON BEHALF OF SUCH COMPANY OR OTHER ENTITY.

BY USING OR ACCESSING ANY OF THE PRODUCTS, YOU ACKNOWLEDGE AND AGREE THAT:
(A) YOU HAVE READ ALL OF THE TERMS OF THIS AGREEMENT;
(B) YOU UNDERSTAND ALL OF THE TERMS OF THIS AGREEMENT;
(C) YOU AGREE TO BE LEGALLY BOUND BY ALL OF THE TERMS SET FORTH IN THIS AGREEMENT.
IF YOU DO NOT AGREE WITH ANY OF THE TERMS OF THIS AGREEMENT, YOU MAY NOT USE OR ACCESS ANY PORTION OF THE
PRODUCTS.
THE "EFFECTIVE DATE" OF THIS AGREEMENT IS THE DATE YOU FIRST DOWNLOAD OR ACCESS ANY OF THE PRODUCTS.

1. Product. For the purpose of this Agreement, "Product" shall mean any of Cloudera"s offerings accompanying
this agreement, including but not limited to Cloudera proprietary software, any hosted or cloud-based service
(a "Cloudera Online Service"), any trial software, and any software related to the foregoing.

2. Entire Agreement. This Agreement includes any exhibits attached hereto and web links referenced herein or
in any exhibit, and the terms set forth on the Cloudera web site at
http://www.cloudera.com/documentation/other/Licenses/Third-Party-Licenses/Third-Party-Licenses.html, all
hereby incorporated by reference into this Agreement in their entirety as they appear on the Effective Date of
this Agreement, and as may be updated by Cloudera in its sole discretion from time to time without amendment
to this Agreement.
This Agreement is the entire agreement of the parties regarding the subject matter hereof, superseding all
other agreements between the parties, whether oral or written, regarding the subject matter hereof.

3. License Delivery. Cloudera grants to Customer a nonexclusive, nontransferable, nonsublicensable, revocable
and limited license to access and use the applicable Product(s) as defined above in Section 1 solely for
                                                                                                  ( 15%)
                                          < Cancel > < Back > < Next >
```



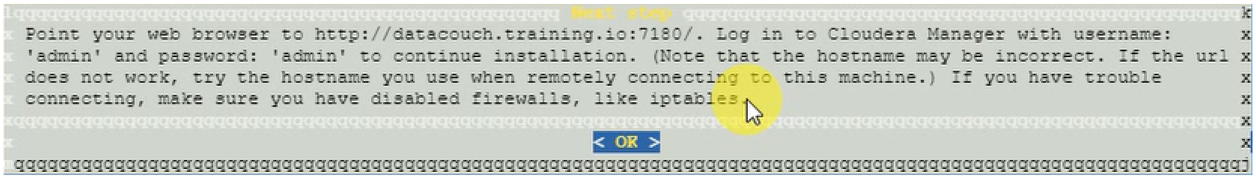
```
 Cloudera Data Center Edition
 Accept this license?

         < No > < Yes >
```



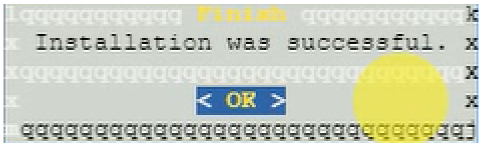
```
                   Installing
 .                   JDK                   .
                     20%
                   openjdk8
```



```
                   Installing
 .           Cloudera Manager Server        .
                     40%
            cloudera-manager-server
```



```
                                   Next step
 Point your web browser to http://datacouch.training.io:7180/. Log in to Cloudera Manager with username:
 'admin' and password: 'admin' to continue installation. (Note that the hostname may be incorrect. If the url
 does not work, try the hostname you use when remotely connecting to this machine.) If you have trouble
 connecting, make sure you have disabled firewalls, like iptables.
                                     < OK >
```



```
            Finish
 Installation was successful.
           < OK >
```

Working with Cloudera Manager

1. Start Cloudera Manager Server:

```
sudo systemctl status cloudera-scm-server
```

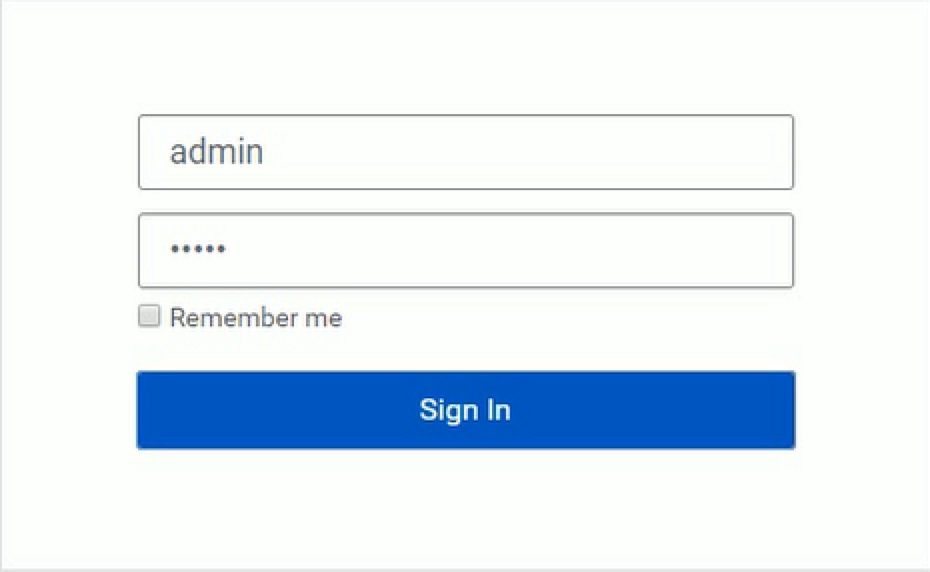
   Note: If it's not running then execute the below command to run the **cloudera-scm-server**

```
sudo systemctl start cloudera-scm-server
```

2. In a web browser, go to ***http://<server_host>:7180***, where <server_host> is the FQDN or IP address of the host where the Cloudera Manager Server is running.

   Note: If "Unable to Connect" message appears, it means that the Cloudera Manager server has not yet fully started. Wait for a few seconds, and then attempt to connect again.

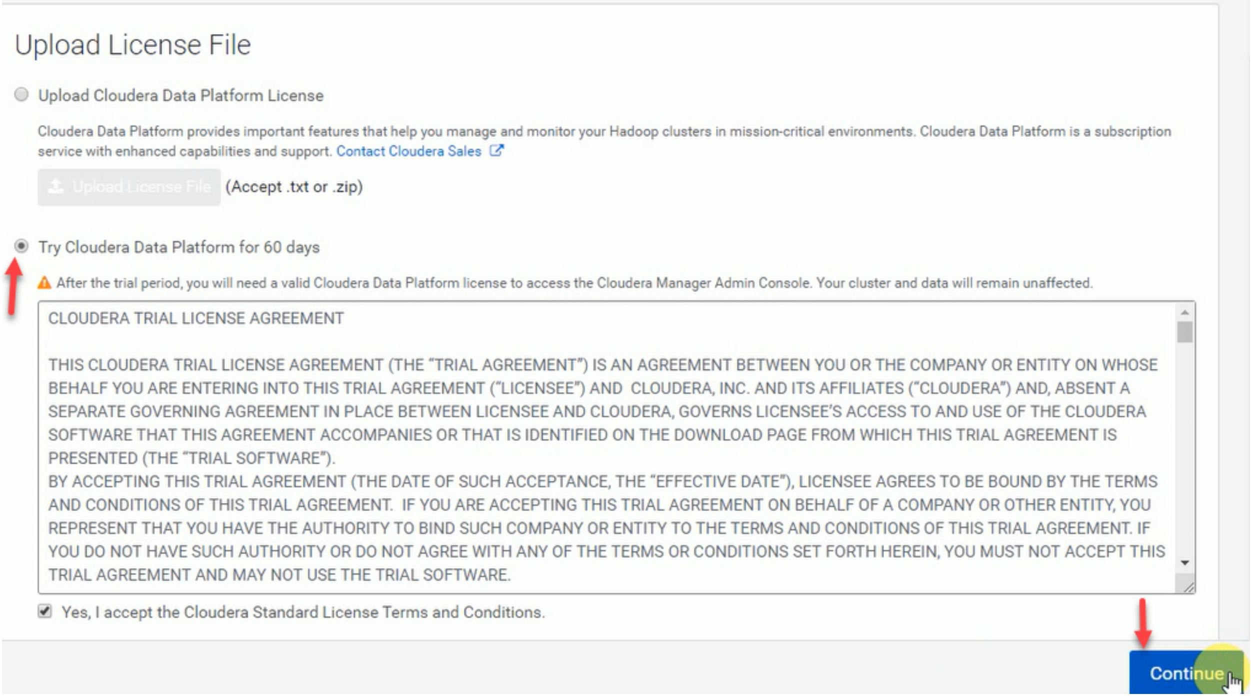
3. Log into Cloudera Manager Admin Console. The default credentials are:
   Username: admin
   Password: admin



Select Edition

On the Select Edition page, you can select the edition of Cloudera Manager to install and, optionally, install a license:

1. Choose which edition to install:
    a. Cloudera Data Platform - Data Center Edition Trial, which does not require a license file, but expires after 60 days.
    b. Cloudera Data Platform Data Center Edition
2. If you choose the CDP Data Center Edition Trial, you can upgrade the license at a later time.
3. Accept the License
4. Click Continue to proceed with the installation.



Upload License File

○ Upload Cloudera Data Platform License

Cloudera Data Platform provides important features that help you manage and monitor your Hadoop clusters in mission-critical environments. Cloudera Data Platform is a subscription service with enhanced capabilities and support. Contact Cloudera Sales

(Accept .txt or .zip)

● Try Cloudera Data Platform for 60 days

⚠ After the trial period, you will need a valid Cloudera Data Platform license to access the Cloudera Manager Admin Console. Your cluster and data will remain unaffected.

CLOUDERA TRIAL LICENSE AGREEMENT

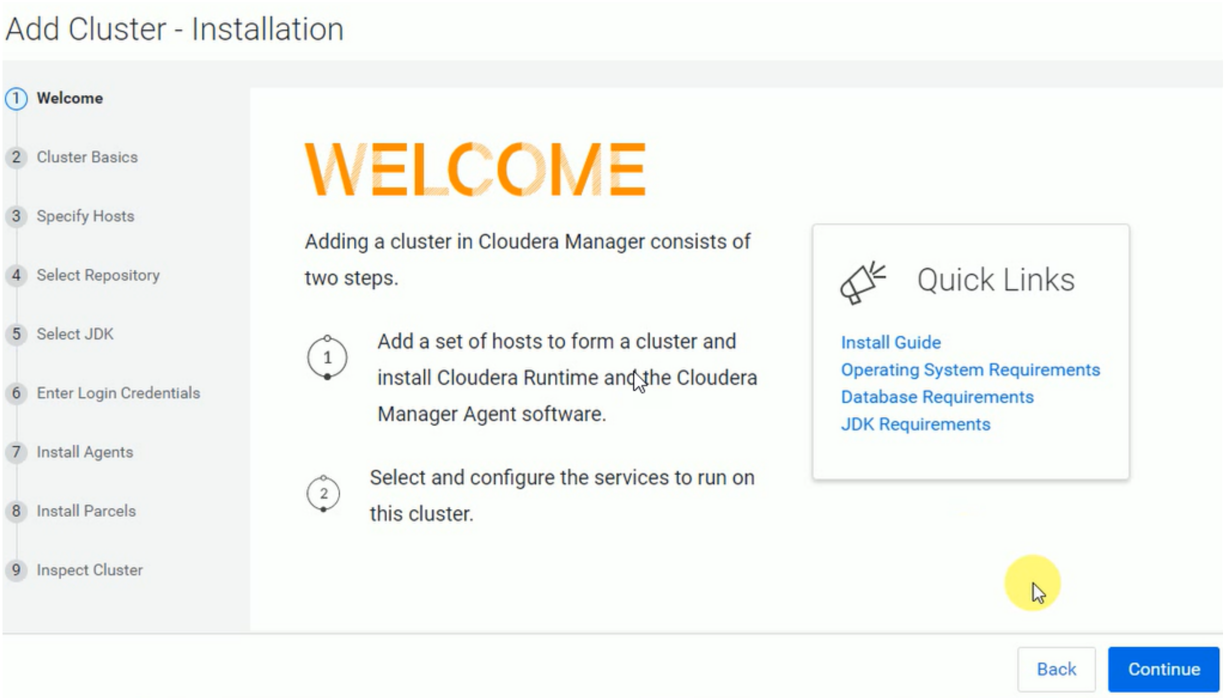
THIS CLOUDERA TRIAL LICENSE AGREEMENT (THE "TRIAL AGREEMENT") IS AN AGREEMENT BETWEEN YOU OR THE COMPANY OR ENTITY ON WHOSE BEHALF YOU ARE ENTERING INTO THIS TRIAL AGREEMENT ("LICENSEE") AND CLOUDERA, INC. AND ITS AFFILIATES ("CLOUDERA") AND, ABSENT A SEPARATE GOVERNING AGREEMENT IN PLACE BETWEEN LICENSEE AND CLOUDERA, GOVERNS LICENSEE'S ACCESS TO AND USE OF THE CLOUDERA SOFTWARE THAT THIS AGREEMENT ACCOMPANIES OR THAT IS IDENTIFIED ON THE DOWNLOAD PAGE FROM WHICH THIS TRIAL AGREEMENT IS PRESENTED (THE "TRIAL SOFTWARE").
BY ACCEPTING THIS TRIAL AGREEMENT (THE DATE OF SUCH ACCEPTANCE, THE "EFFECTIVE DATE"), LICENSEE AGREES TO BE BOUND BY THE TERMS AND CONDITIONS OF THIS TRIAL AGREEMENT. IF YOU ARE ACCEPTING THIS TRIAL AGREEMENT ON BEHALF OF A COMPANY OR OTHER ENTITY, YOU REPRESENT THAT YOU HAVE THE AUTHORITY TO BIND SUCH COMPANY OR ENTITY TO THE TERMS AND CONDITIONS OF THIS TRIAL AGREEMENT. IF YOU DO NOT HAVE SUCH AUTHORITY OR DO NOT AGREE WITH ANY OF THE TERMS OR CONDITIONS SET FORTH HEREIN, YOU MUST NOT ACCEPT THIS TRIAL AGREEMENT AND MAY NOT USE THE TRIAL SOFTWARE.

☑ Yes, I accept the Cloudera Standard License Terms and Conditions.

Welcome (Add Cluster - Installation)

The Welcome page of the Add Cluster - Installation wizard provides a brief overview of the installation and configuration procedure, as well as some links to relevant documentation.

Click Continue to proceed with the installation.



Add Cluster - Installation

WELCOME

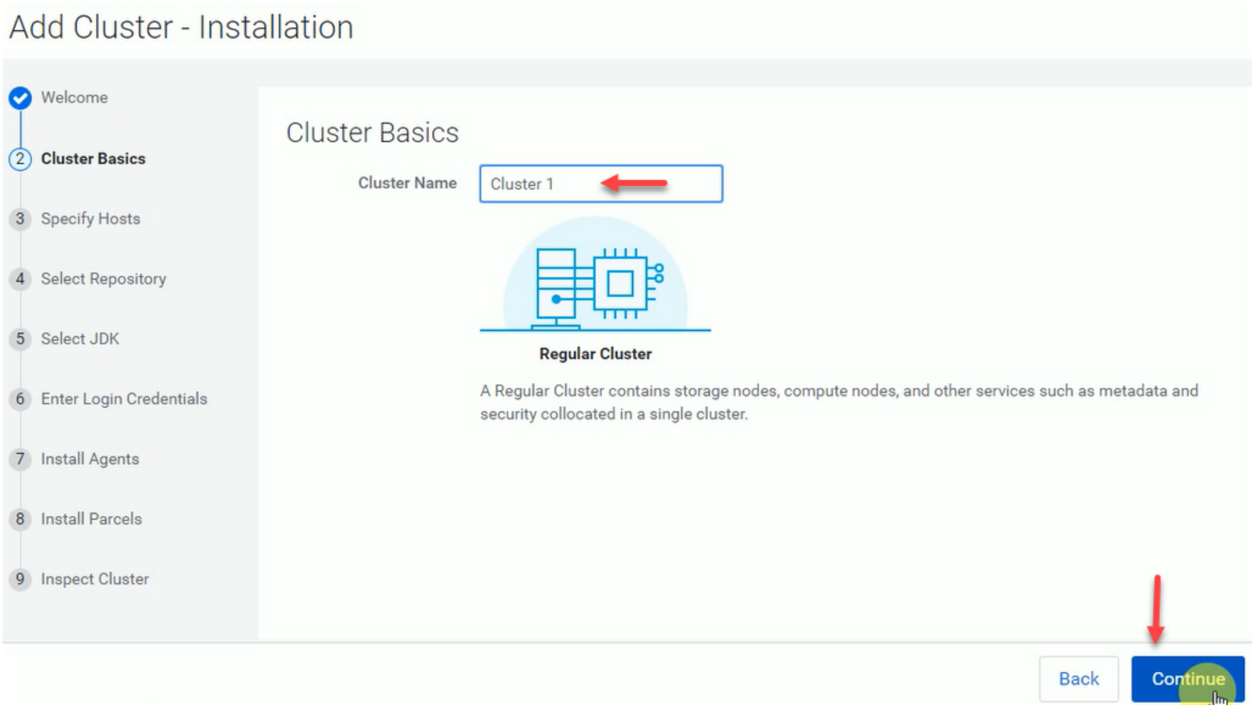
Adding a cluster in Cloudera Manager consists of two steps.

1. Add a set of hosts to form a cluster and install Cloudera Runtime and the Cloudera Manager Agent software.
2. Select and configure the services to run on this cluster.

Cluster Basics

The Cluster Basics page allows you to specify the Cluster Name.

Enter a cluster name and then click Continue.



Add Cluster - Installation

Cluster Basics

Cluster Name: Cluster 1

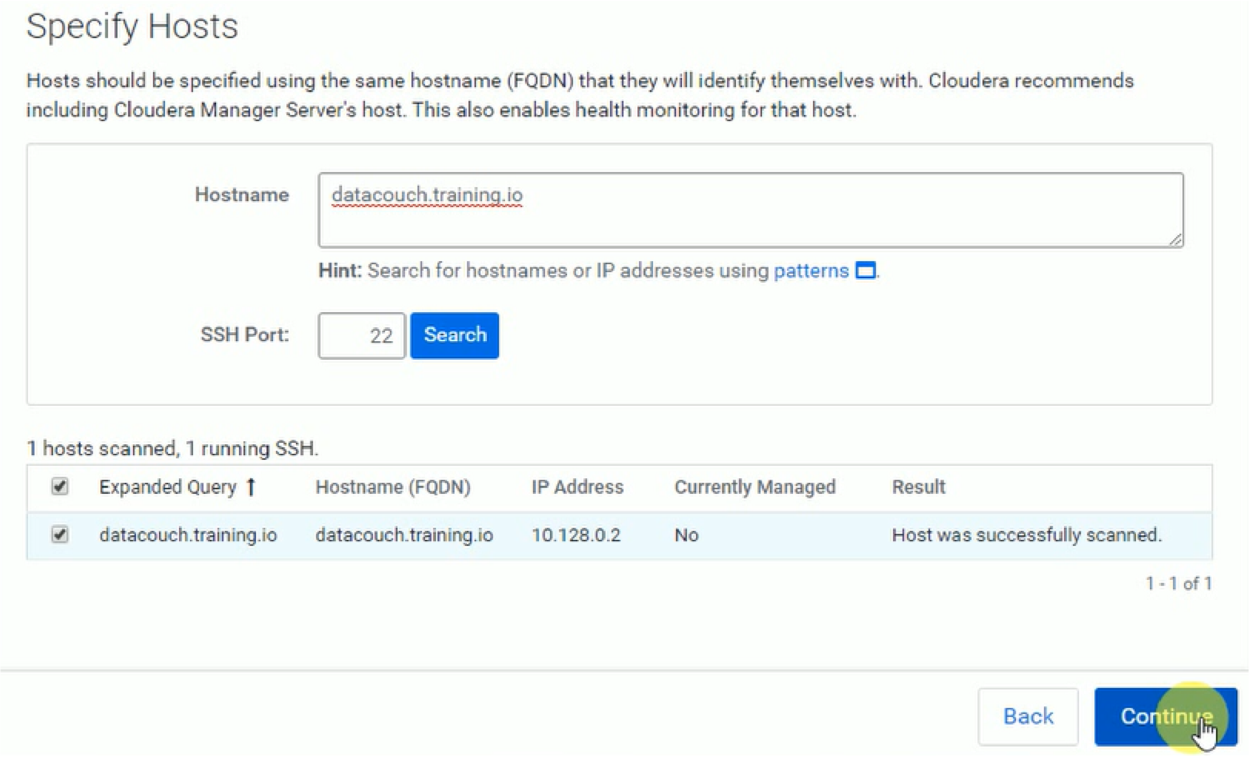
Regular Cluster

A Regular Cluster contains storage nodes, compute nodes, and other services such as metadata and security collocated in a single cluster.

Specify Hosts

Choose which hosts will run Runtime and other managed services.

1. The **“Specify hosts”** page appears. Type the hostnames of a machine: and Click Search.
2. Click Continue.



Specify Hosts

Hosts should be specified using the same hostname (FQDN) that they will identify themselves with. Cloudera recommends including Cloudera Manager Server's host. This also enables health monitoring for that host.

Hostname: datacouch.training.io

Hint: Search for hostnames or IP addresses using patterns.

SSH Port: 22

1 hosts scanned, 1 running SSH.

☑	Expanded Query ↑	Hostname (FQDN)	IP Address	Currently Managed	Result
☑	datacouch.training.io	datacouch.training.io	10.128.0.2	No	Host was successfully scanned.

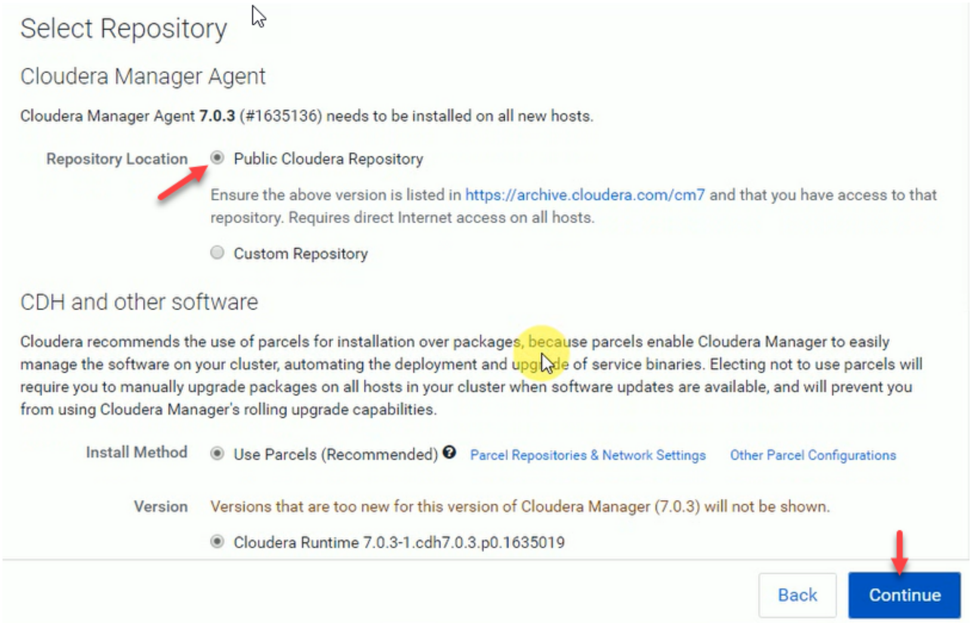
1 - 1 of 1

Select Repository

The Select Repository page allows you to specify repositories for Cloudera Manager Agent and CDH and other software.

In the Cloudera Manager Agent section:

1. Select either Public Cloudera Repository or Custom Repository for the Cloudera Manager Agent software.
2. If you select Custom Repository, do not include the operating system-specific paths in the URL. For instructions on setting up a custom repository, see Configuring a Local Package Repository.



Select Repository

Cloudera Manager Agent

Cloudera Manager Agent **7.0.3** (#1635136) needs to be installed on all new hosts.

Repository Location ● Public Cloudera Repository

Ensure the above version is listed in https://archive.cloudera.com/cm7 and that you have access to that repository. Requires direct Internet access on all hosts.

○ Custom Repository

CDH and other software

Cloudera recommends the use of parcels for installation over packages, because parcels enable Cloudera Manager to easily manage the software on your cluster, automating the deployment and upgrade of service binaries. Electing not to use parcels will require you to manually upgrade packages on all hosts in your cluster when software updates are available, and will prevent you from using Cloudera Manager's rolling upgrade capabilities.

Install Method ● Use Parcels (Recommended)

Version: Versions that are too new for this version of Cloudera Manager (7.0.3) will not be shown.

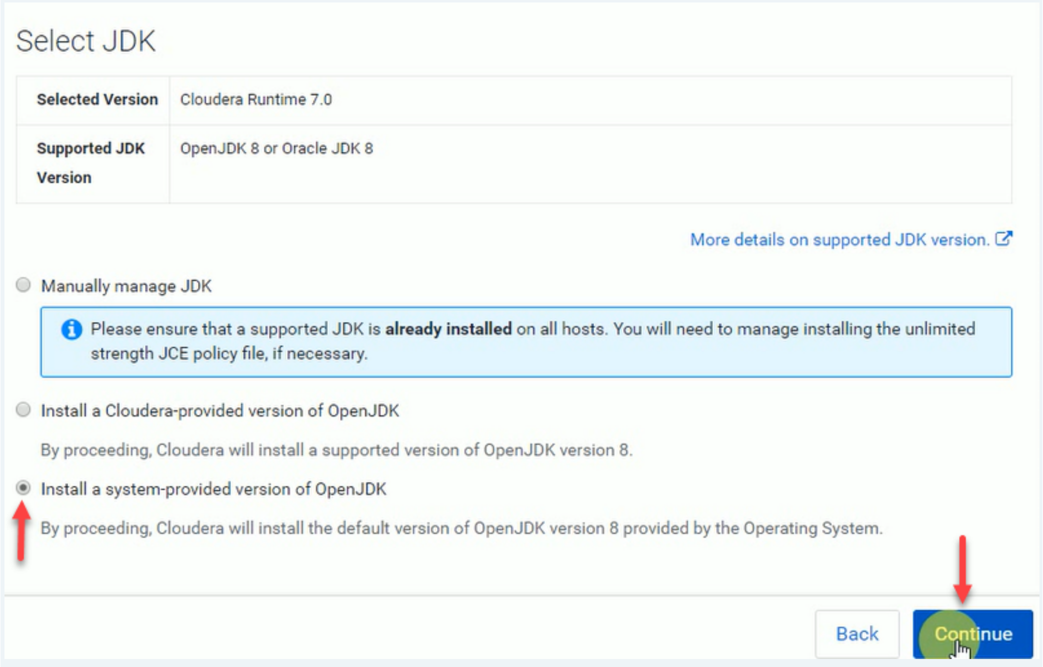
● Cloudera Runtime 7.0.3-1.cdh7.0.3.p0.1635019

Parcel Repositories & Network Settings
Other Parcel Configurations
Back
Continue

Click Continue.

Accept JDK License

Select **Install a system-provided version of OpenJDK**



Select JDK

Selected Version	Cloudera Runtime 7.0
Supported JDK Version	OpenJDK 8 or Oracle JDK 8

○ Manually manage JDK

Please ensure that a supported JDK is **already installed** on all hosts. You will need to manage installing the unlimited strength JCE policy file, if necessary.

○ Install a Cloudera-provided version of OpenJDK

By proceeding, Cloudera will install a supported version of OpenJDK version 8.

● Install a system-provided version of OpenJDK

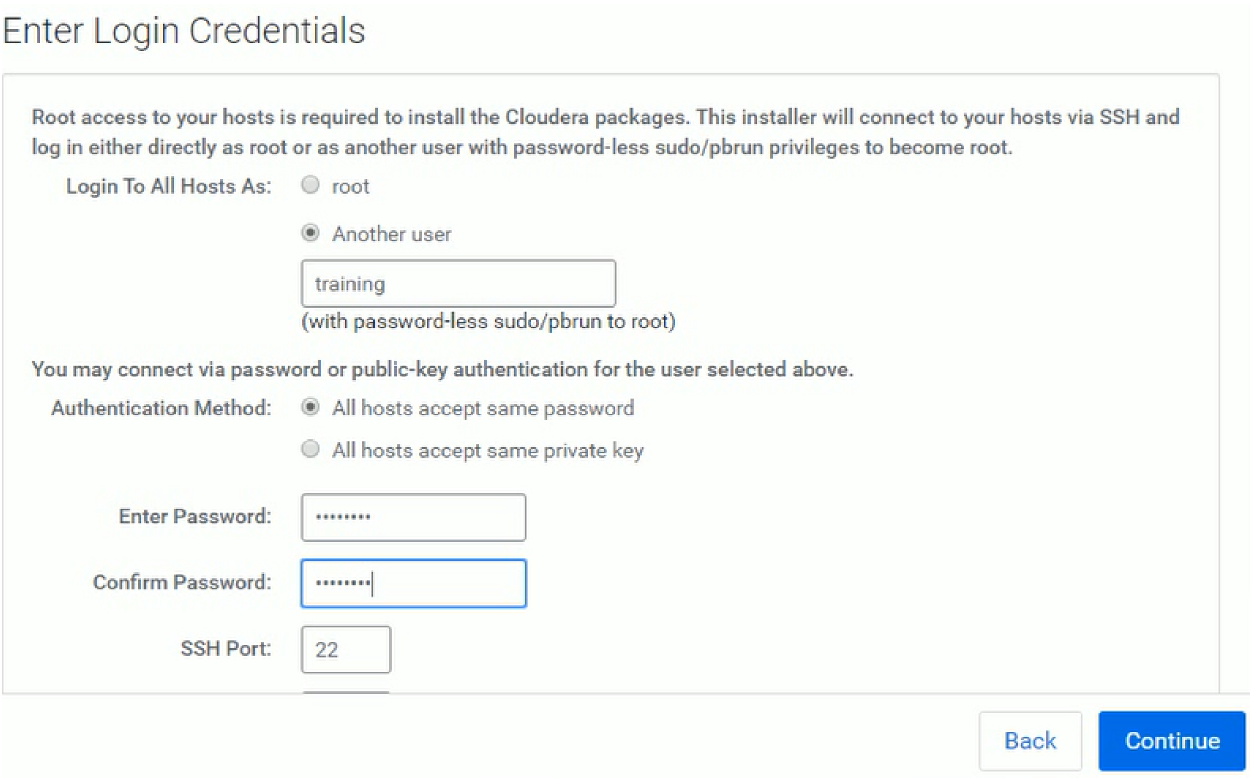
By proceeding, Cloudera will install the default version of OpenJDK version 8 provided by the Operating System.

Click Continue.

Enter Login Credentials

1. Select Another user and enter the username for an account that has password-less sudo privileges.

   Another user: training
   Password: password



Enter Login Credentials

Root access to your hosts is required to install the Cloudera packages. This installer will connect to your hosts via SSH and log in either directly as root or as another user with password-less sudo/pbrun privileges to become root.

Login To All Hosts As: ○ root
● Another user
training
(with password-less sudo/pbrun to root)

You may connect via password or public-key authentication for the user selected above.

Authentication Method: ● All hosts accept same password
○ All hosts accept same private key

Enter Password: ••••••••

Confirm Password: ••••••••

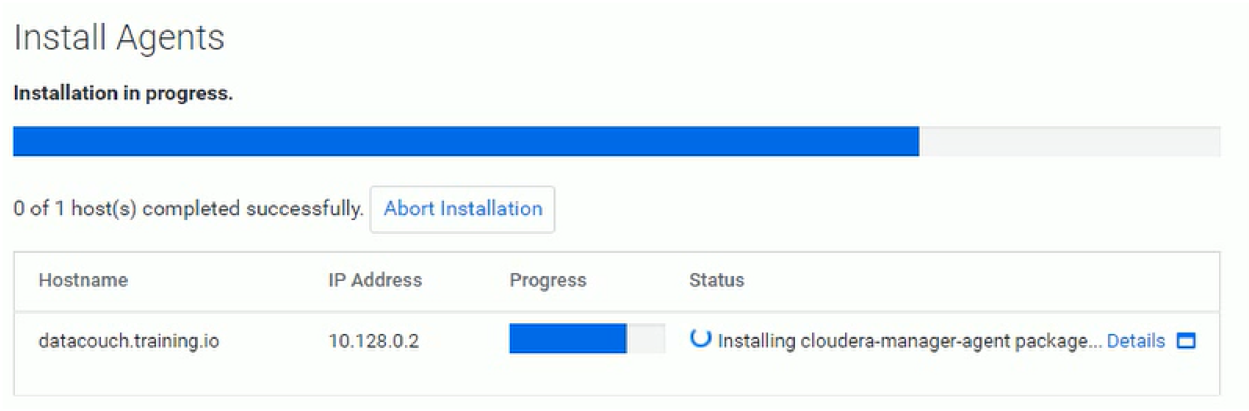
SSH Port: 22

2. Click Continue.

Install Agents

The Install Agents page displays the progress of the installation. You can click on the Details link for any host to view the installation log. If the installation is stalled, you can click the Abort Installation button to cancel the installation and then view the installation logs to troubleshoot the problem.

If the installation fails on any hosts, you can click the Retry Failed Hosts to retry all failed hosts, or you can click the Retry link on a specific host.

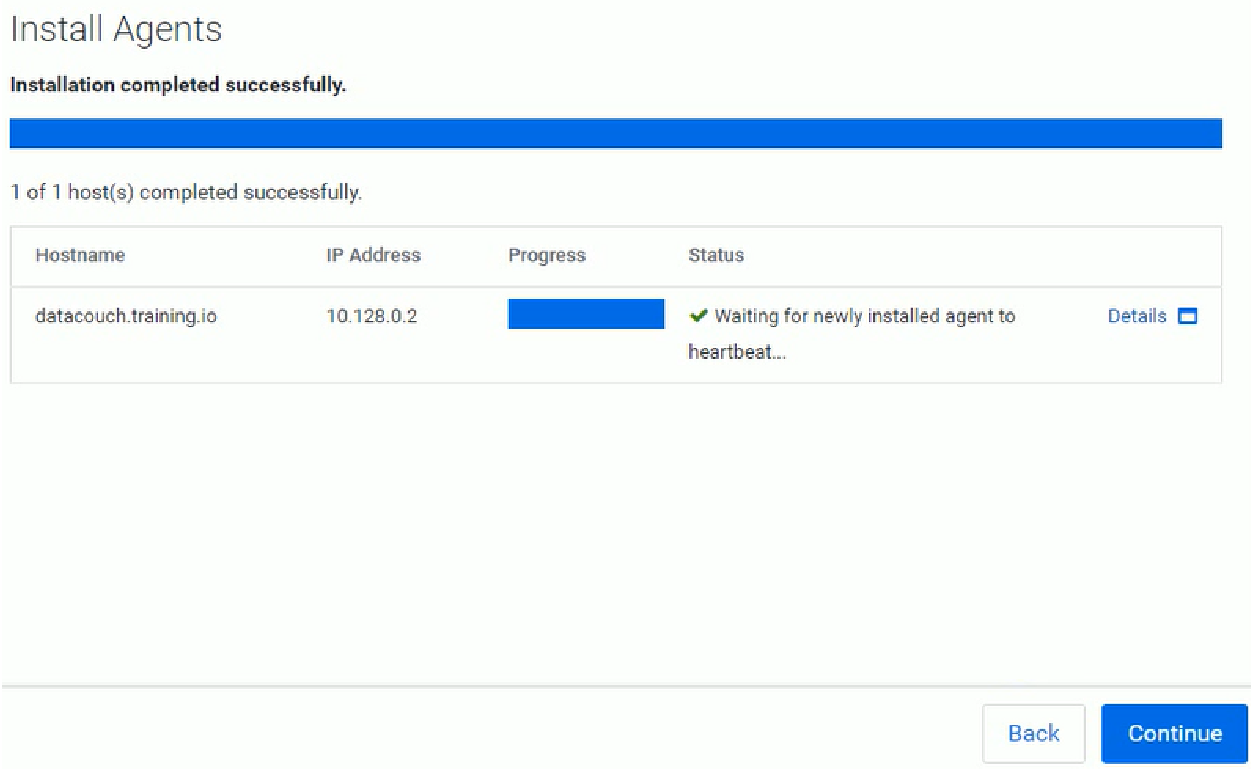


Install Agents

Installation in progress.

0 of 1 host(s) completed successfully.

Hostname	IP Address	Progress	Status
datacouch.training.io	10.128.0.2		Installing cloudera-manager-agent package... Details



Install Agents

Installation completed successfully.

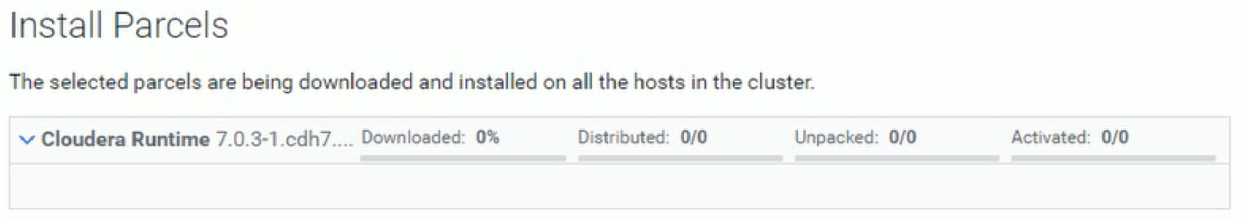
1 of 1 host(s) completed successfully.

Hostname	IP Address	Progress	Status
datacouch.training.io	10.128.0.2		✓ Waiting for newly installed agent to heartbeat... Details

After installing the Cloudera Manager Agent on all hosts, click Continue.

Install Parcels

After the parcels are downloaded, progress bars appear representing each cluster host. You can click on an individual progress bar for details about that host.



Install Parcels

The selected parcels are being downloaded and installed on all the hosts in the cluster.

Cloudera Runtime 7.0.3-1.cdh7....	Downloaded: 0%	Distributed: 0/0	Unpacked: 0/0	Activated: 0/0

After the installation is complete, click Continue.

Inspect Cluster

The Inspect Cluster page provides a tool for inspecting network performance as well as the Host Inspector to search for common configuration problems. Cloudera recommends that you run the inspectors sequentially:

1. Run the Inspect Network Performance tool.
2. After the network inspector completes, click Show Inspector Results to view the results in a new tab.
3. Address any reported issues, and click Run Again (if applicable).
4. Click Inspect Hosts to run the Host Inspector utility.
5. After the host inspector completes, click Show Inspector Results to view the results in a new tab.
6. Address any reported issues, and click Run Again (if applicable).

If the reported issues cannot be resolved in a timely manner, then select the radio button labeled I understand the risks, let me continue with cluster creation, and then click Continue.

This completes the Cluster Installation wizard and launches the Add Cluster - Configuration wizard.

Set Up a Cluster

Select Services

The Select Services page allows you to select the services you want to install and configure.

You can choose from:

Data Engineering

HDFS, YARN, YARN Queue Manager, Ranger, Atlas, Hive, Hive on Tez, Spark, Oozie, Zeppelin, Livy, and Hue

Data Mart

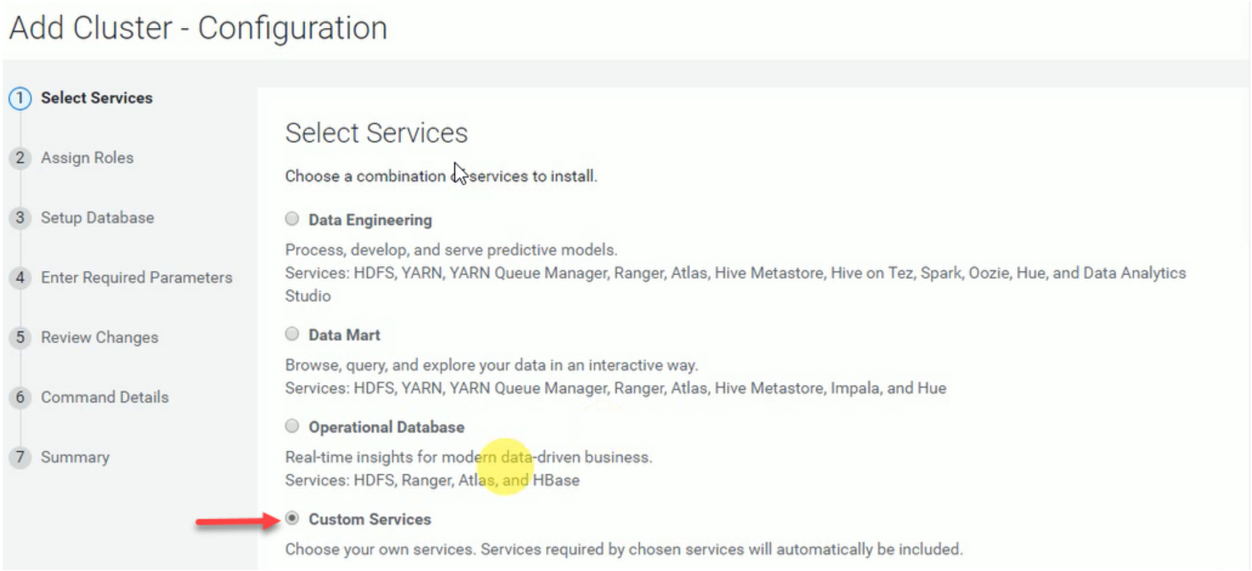
HDFS, YARN, YARN Queue Manager, Ranger, Atlas, Hive, Impala, and Hue

Operational Database

HDFS, Ranger, Atlas, and HBASE

Custom Services

Choose your own services. Services required by chosen services will automatically be included.



Select Services

Choose a combination of services to install.

○ **Data Engineering**

Process, develop, and serve predictive models.
Services: HDFS, YARN, YARN Queue Manager, Ranger, Atlas, Hive Metastore, Hive on Tez, Spark, Oozie, Hue, and Data Analytics Studio

○ **Data Mart**

Browse, query, and explore your data in an interactive way.
Services: HDFS, YARN, YARN Queue Manager, Ranger, Atlas, Hive Metastore, Impala, and Hue

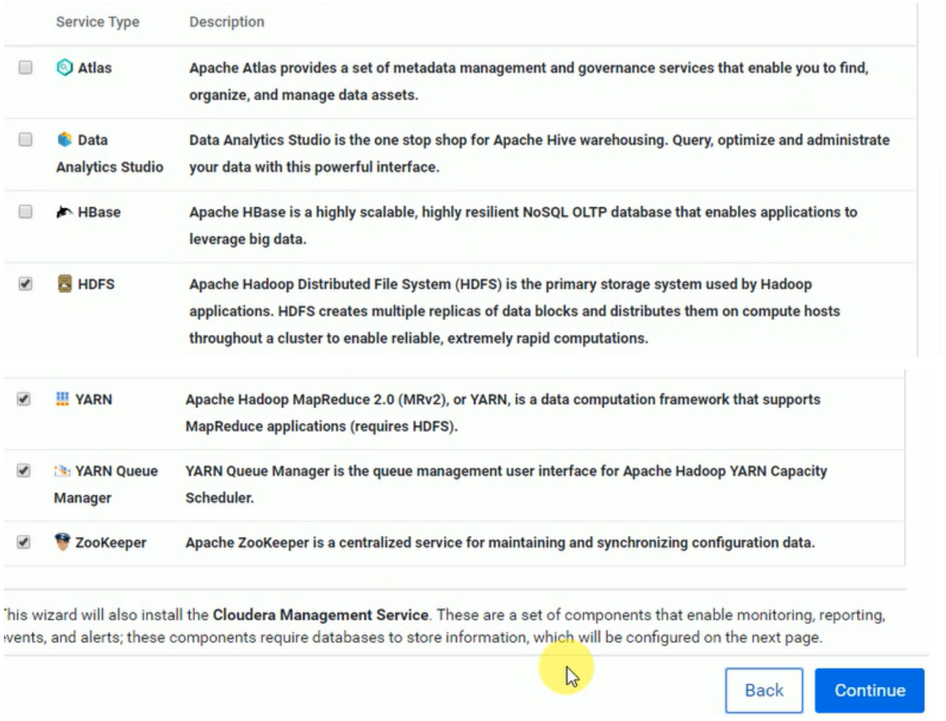
○ **Operational Database**

Real-time insights for modern data-driven business.
Services: HDFS, Ranger, Atlas, and HBase

● **Custom Services**

Choose your own services. Services required by chosen services will automatically be included.

Click on custom services and select your own services



	Service Type	Description
□	Atlas	Apache Atlas provides a set of metadata management and governance services that enable you to find, organize, and manage data assets.
□	Data Analytics Studio	Data Analytics Studio is the one stop shop for Apache Hive warehousing. Query, optimize and administrate your data with this powerful interface.
□	HBase	Apache HBase is a highly scalable, highly resilient NoSQL OLTP database that enables applications to leverage big data.
☑	HDFS	Apache Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute hosts throughout a cluster to enable reliable, extremely rapid computations.
☑	YARN	Apache Hadoop MapReduce 2.0 (MRv2), or YARN, is a data computation framework that supports MapReduce applications (requires HDFS).
☑	YARN Queue Manager	YARN Queue Manager is the queue management user interface for Apache Hadoop YARN Capacity Scheduler.
☑	ZooKeeper	Apache ZooKeeper is a centralized service for maintaining and synchronizing configuration data.

his wizard will also install the **Cloudera Management Service**. These are a set of components that enable monitoring, reporting, vents, and alerts; these components require databases to store information, which will be configured on the next page.

After selecting the services you want to add, click Continue. The Assign Roles page displays.

Customize Role Setups for the Cluster

HDFS:

NameNode: datacouch.training.io
Balancer: datacouch.training.io
Datanode: All Host

Cloudera Management Service:

Service Monitor: datacouch.training.io
Activity Monitor: datacouch.training.io
Host Monitor: datacouch.training.io
Reports Manager: datacouch.training.io
Event Server: datacouch.training.io
Alert Publisher: datacouch.training.io

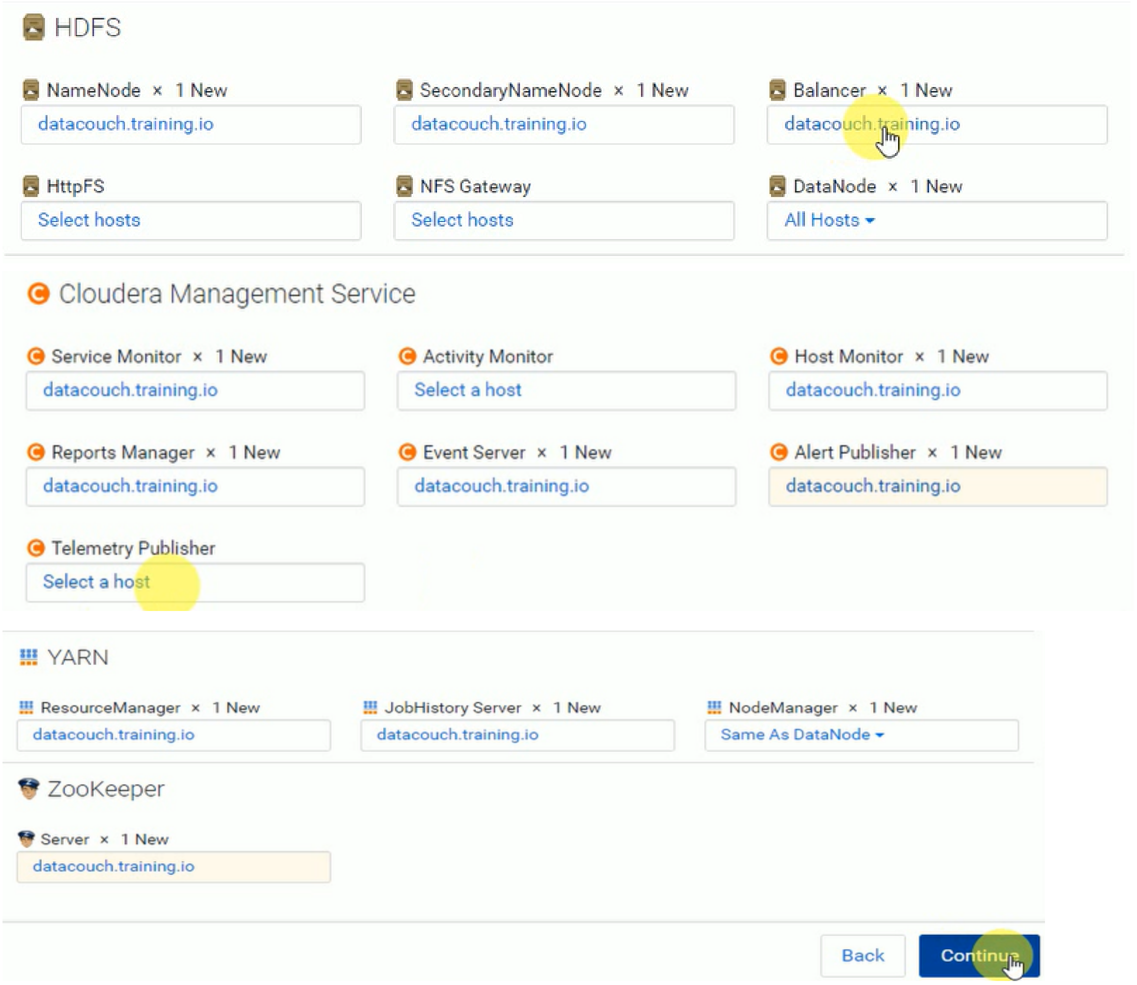
YARN:

ResourceManager: datacouch.training.io
Job History Server: datacouch.training.io

Assign Roles

The Assign Roles page suggests role assignments for the hosts in your cluster.

You can click on the hostname for a role to select a different host. You can also click the View By Host button to see all the roles assigned to a host.



HDFS

NameNode × 1 New: datacouch.training.io

SecondaryNameNode × 1 New: datacouch.training.io

Balancer × 1 New: datacouch.training.io

HttpFS: Select hosts

NFS Gateway: Select hosts

DataNode × 1 New: All Hosts

Cloudera Management Service

Service Monitor × 1 New: datacouch.training.io

Activity Monitor: Select a host

Host Monitor × 1 New: datacouch.training.io

Reports Manager × 1 New: datacouch.training.io

Event Server × 1 New: datacouch.training.io

Alert Publisher × 1 New: datacouch.training.io

Telemetry Publisher: Select a host

YARN

ResourceManager × 1 New: datacouch.training.io

JobHistory Server × 1 New: datacouch.training.io

NodeManager × 1 New: Same As DataNode

ZooKeeper

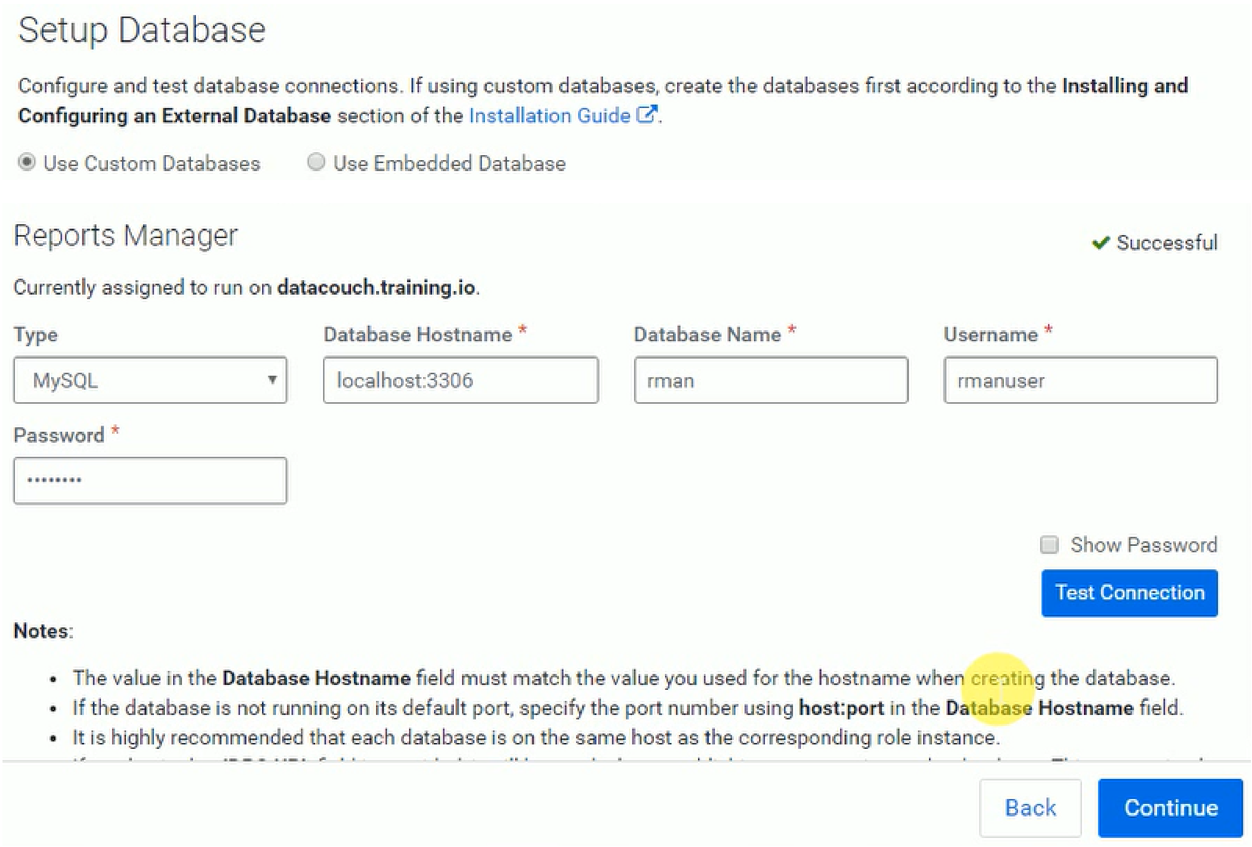
Server × 1 New: datacouch.training.io

Click on each service and select the host. After assigning all of the roles for your services, click Continue.

Setup Database

On the Setup Database page, you can enter the database hosts, names, usernames, and passwords you created.

Select the database type and enter the database name, username, and password for each service.



Setup Database

Configure and test database connections. If using custom databases, create the databases first according to the **Installing and Configuring an External Database** section of the Installation Guide.

● Use Custom Databases ○ Use Embedded Database

Reports Manager

✓ Successful

Currently assigned to run on **datacouch.training.io**.

Type	Database Hostname *	Database Name *	Username *
MySQL	localhost:3306	rman	rmanuser

Password * ••••••••

□ Show Password

Notes:

- The value in the **Database Hostname** field must match the value you used for the hostname when creating the database.
- If the database is not running on its default port, specify the port number using **host:port** in the **Database Hostname** field.
- It is highly recommended that each database is on the same host as the corresponding role instance.

Click Test Connection to validate the settings. If the connection is successful, a green checkmark and the word Successful appears next to each service. If there are any problems, the error is reported next to the service that failed to connect.

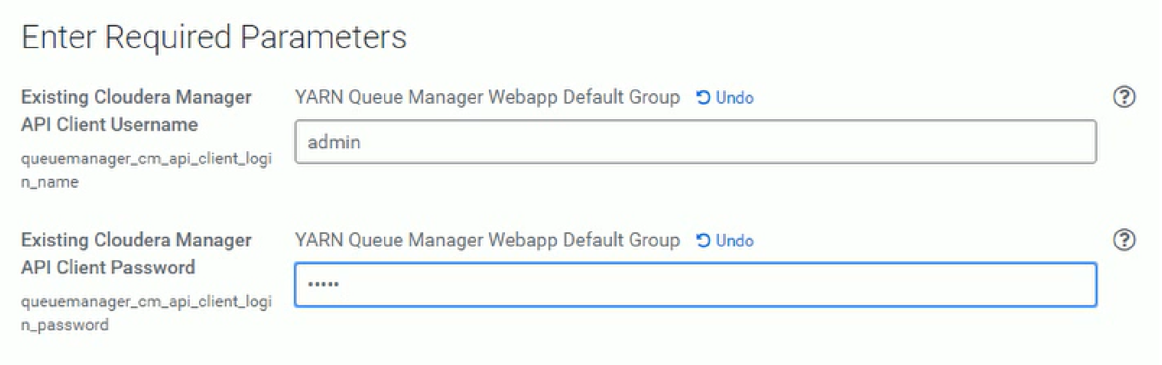
After verifying that each connection is successful, click Continue.

Yarn Queue Manager

In Yarn Queue Manager, in section “**Enter the Required Parameter”** specify username and password

queuemanager_cm_api_client_login_name: admin

queuemanager_cm_api_client_login_password: admin



Enter Required Parameters

Existing Cloudera Manager API Client Username
queuemanager_cm_api_client_login_name
YARN Queue Manager Webapp Default Group ⟲ Undo
admin

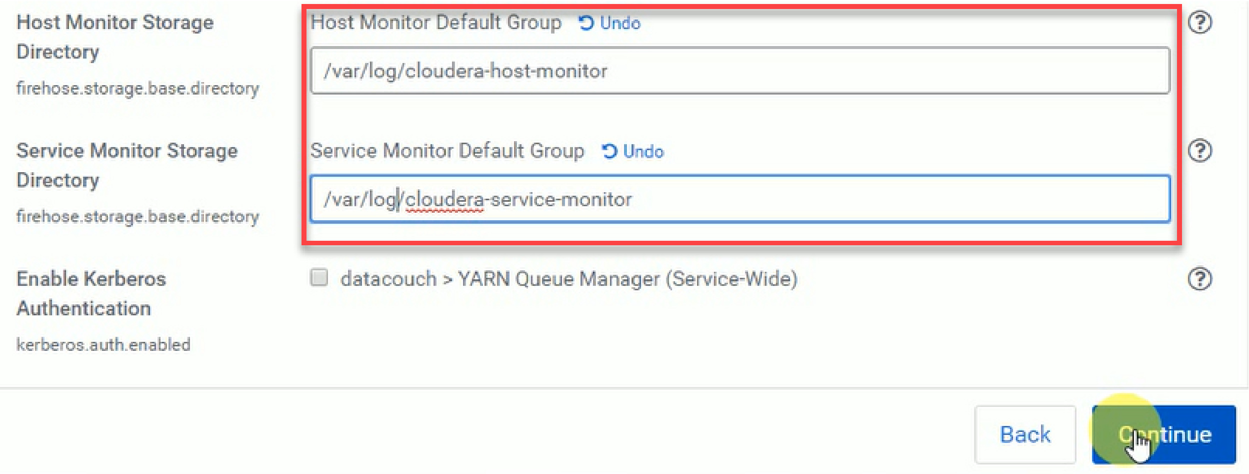
Existing Cloudera Manager API Client Password
queuemanager_cm_api_client_login_password
YARN Queue Manager Webapp Default Group ⟲ Undo
•••••

Click Continue.

In Review Changes

Host Monitor Storage Directory:- /var/**log**/cloudera-host-monitor

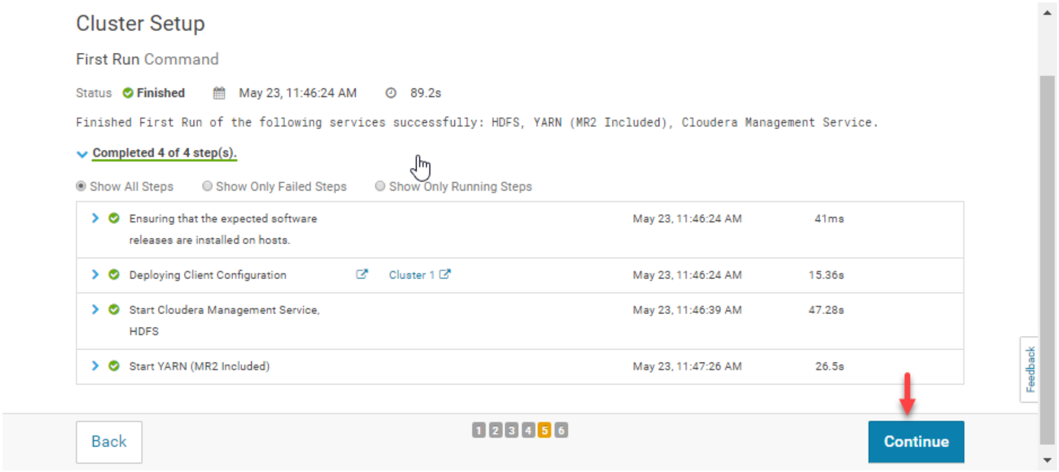
Service Monitor Storage Directory:- /var/**log**/cloudera-service-monitor



Host Monitor Storage Directory
firehose.storage.base.directory
Host Monitor Default Group ⟲ Undo
/var/log/cloudera-host-monitor

Service Monitor Storage Directory
firehose.storage.base.directory
Service Monitor Default Group ⟲ Undo
/var/log/cloudera-service-monitor

Enable Kerberos Authentication
kerberos.auth.enabled
□ datacouch > YARN Queue Manager (Service-Wide)



Cluster Setup

First Run Command

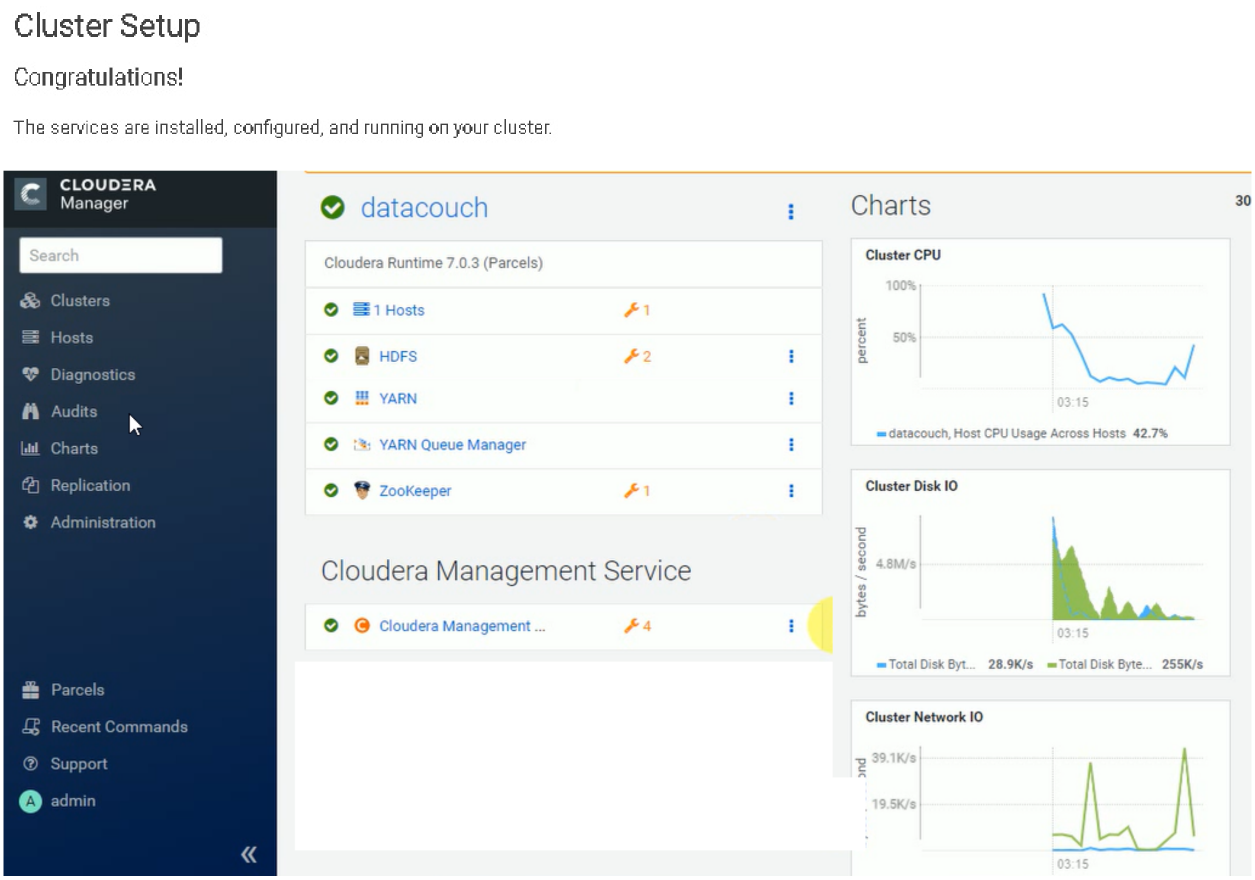
Status ✓ Finished May 23, 11:46:24 AM 89.2s

```
Finished First Run of the following services successfully: HDFS, YARN (MR2 Included), Cloudera Management Service.
```

Completed 4 of 4 step(s).

● Show All Steps ○ Show Only Failed Steps ○ Show Only Running Steps

✓ Ensuring that the expected software releases are installed on hosts.		May 23, 11:46:24 AM	41ms
✓ Deploying Client Configuration	Cluster 1	May 23, 11:46:24 AM	15.36s
✓ Start Cloudera Management Service, HDFS		May 23, 11:46:39 AM	47.28s
✓ Start YARN (MR2 Included)		May 23, 11:47:26 AM	26.5s



Cluster Setup

Congratulations!

The services are installed, configured, and running on your cluster.

As the hdfs user, create a home directory for the training user on HDFS and give the training user ownership of it's home directory.

```
$ sudo -u hdfs hdfs dfs -mkdir -p /user/training/
$ sudo -u hdfs hdfs dfs -chown training /user/training
```

Testing Your Hadoop Installation

You will now test the Hadoop installation in *center* machine by uploading some data from local machine.

1. Git clone **“hadoop-admin”** file

```
$ sudo yum install git -y
$ git clone https://github.com/datacouch16/hadoop-admin.git
```

2. Upload **hadoop-admin/data/sherlock.txt** / in HDFS.

```
$ cd hadoop-admin/data
$ hdfs dfs -mkdir data/
$ hdfs dfs -put sherlock.txt data/
```

3. Verify that the file is now in HDFS.

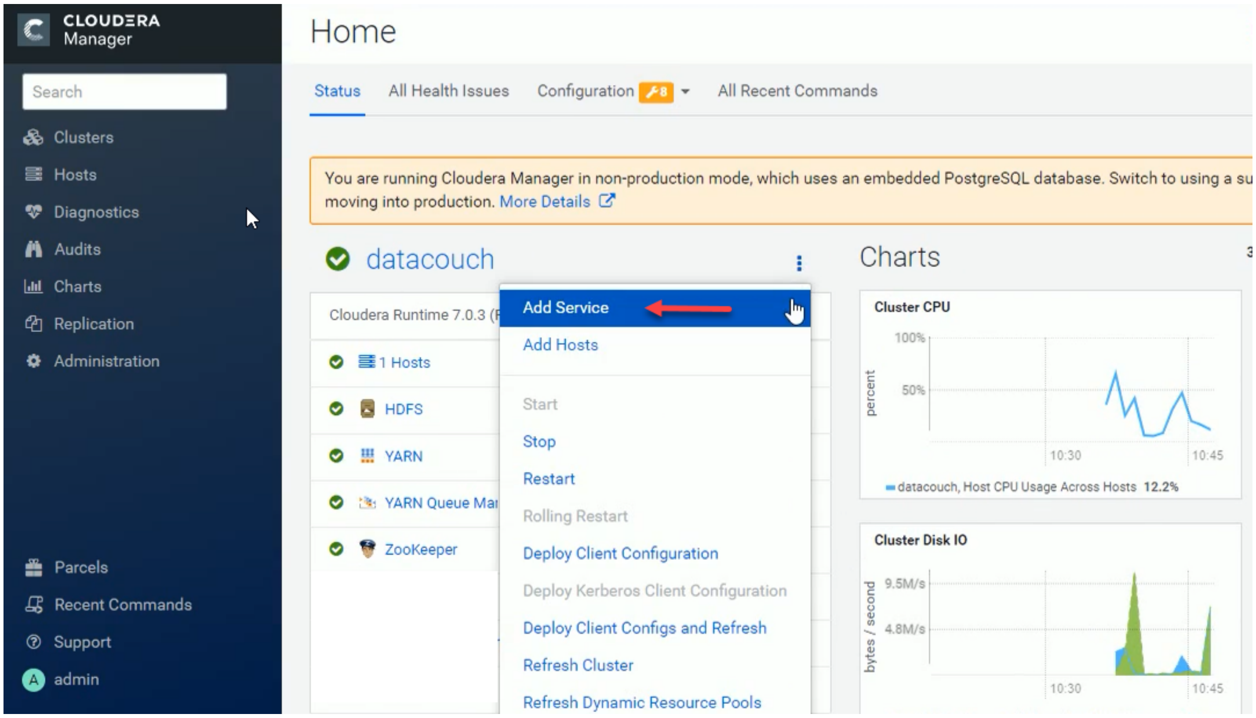
   In Cloudera Manager choose Clusters > HDFS. Then click on File Browser. Browse into tmp and confirm that sherlock.txt appears.

```
$ hdfs dfs -ls data/
```

Hands-On-Exercise: Installing Hive

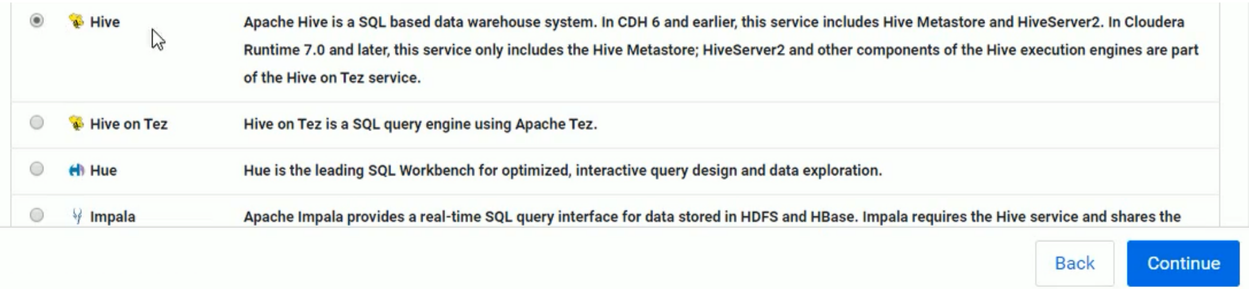
1. From the Cloudera Manager Home page, select the ‘Add a Service’ menu option from the drop-down menu to the right of Cluster Name.

   The Add Service Wizard appears.



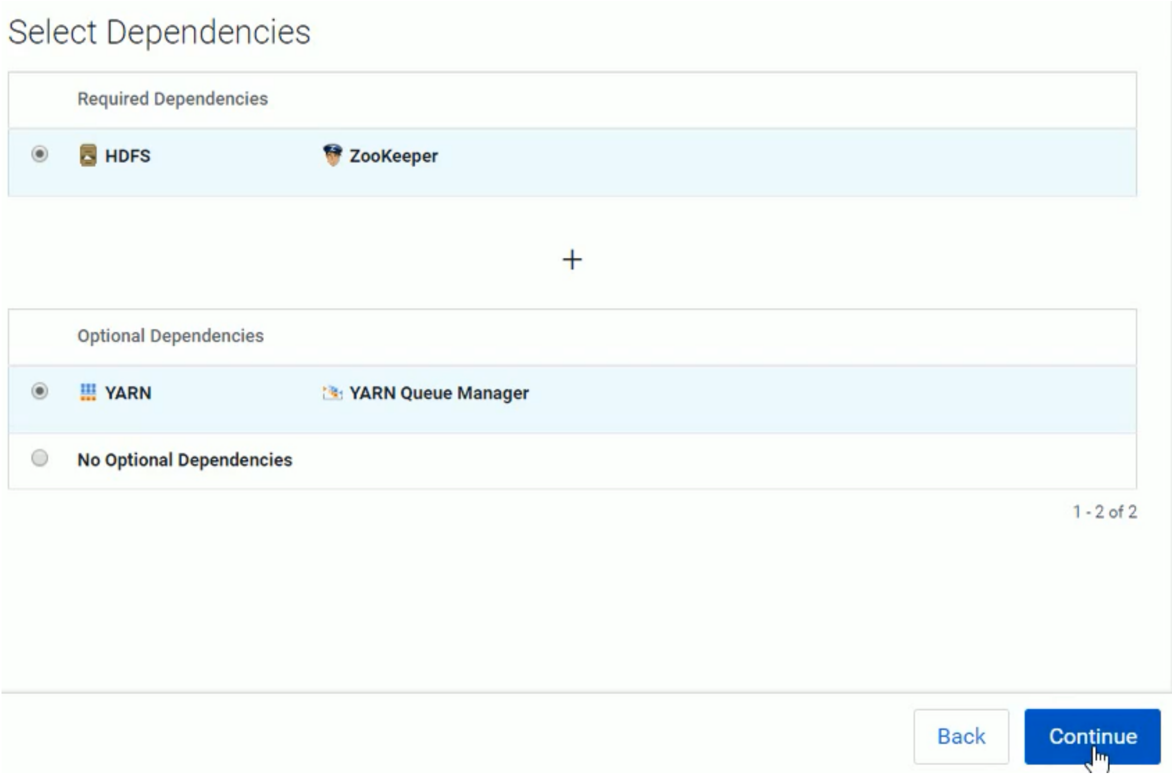
You are running Cloudera Manager in non-production mode, which uses an embedded PostgreSQL database. Switch to using a su moving into production. More Details

2. Select the Hive Service



●	Hive	Apache Hive is a SQL based data warehouse system. In CDH 6 and earlier, this service includes Hive Metastore and HiveServer2. In Cloudera Runtime 7.0 and later, this service only includes the Hive Metastore; HiveServer2 and other components of the Hive execution engines are part of the Hive on Tez service.
○	Hive on Tez	Hive on Tez is a SQL query engine using Apache Tez.
○	Hue	Hue is the leading SQL Workbench for optimized, interactive query design and data exploration.
○	Impala	Apache Impala provides a real-time SQL query interface for data stored in HDFS and HBase. Impala requires the Hive service and shares the

3. Select dependencies and click on continue.



Select Dependencies

	Required Dependencies	
●	HDFS	ZooKeeper

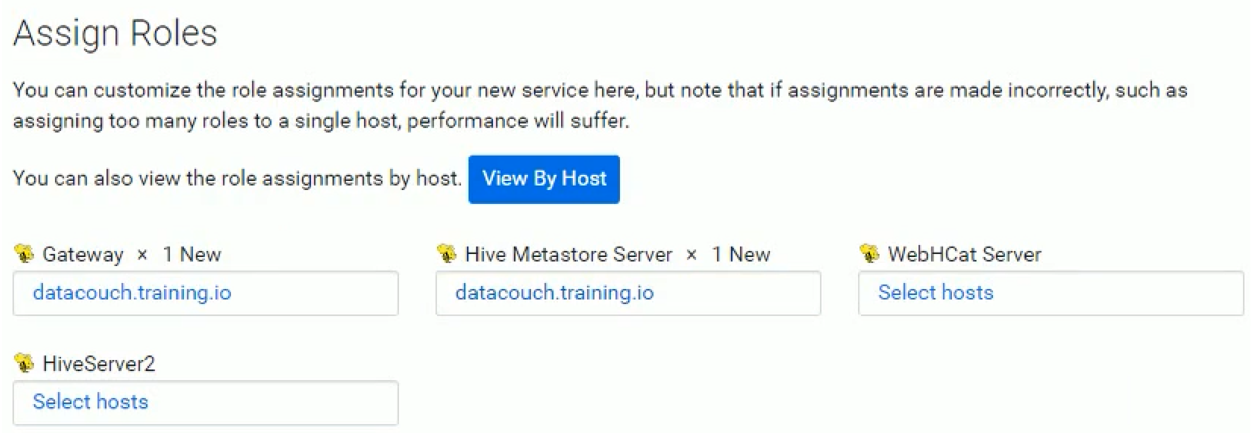
+

	Optional Dependencies	
●	YARN	YARN Queue Manager
○	No Optional Dependencies	

1 - 2 of 2

4. Choose center machine for your Gateway, Hive MetaStore and HiveServer2.

   Gateway: datacouch.training.io
   Hive Metastore Server: datacouch.training.io
   Hive Server2: datacouch.training.io



Assign Roles

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host.

Gateway × 1 New: datacouch.training.io

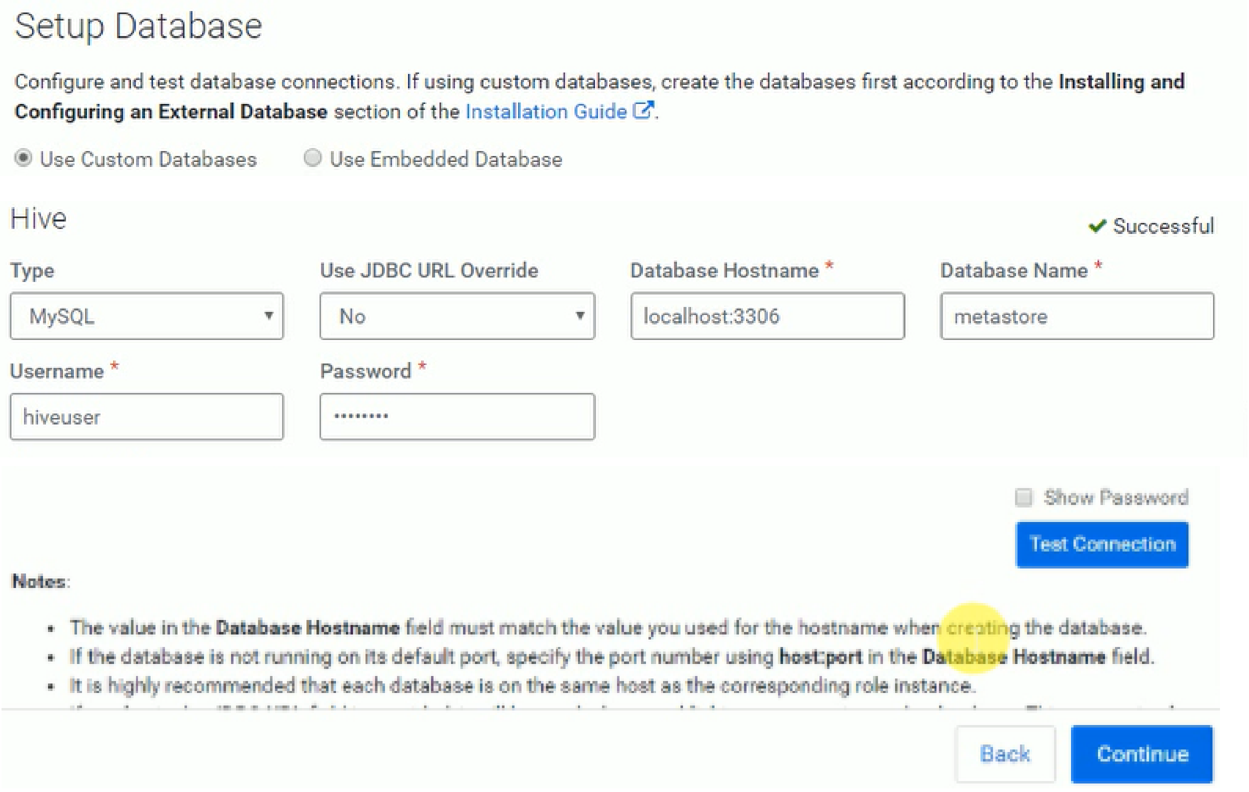
Hive Metastore Server × 1 New: datacouch.training.io

WebHCat Server: Select hosts

HiveServer2: Select hosts

5. Configure and test database connection

Select the database type and enter the database name, username, and password for each service.



Setup Database

Configure and test database connections. If using custom databases, create the databases first according to the **Installing and Configuring an External Database** section of the Installation Guide.

● Use Custom Databases ○ Use Embedded Database

Hive

✓ Successful

Type	Use JDBC URL Override	Database Hostname *	Database Name *
MySQL	No	localhost:3306	metastore

Username *	Password *
hiveuser	••••••••

□ Show Password

Notes:

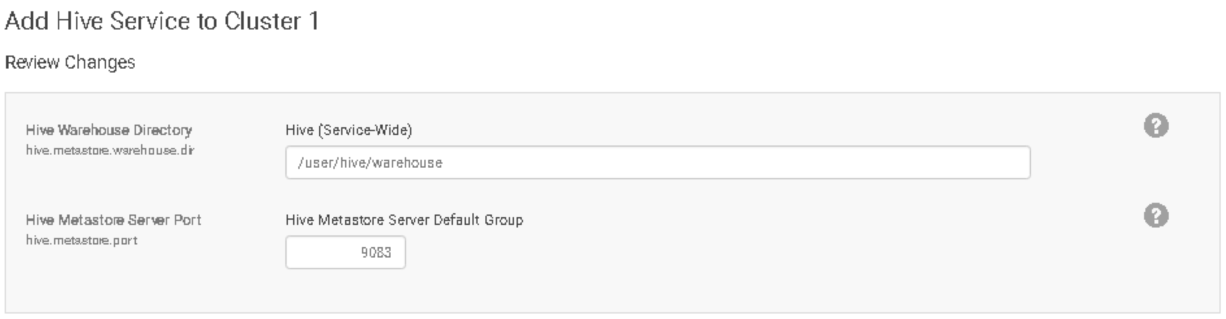
- The value in the **Database Hostname** field must match the value you used for the hostname when creating the database.
- If the database is not running on its default port, specify the port number using **host:port** in the **Database Hostname** field.
- It is highly recommended that each database is on the same host as the corresponding role instance.

Click Test Connection to validate the settings. If the connection is successful, a green checkmark and the word Successful appears next to each service. If there are any problems, the error is reported next to the service that failed to connect.

After verifying that each connection is successful, click Continue.

6. After testing the connection click Continue

   Mention the Hive Warehouse Directory and the Hive Metastore Port Number and Click Continue



Add Hive Service to Cluster 1

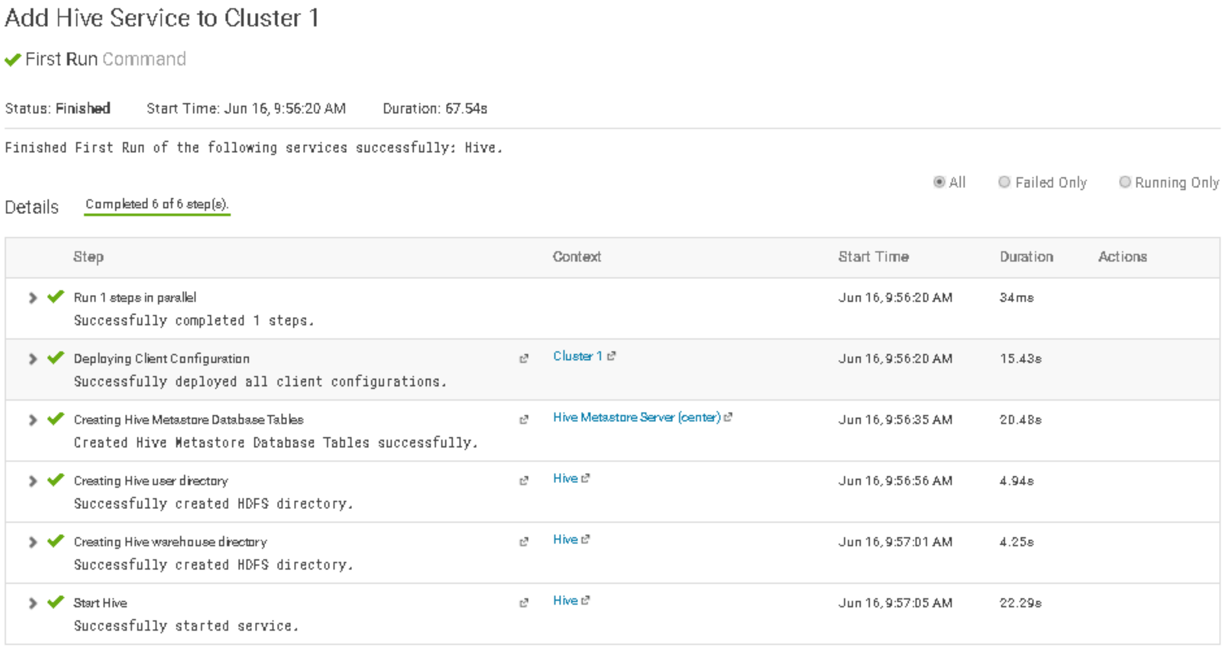
Review Changes

Hive Warehouse Directory
hive.metastore.warehouse.dir

Hive (Service-Wide)
/user/hive/warehouse

Hive Metastore Server Port
hive.metastore.port

Hive Metastore Server Default Group
9083



Add Hive Service to Cluster 1

✓ First Run Command

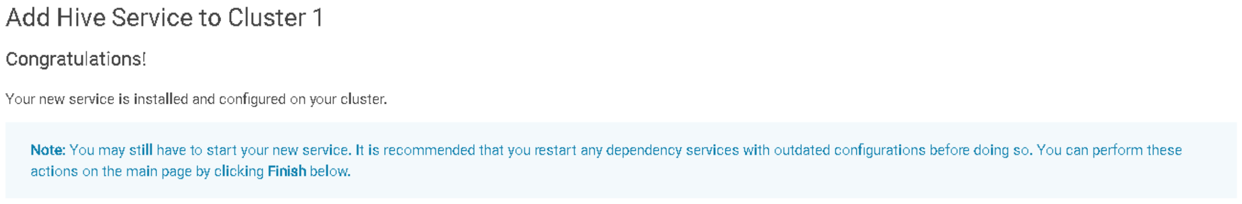
Status: Finished   Start Time: Jun 16, 9:56:20 AM   Duration: 67.54s

```
Finished First Run of the following services successfully: Hive.
```

● All ○ Failed Only ○ Running Only

Details   Completed 6 of 6 step(s).

Step	Context	Start Time	Duration	Actions
✓ Run 1 steps in parallel Successfully completed 1 steps.		Jun 16, 9:56:20 AM	34ms	
✓ Deploying Client Configuration Successfully deployed all client configurations.	Cluster 1	Jun 16, 9:56:20 AM	15.43s	
✓ Creating Hive Metastore Database Tables Created Hive Metastore Database Tables successfully.	Hive Metastore Server (center)	Jun 16, 9:56:35 AM	20.48s	
✓ Creating Hive user directory Successfully created HDFS directory.	Hive	Jun 16, 9:56:56 AM	4.94s	
✓ Creating Hive warehouse directory Successfully created HDFS directory.	Hive	Jun 16, 9:57:01 AM	4.25s	
✓ Start Hive Successfully started service.	Hive	Jun 16, 9:57:05 AM	22.29s	



Add Hive Service to Cluster 1

Congratulations!

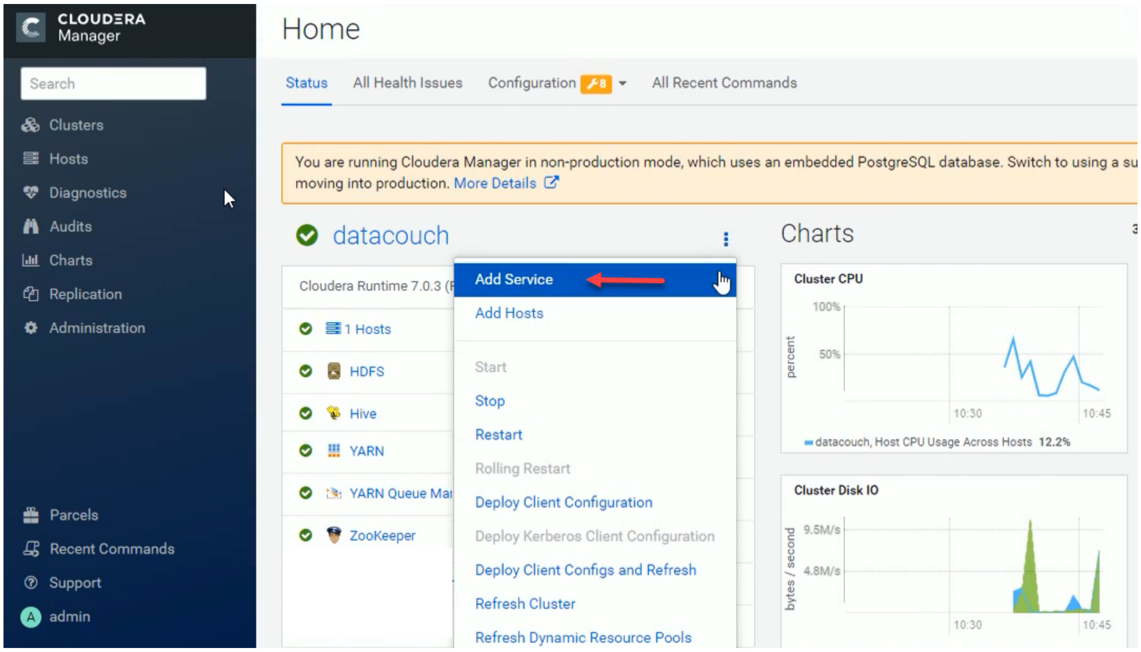
Your new service is installed and configured on your cluster.

Note: You may still have to start your new service. It is recommended that you restart any dependency services with outdated configurations before doing so. You can perform these actions on the main page by clicking **Finish** below.

Hands-On-Exercise: Installing Tez

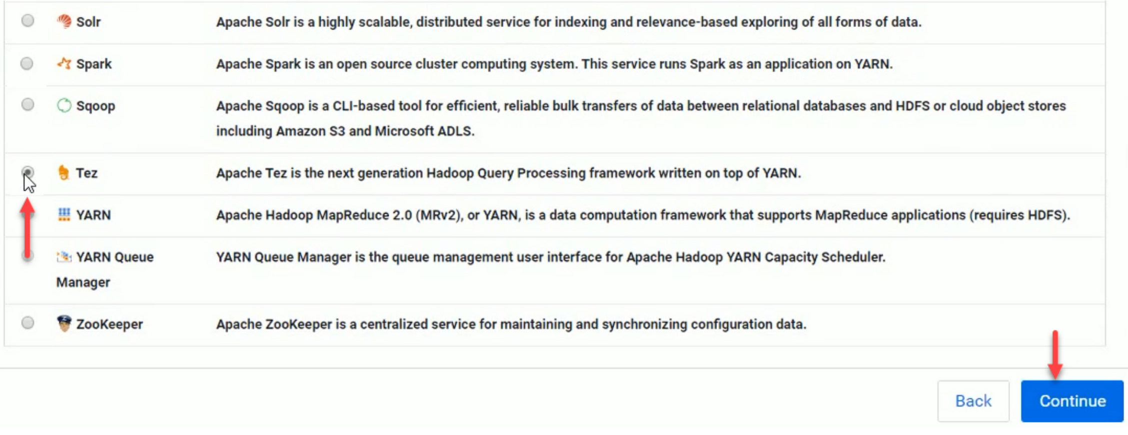
1. From the Cloudera Manager Home page, select the 'Add a Service' menu option from the drop-down menu to the right of Cluster Name.

   The Add Service Wizard appears.



You are running Cloudera Manager in non-production mode, which uses an embedded PostgreSQL database. Switch to using a su moving into production. More Details

2. Select the Tez Service

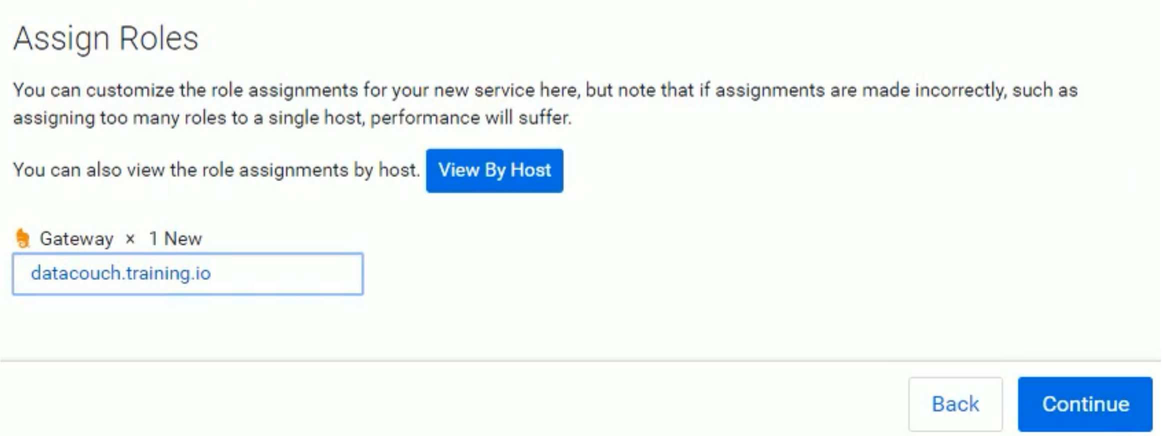


○	Solr	Apache Solr is a highly scalable, distributed service for indexing and relevance-based exploring of all forms of data.
○	Spark	Apache Spark is an open source cluster computing system. This service runs Spark as an application on YARN.
○	Sqoop	Apache Sqoop is a CLI-based tool for efficient, reliable bulk transfers of data between relational databases and HDFS or cloud object stores including Amazon S3 and Microsoft ADLS.
●	Tez	Apache Tez is the next generation Hadoop Query Processing framework written on top of YARN.
	YARN	Apache Hadoop MapReduce 2.0 (MRv2), or YARN, is a data computation framework that supports MapReduce applications (requires HDFS).
	YARN Queue Manager	YARN Queue Manager is the queue management user interface for Apache Hadoop YARN Capacity Scheduler.
○	ZooKeeper	Apache ZooKeeper is a centralized service for maintaining and synchronizing configuration data.

3. Customize Roles Setup for the Cluster

   Tez:

   Gateway: datacouch.training.io



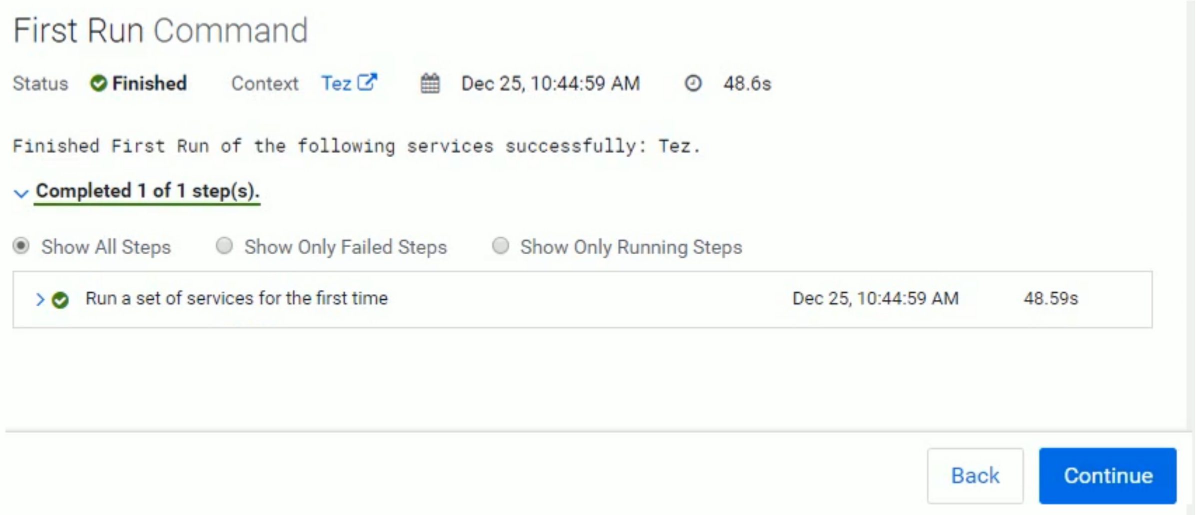
Assign Roles

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host.

Gateway × 1 New

datacouch.training.io



First Run Command

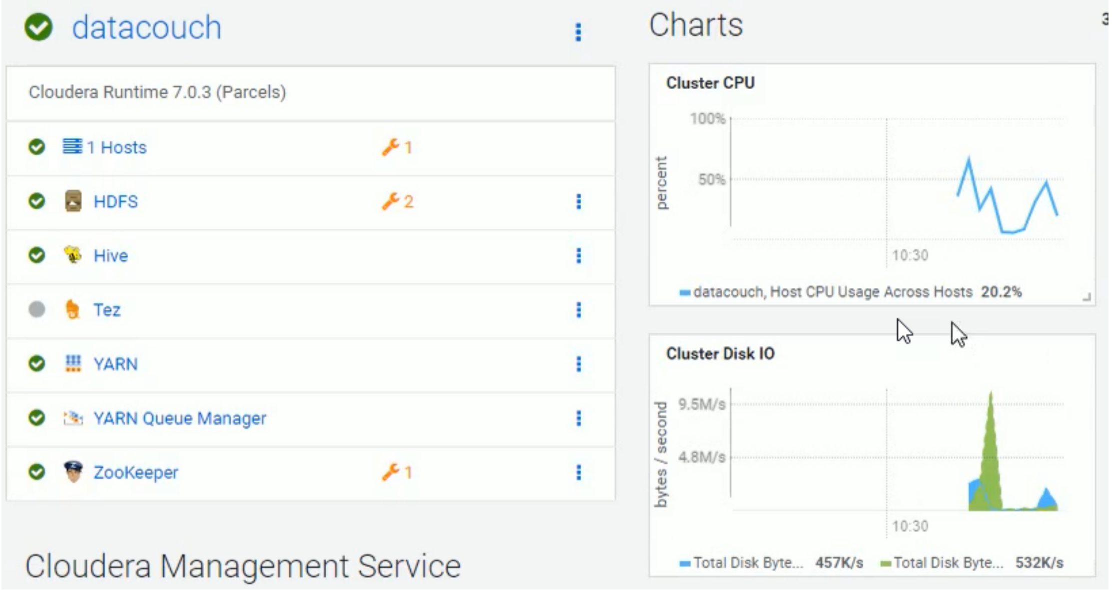
Status ✓ **Finished** Context Tez Dec 25, 10:44:59 AM 48.6s

```
Finished First Run of the following services successfully: Tez.
```

Completed 1 of 1 step(s).

● Show All Steps ○ Show Only Failed Steps ○ Show Only Running Steps

> ✓ Run a set of services for the first time	Dec 25, 10:44:59 AM	48.59s



✓ datacouch

Cloudera Runtime 7.0.3 (Parcels)

- ✓ 1 Hosts 1
- ✓ HDFS 2
- ✓ Hive
- ● Tez
- ✓ YARN
- ✓ YARN Queue Manager
- ✓ ZooKeeper 1

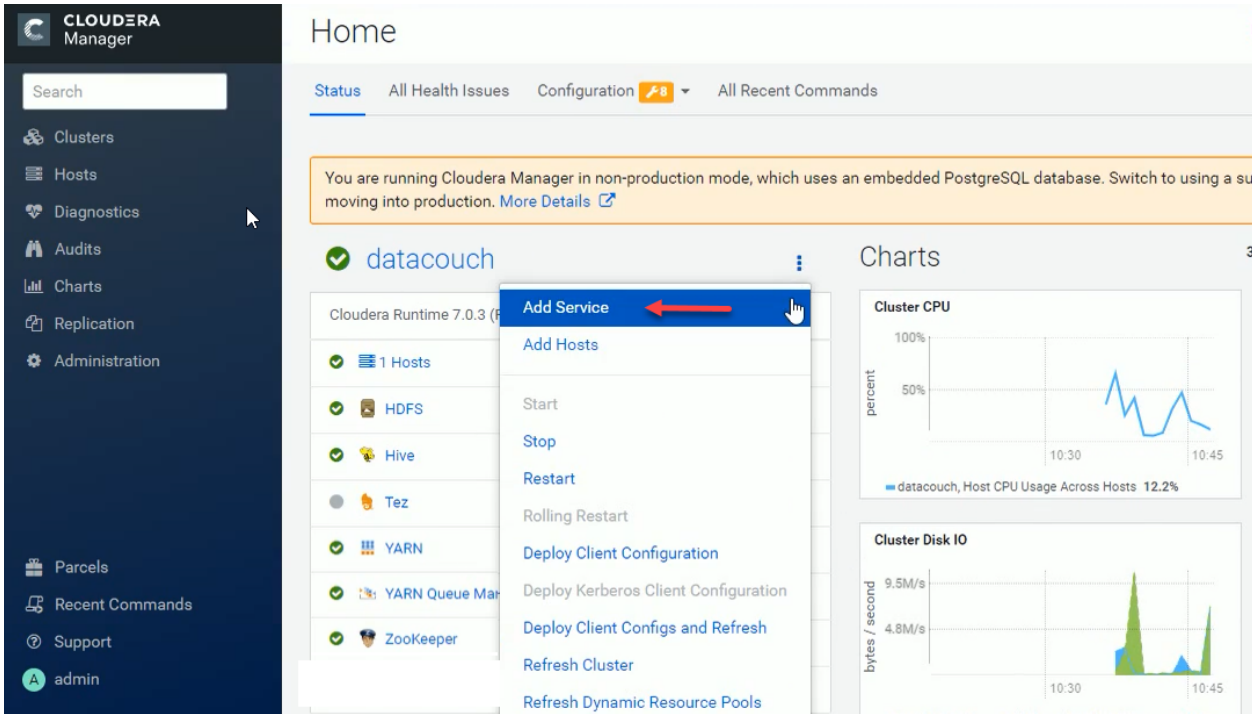
Cloudera Management Service

Charts
Cluster CPU
100%
50%
percent
10:30
datacouch, Host CPU Usage Across Hosts 20.2%
Cluster Disk IO
9.5M/s
4.8M/s
bytes / second
10:30
Total Disk Byte... 457K/s
Total Disk Byte... 532K/s

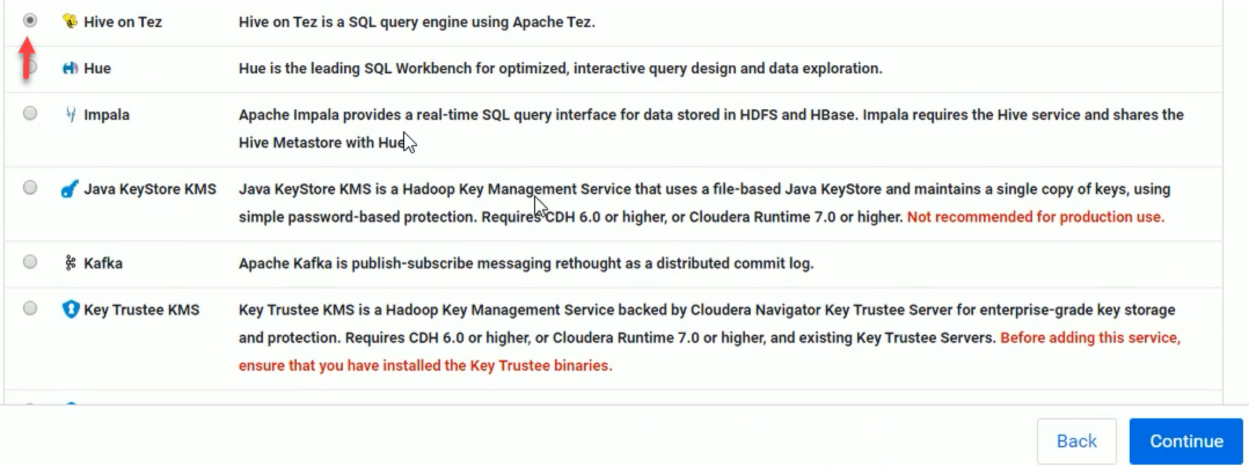
Hands-On-Exercise: Installing Hive on Tez

1. From the Cloudera Manager Home page, select the ‘Add a Service’ menu option from the drop-down menu to the right of Cluster Name.

   The Add Service Wizard appears.



2. Select the Hive on Tez Service

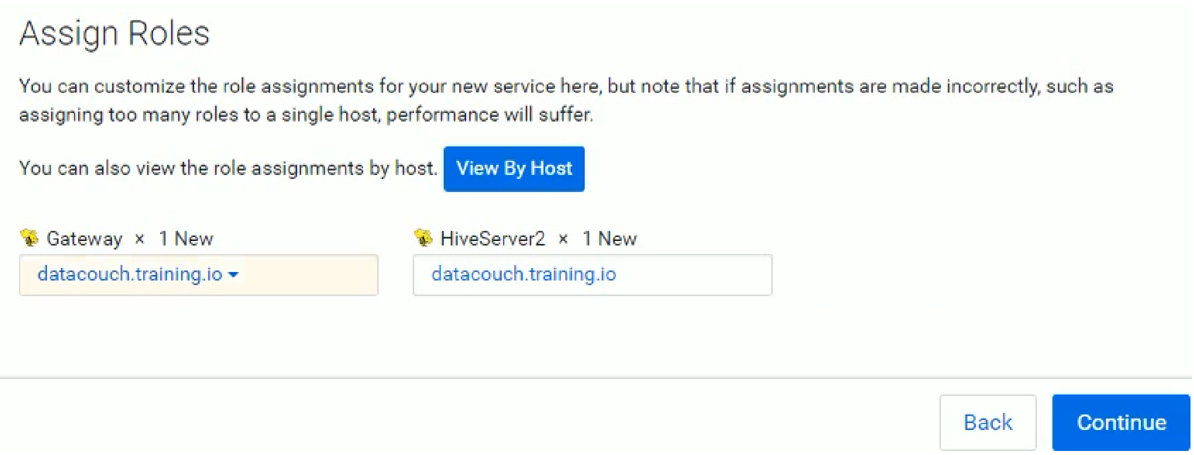


●	Hive on Tez	Hive on Tez is a SQL query engine using Apache Tez.
○	Hue	Hue is the leading SQL Workbench for optimized, interactive query design and data exploration.
○	Impala	Apache Impala provides a real-time SQL query interface for data stored in HDFS and HBase. Impala requires the Hive service and shares the Hive Metastore with Hue.
○	Java KeyStore KMS	Java KeyStore KMS is a Hadoop Key Management Service that uses a file-based Java KeyStore and maintains a single copy of keys, using simple password-based protection. Requires CDH 6.0 or higher, or Cloudera Runtime 7.0 or higher. Not recommended for production use.
○	Kafka	Apache Kafka is publish-subscribe messaging rethought as a distributed commit log.
○	Key Trustee KMS	Key Trustee KMS is a Hadoop Key Management Service backed by Cloudera Navigator Key Trustee Server for enterprise-grade key storage and protection. Requires CDH 6.0 or higher, or Cloudera Runtime 7.0 or higher, and existing Key Trustee Servers. Before adding this service, ensure that you have installed the Key Trustee binaries.

3. Customize Roles Setup for the Cluster

   Hive on Tez:

HiveServer2: datacouch.training.io



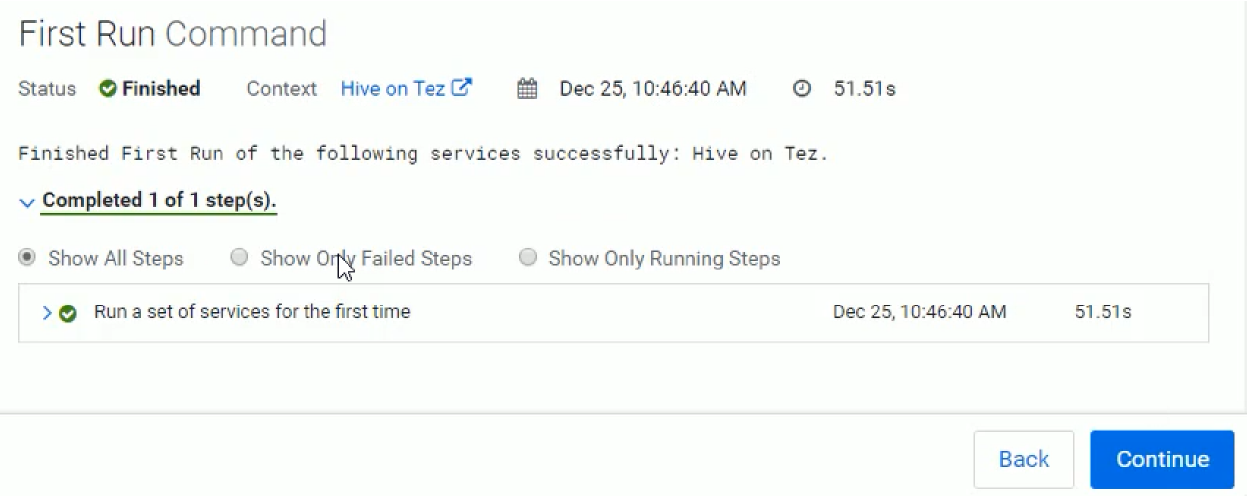
Assign Roles

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host. View By Host

Gateway × 1 New
datacouch.training.io

HiveServer2 × 1 New
datacouch.training.io



First Run Command

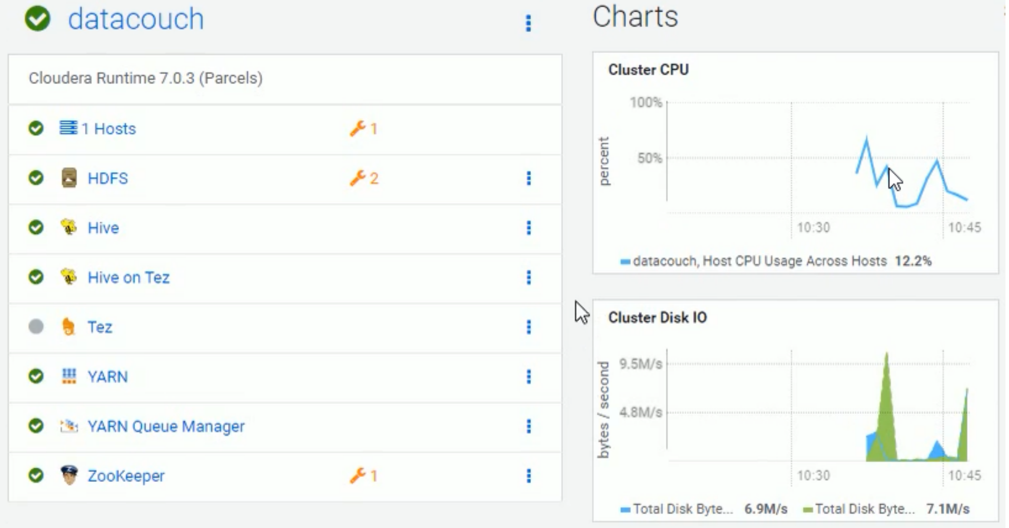
Status ✓ Finished Context Hive on Tez Dec 25, 10:46:40 AM 51.51s

```
Finished First Run of the following services successfully: Hive on Tez.
```

Completed 1 of 1 step(s).

● Show All Steps ○ Show Only Failed Steps ○ Show Only Running Steps

✓ Run a set of services for the first time	Dec 25, 10:46:40 AM	51.51s



✓ datacouch

Cloudera Runtime 7.0.3 (Parcels)

✓ 1 Hosts	1
✓ HDFS	2
✓ Hive	
✓ Hive on Tez	
○ Tez	
✓ YARN	
✓ YARN Queue Manager	
✓ ZooKeeper	1

Charts

Cluster CPU
100%
50%
percent
10:30
10:45
datacouch, Host CPU Usage Across Hosts 12.2%
Cluster Disk IO
9.5M/s
4.8M/s
bytes / second
10:30
10:45
Total Disk Byte... 6.9M/s
Total Disk Byte... 7.1M/s

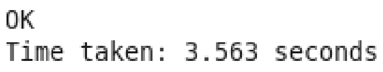
Hive Validation

Step 1 : Invoke Hive shell

```
$ hive
```

Step 2 : Create a Database

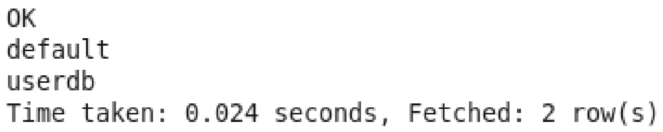
```
hive> CREATE DATABASE userdb;
```



```
OK
Time taken: 3.563 seconds
```

Step 3 : Verify an existing Databases

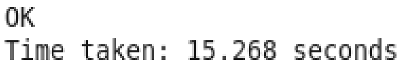
```
hive> SHOW DATABASES;
```



```
OK
default
userdb
Time taken: 0.024 seconds, Fetched: 2 row(s)
```

Step 4 : Drop Database

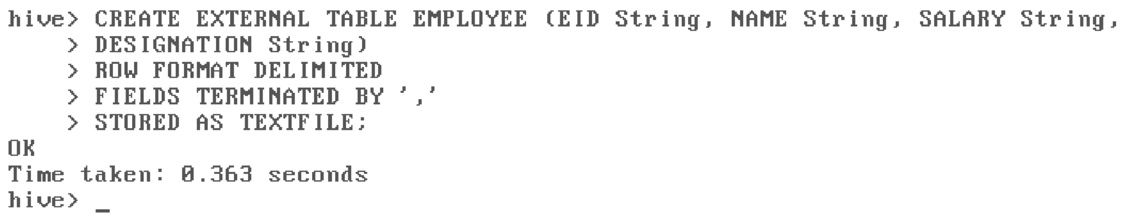
```
hive> DROP DATABASE userdb;
```



```
OK
Time taken: 15.268 seconds
```

Step 5 : Create Table

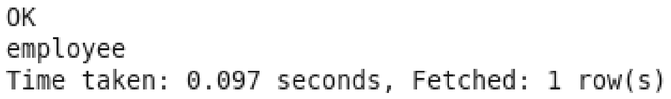
```
hive> CREATE EXTERNAL TABLE EMPLOYEE (EID String, NAME String, SALARY
String,Designation String)
 ROW FORMAT DELIMITED
 FIELDS TERMINATED BY ','
 STORED AS TEXTFILE;
```



```
hive> CREATE EXTERNAL TABLE EMPLOYEE (EID String, NAME String, SALARY String,
    > DESIGNATION String)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.363 seconds
hive> _
```

Verify the tables has been created

```
hive> SHOW TABLES;
```

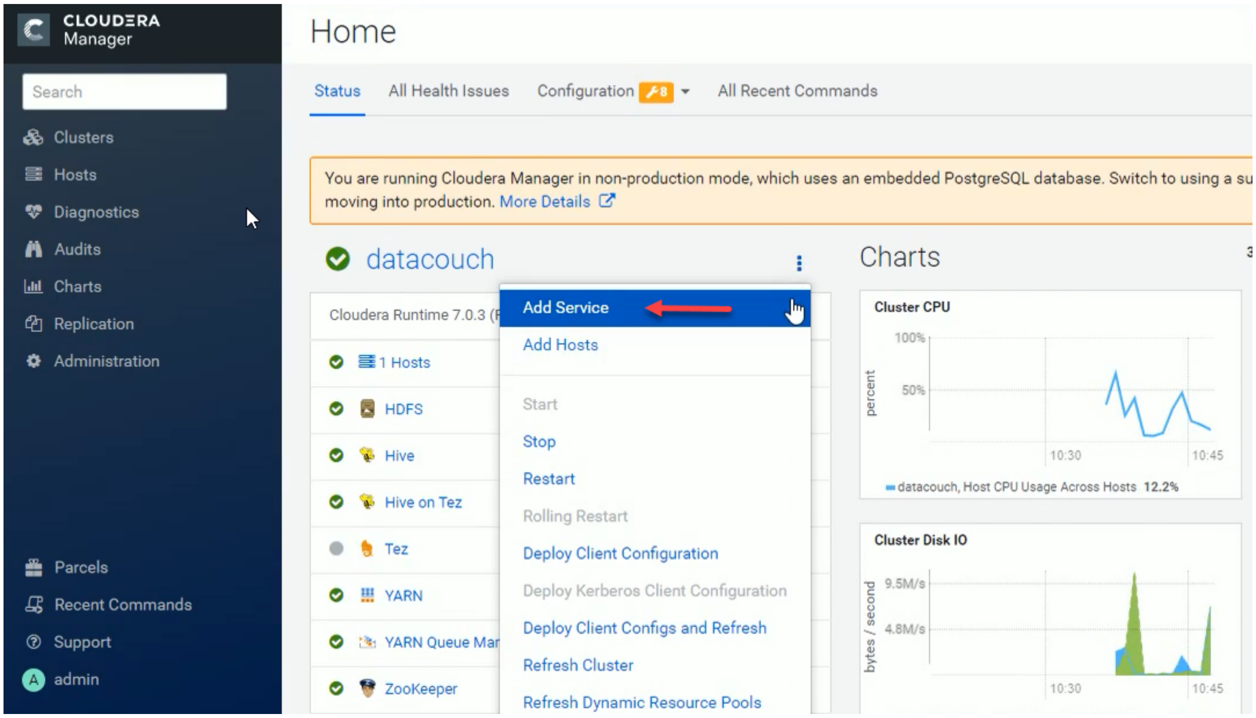


```
OK
employee
Time taken: 0.097 seconds, Fetched: 1 row(s)
```

Hands-On-Exercise: Deploying Spark 2.4

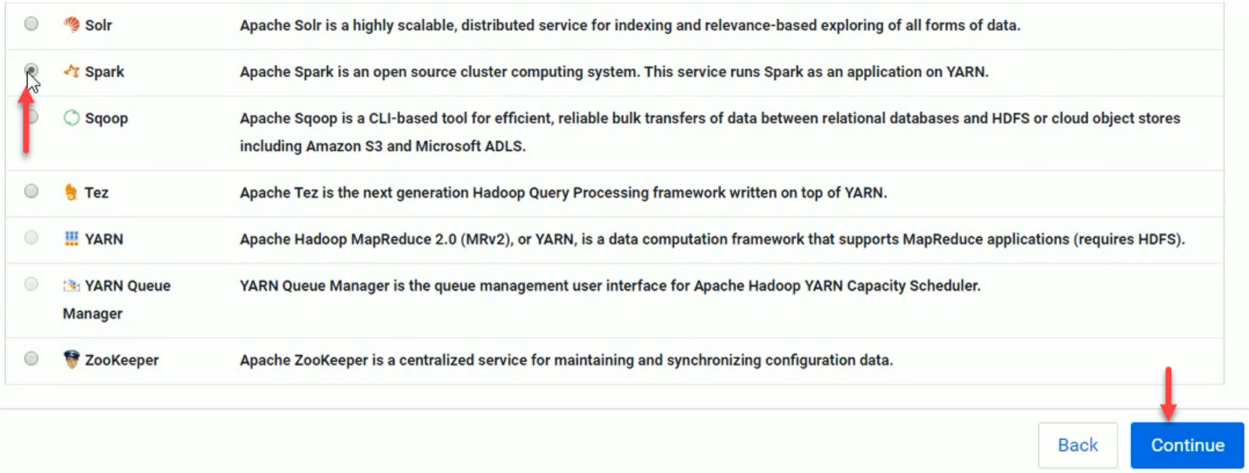
1. From the Cloudera Manager Home page, select the ‘Add a Service’ menu option from the drop-down menu to the right of Cluster Name.

   The Add Service Wizard appears.



You are running Cloudera Manager in non-production mode, which uses an embedded PostgreSQL database. Switch to using a su moving into production. More Details

2. Select Spark and Continue

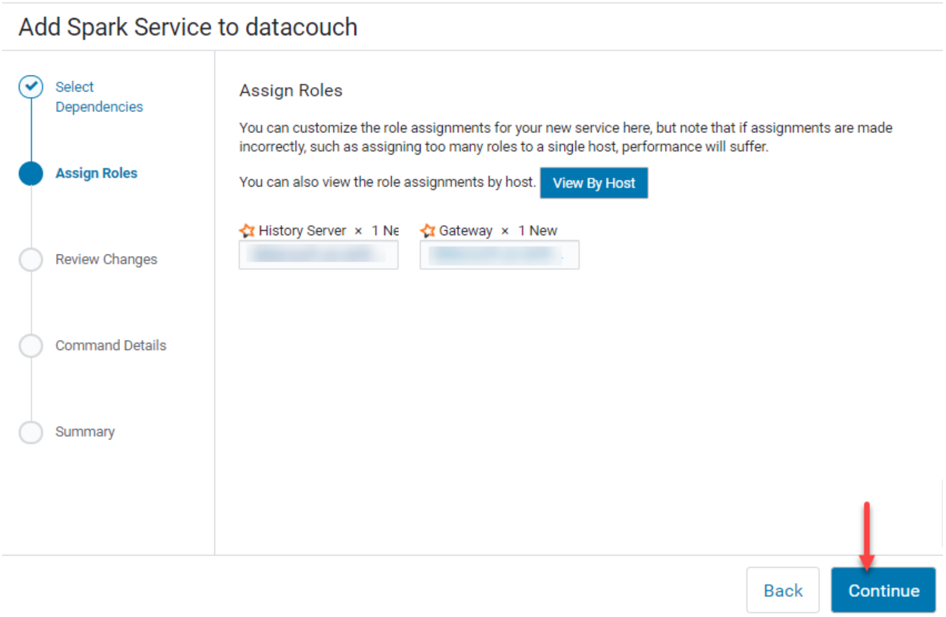


○	Solr	Apache Solr is a highly scalable, distributed service for indexing and relevance-based exploring of all forms of data.
●	Spark	Apache Spark is an open source cluster computing system. This service runs Spark as an application on YARN.
○	Sqoop	Apache Sqoop is a CLI-based tool for efficient, reliable bulk transfers of data between relational databases and HDFS or cloud object stores including Amazon S3 and Microsoft ADLS.
○	Tez	Apache Tez is the next generation Hadoop Query Processing framework written on top of YARN.
○	YARN	Apache Hadoop MapReduce 2.0 (MRv2), or YARN, is a data computation framework that supports MapReduce applications (requires HDFS).
○	YARN Queue Manager	YARN Queue Manager is the queue management user interface for Apache Hadoop YARN Capacity Scheduler.
○	ZooKeeper	Apache ZooKeeper is a centralized service for maintaining and synchronizing configuration data.

3. Customize Role Assignments

   Spark History server: datacouch.training.io

Spark Gateway: datacouch.training.io



Add Spark Service to datacouch

Assign Roles

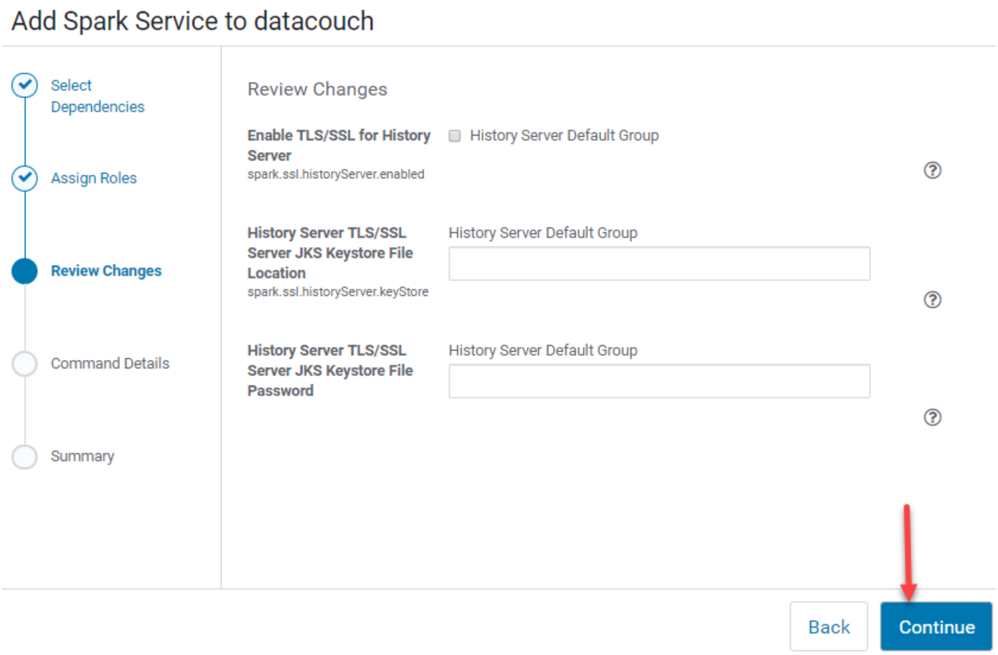
You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host.

History Server × 1 Ne [redacted]

Gateway × 1 New [redacted]

4. Use default settings click on continue



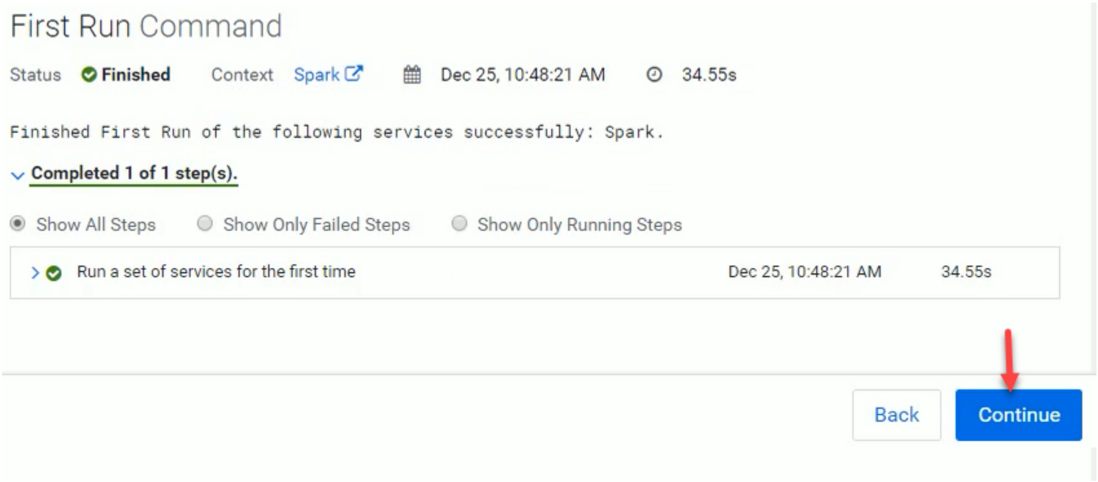
Add Spark Service to datacouch

Review Changes

Enable TLS/SSL for History Server spark.ssl.historyServer.enabled — □ History Server Default Group

History Server TLS/SSL Server JKS Keystore File Location spark.ssl.historyServer.keyStore — History Server Default Group

History Server TLS/SSL Server JKS Keystore File Password — History Server Default Group



First Run Command

Status ✓ Finished   Context Spark   Dec 25, 10:48:21 AM   34.55s

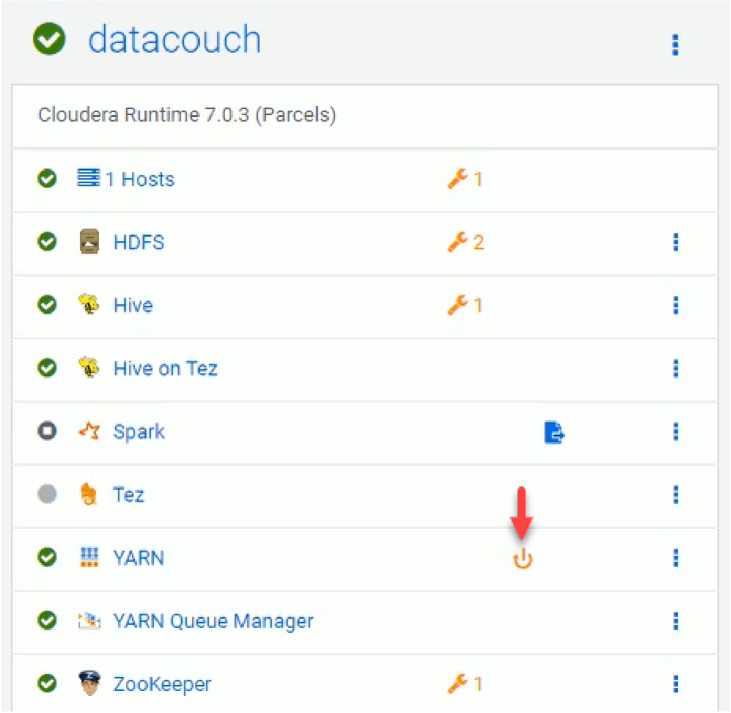
```
Finished First Run of the following services successfully: Spark.
```

Completed 1 of 1 step(s).

● Show All Steps ○ Show Only Failed Steps ○ Show Only Running Steps

> ✓ Run a set of services for the first time	Dec 25, 10:48:21 AM	34.55s

5. Click Restart Stale Services.



datacouch

Cloudera Runtime 7.0.3 (Parcels)

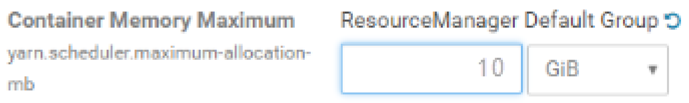
✓	1 Hosts	1
✓	HDFS	2
✓	Hive	1
✓	Hive on Tez	
○	Spark	
●	Tez	
✓	YARN	
✓	YARN Queue Manager	
✓	ZooKeeper	1

Before Running pyspark2

Set:

1. Click on search box

   Set **“yarn.scheduler.maximum-allocation-mb” 10 GB** then click on save

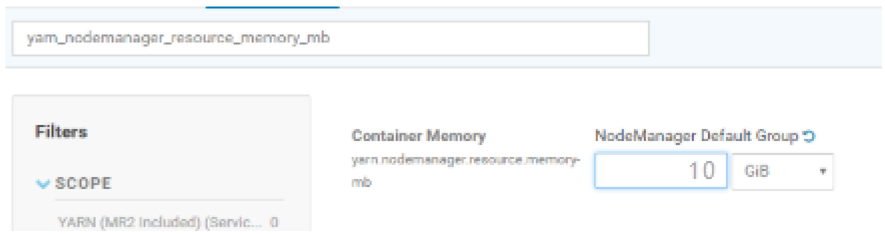


Container Memory Maximum
yarn.scheduler.maximum-allocation-mb

ResourceManager Default Group: 10 GiB

   Restart stale service

   Set **“yarn.nodemanager.resource.memory-mb” 10 GB** then click on save



yarn_nodemanager_resource_memory_mb

Filters

SCOPE

YARN (MR2 Included) (Servic... 0

Container Memory
yarn.nodemanager.resource.memory-mb

NodeManager Default Group: 10 GiB

   Restart stale service

Running Job on Apache Spark2

1. Upload **sherlock.txt** in ***~/hadoop-admin/data*** to HDFS

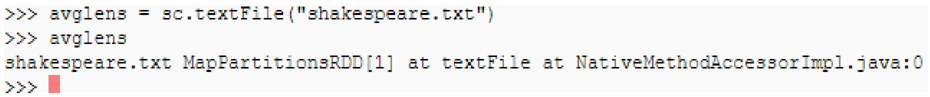
```
$ hdfs dfs -put sherlock.txt /user/training/
```

2. Open the spark shell

```
$ pyspark --master yarn
```

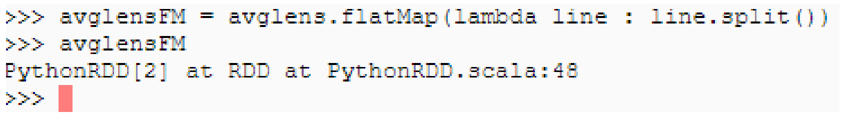
3. Making RDD from the textFile

```
>>> avglens = sc.textFile("sherlock.txt")
>>> avglens
```



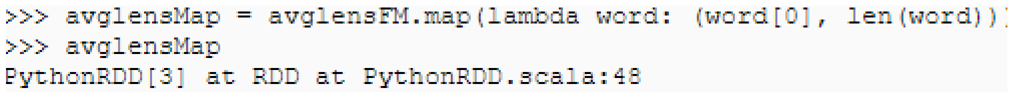
```
>>> avglens = sc.textFile("shakespeare.txt")
>>> avglens
shakespeare.txt MapPartitionsRDD[1] at textFile at NativeMethodAccessorImpl.java:0
>>> 
```

```
>>> avglensFM = avglens.flatMap(lambda line : line.split())
>>> avglensFM
```



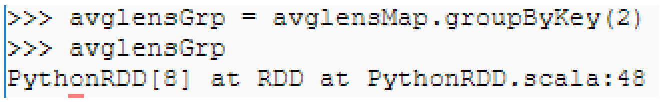
```
>>> avglensFM = avglens.flatMap(lambda line : line.split())
>>> avglensFM
PythonRDD[2] at RDD at PythonRDD.scala:48
>>> 
```

```
>>> avglensMap = avglensFM.map(lambda word: (word[0], len(word)))
>>> avglensMap
```



```
>>> avglensMap = avglensFM.map(lambda word: (word[0], len(word))
>>> avglensMap
PythonRDD[3] at RDD at PythonRDD.scala:48
```

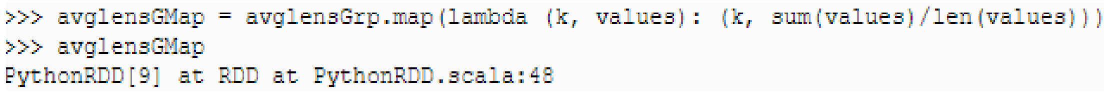
```
>>> avglensGrp = avglensMap.groupByKey(2)
>>> avglensGrp
```



```
>>> avglensGrp = avglensMap.groupByKey(2)
>>> avglensGrp
PythonRDD[8] at RDD at PythonRDD.scala:48
```

```
>>>avglensGMap = avglensGrp.map(lambda (k, values): (k,
sum(values)/len(values)))
```

```
>>>avglensGMap
```

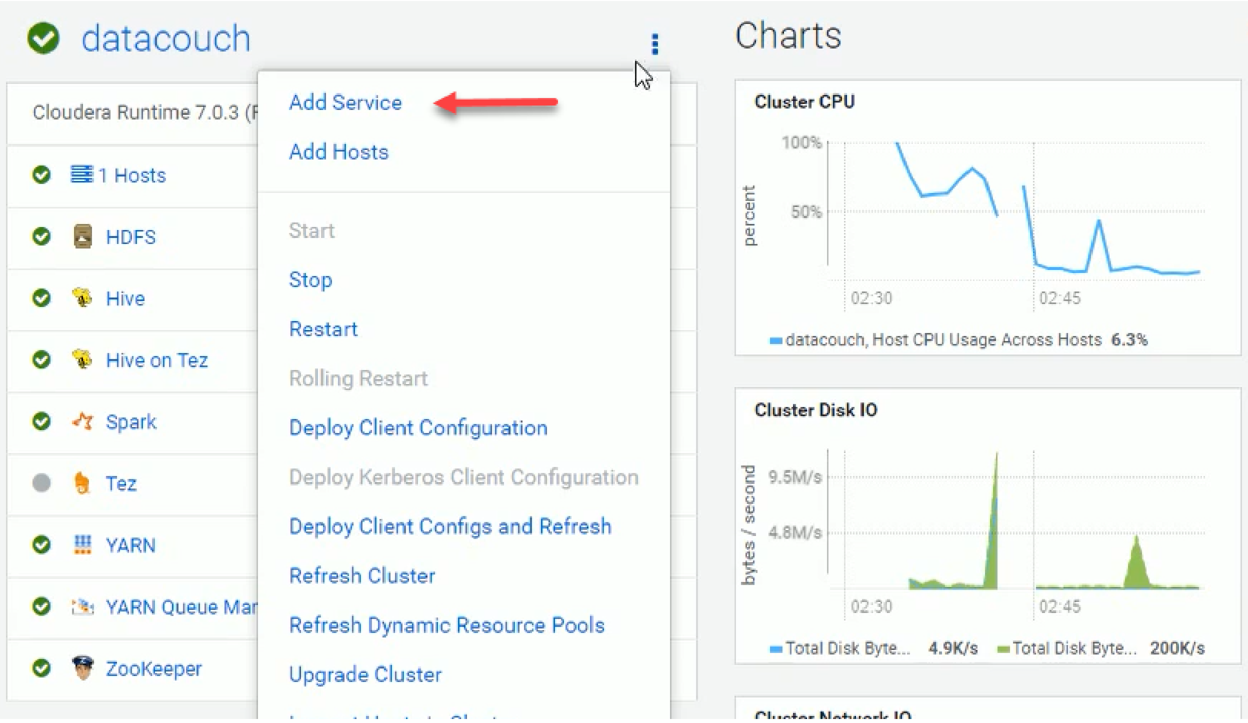


```
>>> avglensGMap = avglensGrp.map(lambda (k, values): (k, sum(values)/len(values)))
>>> avglensGMap
PythonRDD[9] at RDD at PythonRDD.scala:48
```

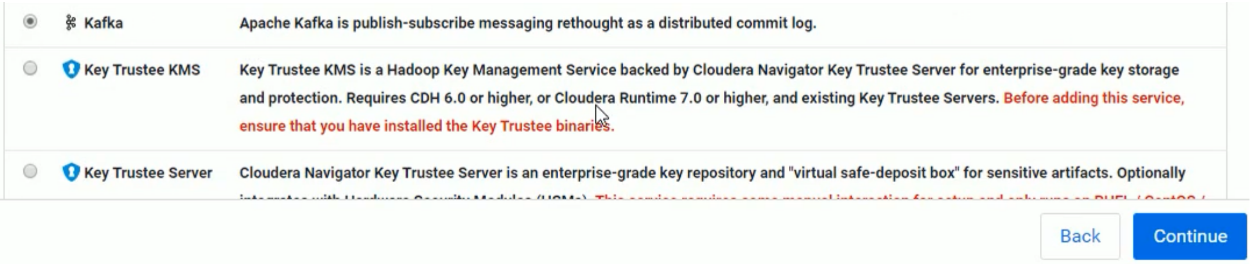
Hands-On-Exercise: Installing Kafka

1. From the Cloudera Manager Home page, select the ‘Add a Service’ menu option from the drop-down menu to the right of Cluster Name.

   The Add Service Wizard appears.

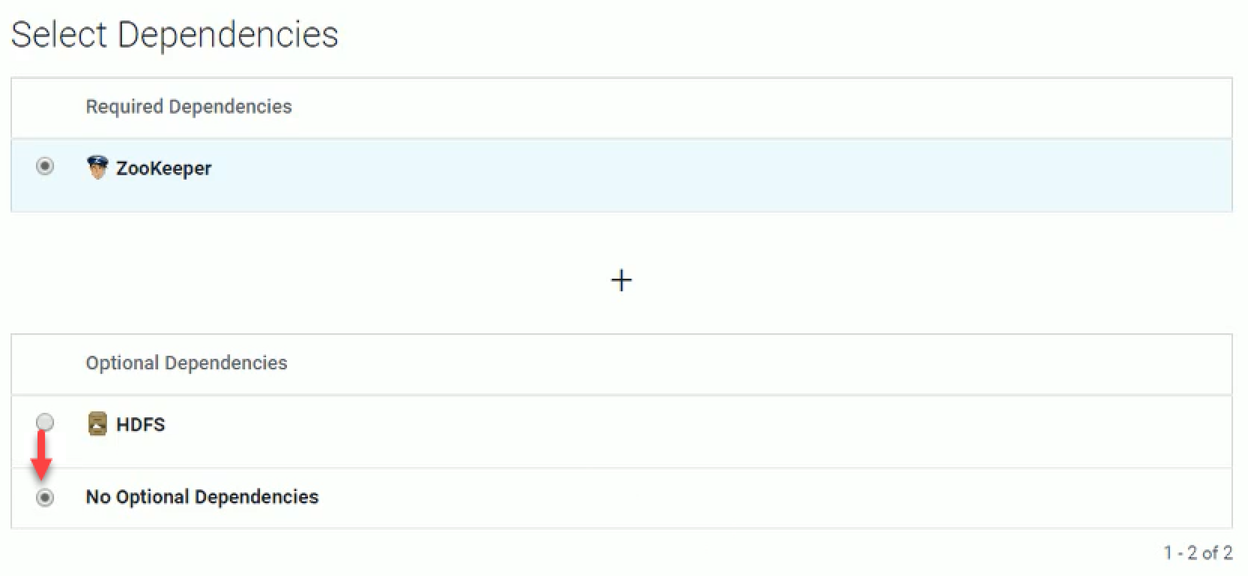

datacouch
Charts
Cloudera Runtime 7.0.3 (F
1 Hosts
HDFS
Hive
Hive on Tez
Spark
Tez
YARN
YARN Queue Mar
ZooKeeper
Add Service
Add Hosts
Start
Stop
Restart
Rolling Restart
Deploy Client Configuration
Deploy Kerberos Client Configuration
Deploy Client Configs and Refresh
Refresh Cluster
Refresh Dynamic Resource Pools
Upgrade Cluster
Cluster CPU
100%
50%
percent
02:30
02:45
datacouch, Host CPU Usage Across Hosts 6.3%
Cluster Disk IO
9.5M/s
4.8M/s
bytes / second
02:30
02:45
Total Disk Byte... 4.9K/s
Total Disk Byte... 200K/s

2. Select Kafka and Continue



●	Kafka	Apache Kafka is publish-subscribe messaging rethought as a distributed commit log.
○	Key Trustee KMS	Key Trustee KMS is a Hadoop Key Management Service backed by Cloudera Navigator Key Trustee Server for enterprise-grade key storage and protection. Requires CDH 6.0 or higher, or Cloudera Runtime 7.0 or higher, and existing Key Trustee Servers. Before adding this service, ensure that you have installed the Key Trustee binaries.
○	Key Trustee Server	Cloudera Navigator Key Trustee Server is an enterprise-grade key repository and "virtual safe-deposit box" for sensitive artifacts. Optionally

3. Select No Optional Dependencies



Select Dependencies

Required Dependencies

● ZooKeeper

+

Optional Dependencies

○ HDFS

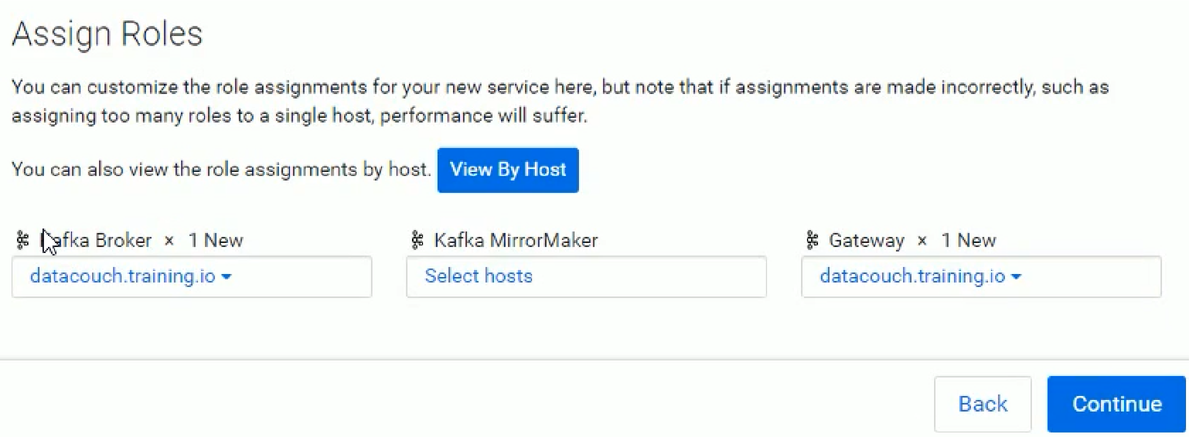
● No Optional Dependencies

1 - 2 of 2

4. Customize Role Assignments

Kafka

Kafka Broker: datacouch.training.io
Gateway: datacouch.training.io



Assign Roles

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

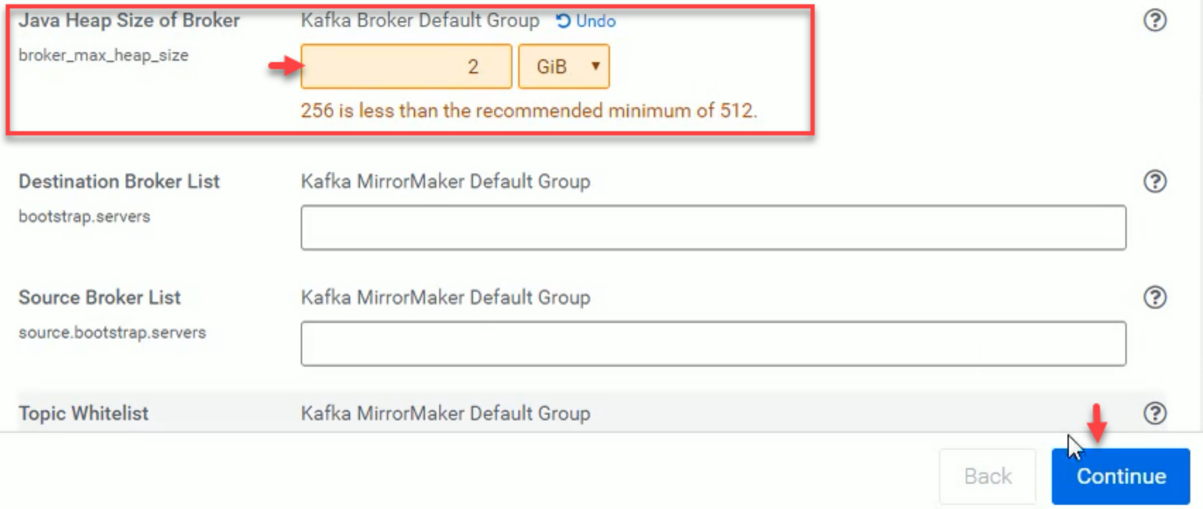
You can also view the role assignments by host. View By Host

Kafka Broker × 1 New: datacouch.training.io

Kafka MirrorMaker: Select hosts

Gateway × 1 New: datacouch.training.io

5. Set Java Heap Size of Broker 2 GB



Java Heap Size of Broker
broker_max_heap_size

Kafka Broker Default Group ⟲ Undo

2 GiB ▾

256 is less than the recommended minimum of 512.

Destination Broker List
bootstrap.servers

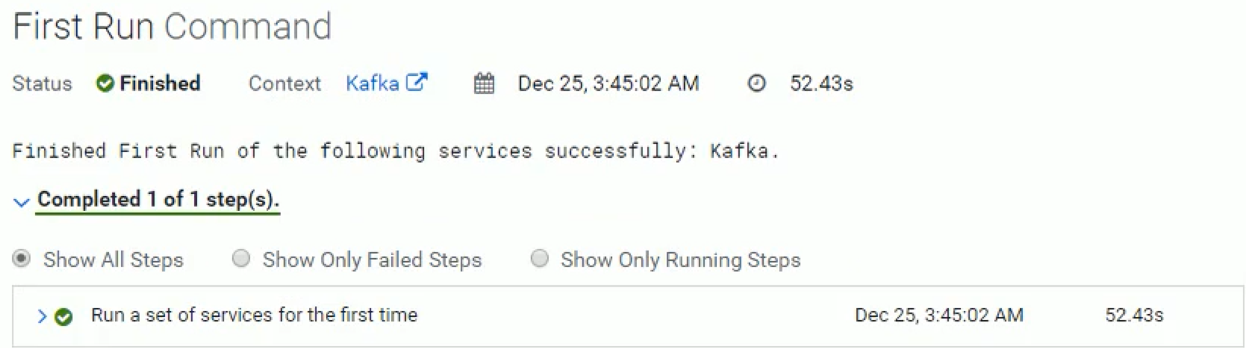
Kafka MirrorMaker Default Group

Source Broker List
source.bootstrap.servers

Kafka MirrorMaker Default Group

Topic Whitelist

Kafka MirrorMaker Default Group



First Run Command

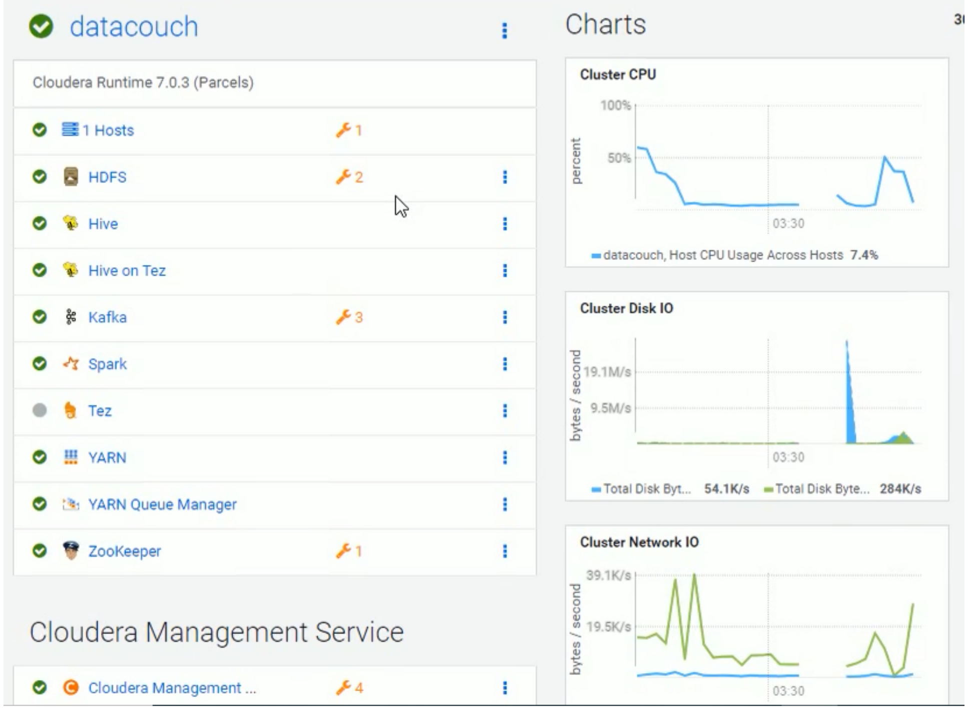
Status ✓ Finished   Context Kafka   Dec 25, 3:45:02 AM   52.43s

```
Finished First Run of the following services successfully: Kafka.
```

Completed 1 of 1 step(s).

● Show All Steps   ○ Show Only Failed Steps   ○ Show Only Running Steps

✓ Run a set of services for the first time	Dec 25, 3:45:02 AM	52.43s



datacouch

Cloudera Runtime 7.0.3 (Parcels)

✓ 1 Hosts	1
✓ HDFS	2
✓ Hive	
✓ Hive on Tez	
✓ Kafka	3
✓ Spark	
● Tez	
✓ YARN	
✓ YARN Queue Manager	
✓ ZooKeeper	1

Cloudera Management Service

✓ Cloudera Management ...	4

Kafka Validation

This section describes ways you can use Kafka tools for data capture for analysis.

```
kafka-topics --create --zookeeper datacouch.training.io:2181/kafka
--replication-factor 1 --partitions 1 --topic test
```

Let's create a topic named "test" with a single partition and only one replica:

```
kafka-topics --list --zookeeper datacouch.training.io:2181/kafka
```

kafka-console-producer
Read data from standard output and write it to a Kafka topic. For example:

```
kafka-console-producer --broker-list datacouch.training.io:9092 --topic
test
```

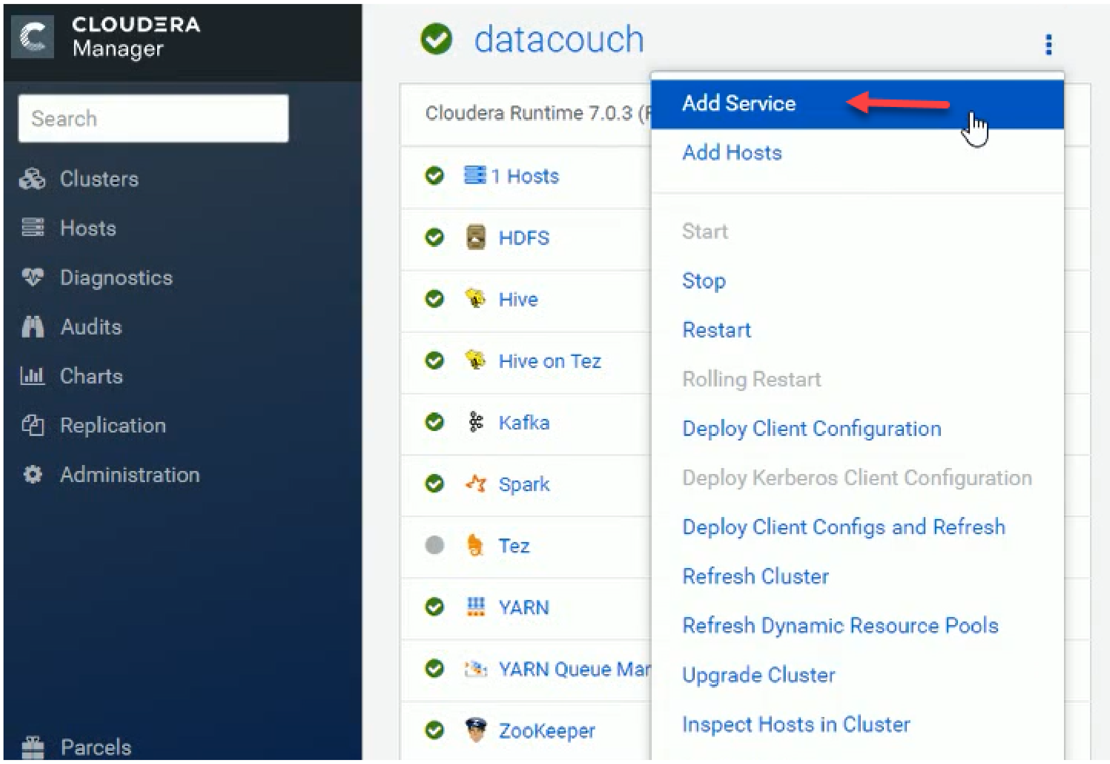
Kafka also has a command line consumer that will dump out messages to standard output.

```
kafka-console-consumer --bootstrap-server datacouch.training.io:9092
--topic test --from-beginning
```

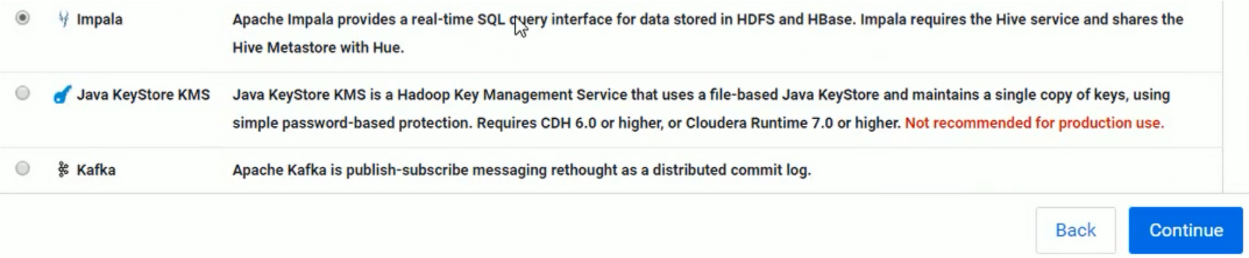
Hands-On-Exercise: Installing Impala

1. From the Cloudera Manager Home page, select the ‘Add a Service’ menu option from the drop-down menu to the right of Cluster Name.

   The Add Service Wizard appears.



2. Select Impala and Continue



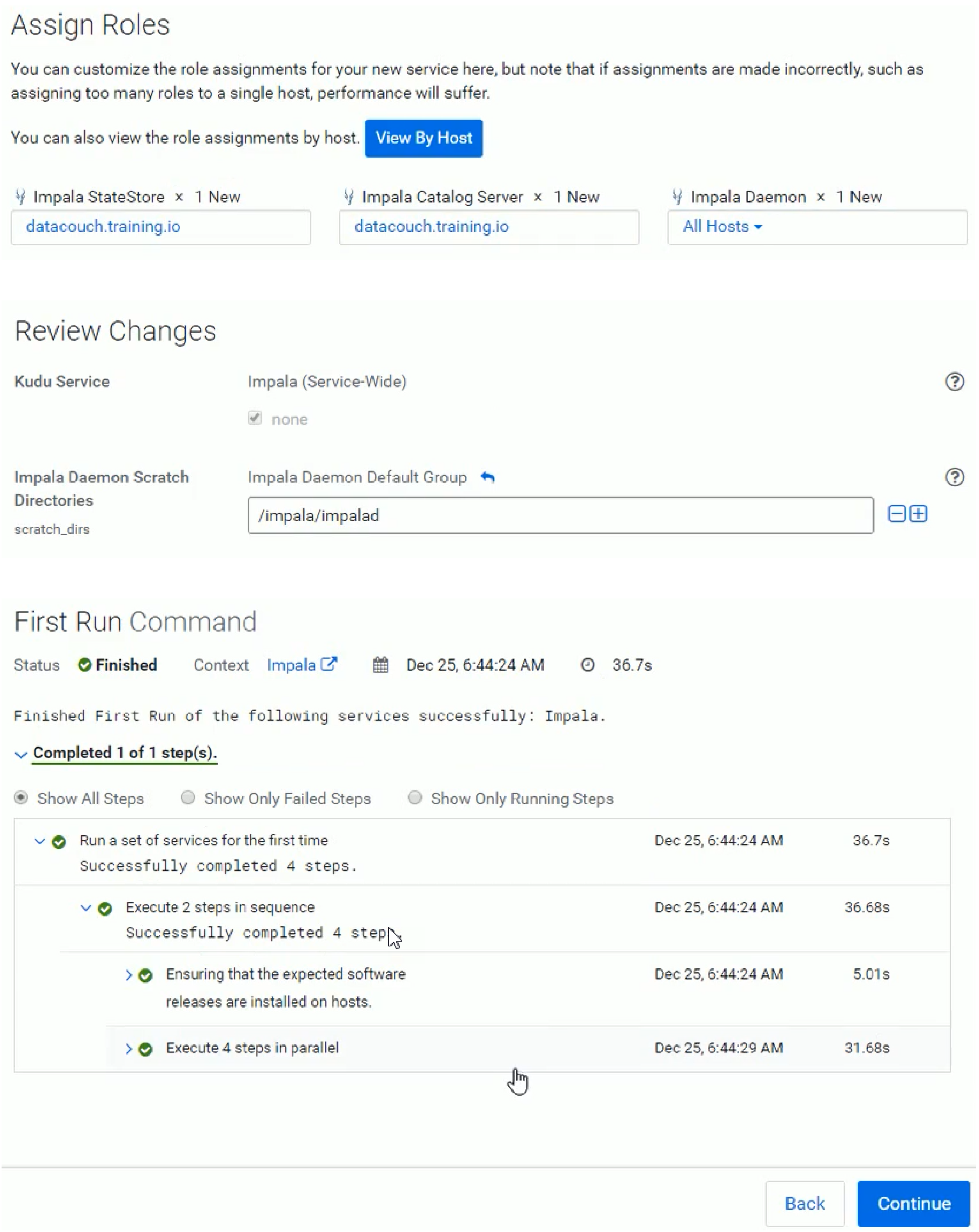
●	Impala	Apache Impala provides a real-time SQL query interface for data stored in HDFS and HBase. Impala requires the Hive service and shares the Hive Metastore with Hue.
○	Java KeyStore KMS	Java KeyStore KMS is a Hadoop Key Management Service that uses a file-based Java KeyStore and maintains a single copy of keys, using simple password-based protection. Requires CDH 6.0 or higher, or Cloudera Runtime 7.0 or higher. Not recommended for production use.
○	Kafka	Apache Kafka is publish-subscribe messaging rethought as a distributed commit log.

6. Customize Role Assignments

   Impala

   Impala SateStore: datacouch.training.io
   Impala Catalog Server:datacouch.training.io

Impala Daemon:datacouch.training.io



Assign Roles

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host. View By Host

Impala StateStore × 1 New	Impala Catalog Server × 1 New	Impala Daemon × 1 New
datacouch.training.io	datacouch.training.io	All Hosts ▾

Review Changes

Kudu Service — Impala (Service-Wide)

☑ none

Impala Daemon Scratch Directories
scratch_dirs — Impala Daemon Default Group

/impala/impalad

First Run Command

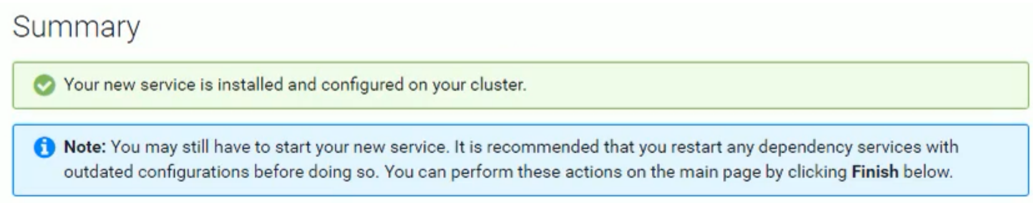
Status ✓ **Finished** Context Impala Dec 25, 6:44:24 AM 36.7s

```
Finished First Run of the following services successfully: Impala.
```

Completed 1 of 1 step(s).

● Show All Steps ○ Show Only Failed Steps ○ Show Only Running Steps

Step	Start	Duration
✓ Run a set of services for the first time `Successfully completed 4 steps.`	Dec 25, 6:44:24 AM	36.7s
✓ Execute 2 steps in sequence `Successfully completed 4 step`	Dec 25, 6:44:24 AM	36.68s
✓ Ensuring that the expected software releases are installed on hosts.	Dec 25, 6:44:24 AM	5.01s
✓ Execute 4 steps in parallel	Dec 25, 6:44:29 AM	31.68s

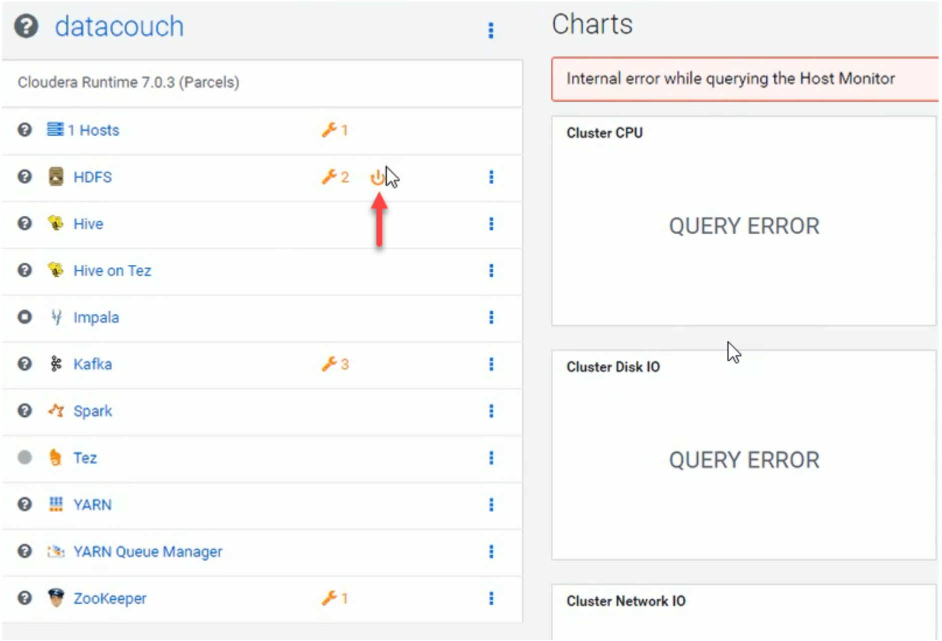


Summary

Your new service is installed and configured on your cluster.

Note: You may still have to start your new service. It is recommended that you restart any dependency services with outdated configurations before doing so. You can perform these actions on the main page by clicking **Finish** below.

7. Restart Stale Configuration



datacouch

Cloudera Runtime 7.0.3 (Parcels)

1 Hosts	1	
HDFS	2	⏻
Hive		
Hive on Tez		
Impala		
Kafka	3	
Spark		
Tez		
YARN		
YARN Queue Manager		
ZooKeeper	1	

Charts

Internal error while querying the Host Monitor

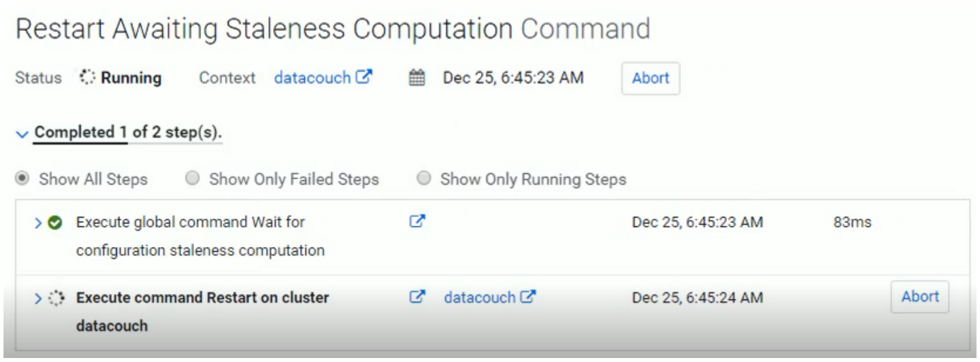
Cluster CPU

QUERY ERROR

Cluster Disk IO

QUERY ERROR

Cluster Network IO



Restart Awaiting Staleness Computation Command

Status **Running** Context datacouch Dec 25, 6:45:23 AM Abort

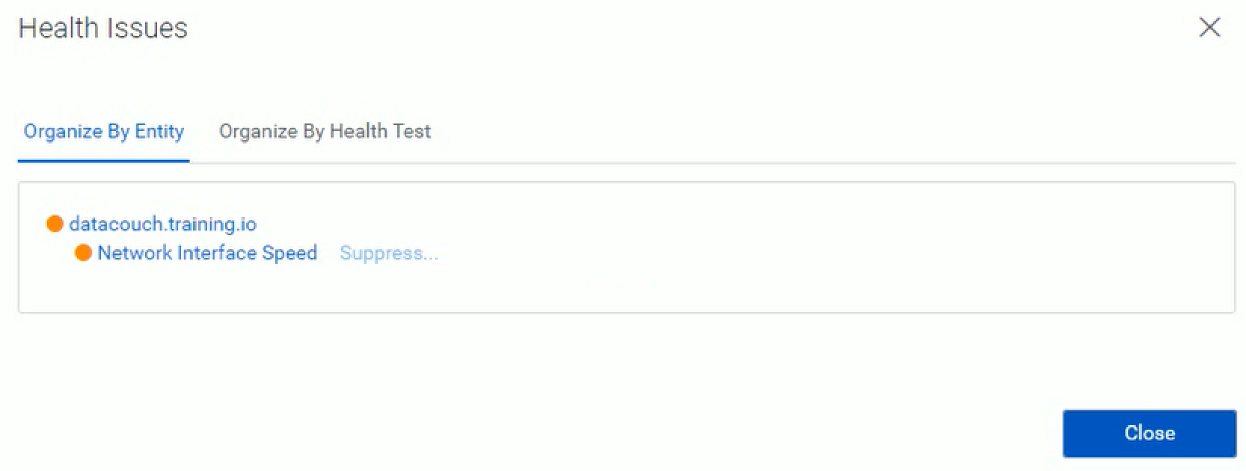
Completed 1 of 2 step(s).

● Show All Steps ○ Show Only Failed Steps ○ Show Only Running Steps

Execute global command Wait for configuration staleness computation		Dec 25, 6:45:23 AM	83ms
Execute command Restart on cluster datacouch	datacouch	Dec 25, 6:45:24 AM	Abort

Common Warnings and Errors

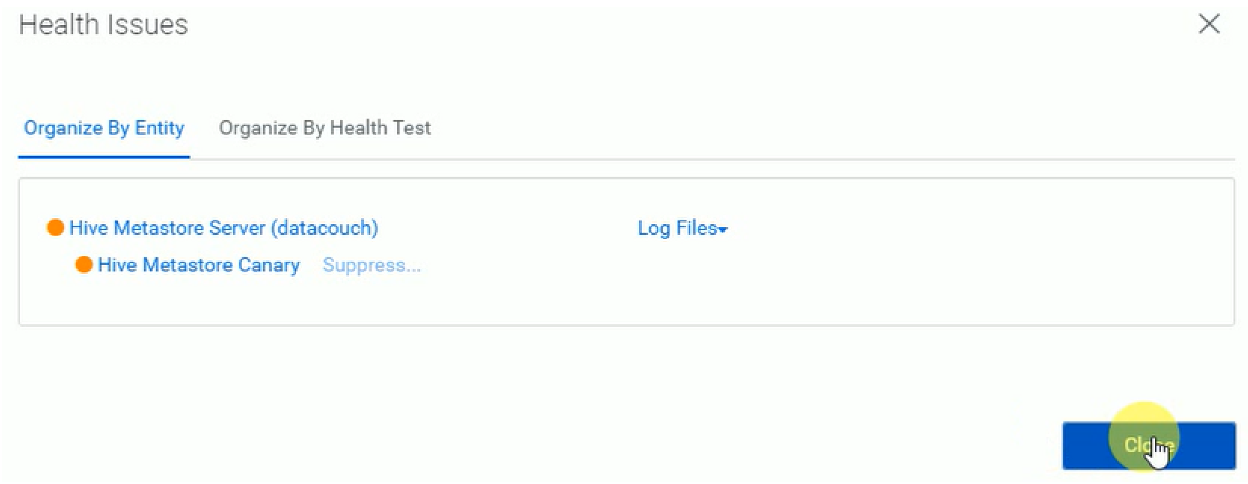
1. Click on Suppress... **Network Interface Speed**



Health Issues

● datacouch.training.io
● Network Interface Speed Suppress...

2. Click on Suppress... **Hive Metastore Canary**



Health Issues

● Hive Metastore Server (datacouch) Log Files▾
● Hive Metastore Canary Suppress...

3. Change the property of **server_host** inside **/etc/cloudera-scm-agent/config.ini**

```
sudo vi /etc/cloudera-scm-agent/config.ini
```

Enter the hostname of your machine.

Restart stale service

References

1. https://www.cloudera.com/documentation/enterprise/latest/topics/cm_ig_mysql.html#cmig_topic_5_5_3
2. https://docs.cloudera.com/cdpdc/7.0/installation/topics/cdpdc-installation.html