PowerScale OneFS with Hadoop and Cloudera Installation Guide

8.1.2 – 9.0.0

Abstract

This guide walks you through the process of installing PowerScale OneFS with the Cloudera distribution of Hadoop.

Published in the USA.

EMC Corporation
Hopkinton, Massachusetts 01748-9103
1-508-435-1000 In North America 1-866-464-7381
www.EMC.com

Publication History

Version	Date	Description
1.01	January 27, 2017	Initial version
1.02	July 31, 2017	• Updated list of users that the Isilon_hadoop_tools script creates. • Removed instructions to modify the block size that is used for reading from Isilon since the block size is 128M by default. • Removed the "`--fixperm`" flag when using the Isilon_create_users.sh script. That flag is intended for user on existing deployments where directory ownership is wrong on OneFS. • Updated various screenshots and the troubleshooting section.
1.03	October 5, 2017	• Added "Updates and Additional Information on Isilon Hadoop Installs" section.
2.00	August 19, 2019	• Added OneFS 8.1.2 related content
3.00	June 16, 2020	• Added OneFS 8.2-9.0 related content.

DELL EMC

CONTENTS

Introduction 5
Audience 5
Overview 5
Updates and additional information about OneFS Hadoop installs 7
Prerequisites 7
Cloudera Distribution with Apache Hadoop (CDH) 7
OneFS cluster configuration 7
Installing OneFS with Cloudera Manager 8
Preparing OneFS 8
Validate OneFS version and license activation 9
Configure OneFS components 9
Create an Access Zone 10
Configure SmartConnect 11
Configure DNS for OneFS 12
Verify the SmartConnect configuration 12
Create HDFS users and groups 13
Create users and directories on the OneFS cluster using Tools for Using Hadoop with OneFS 13
Create users on the OneFS cluster manually 13
Create and configure HDFS for OneFS 8.1.2 and previous versions 14
Configure HDFS user for OneFS 8.2 and later versions 15
Preparing Cloudera 16
Deploy Cloudera Manager 17
Configure the Hadoop cluster 22
Troubleshoot Cloudera Manager 30
Verify the deployment 32
Contacting Dell EMC PowerScale Technical Support 34

Introduction

Hadoop is an open-source framework that enables the distributed processing of large sets of data across clusters of systems. You can follow the steps in this guide to install PowerScale OneFS with Hadoop for use with Cloudera.

Before you begin, you must install a PowerScale OneFS cluster.

Audience

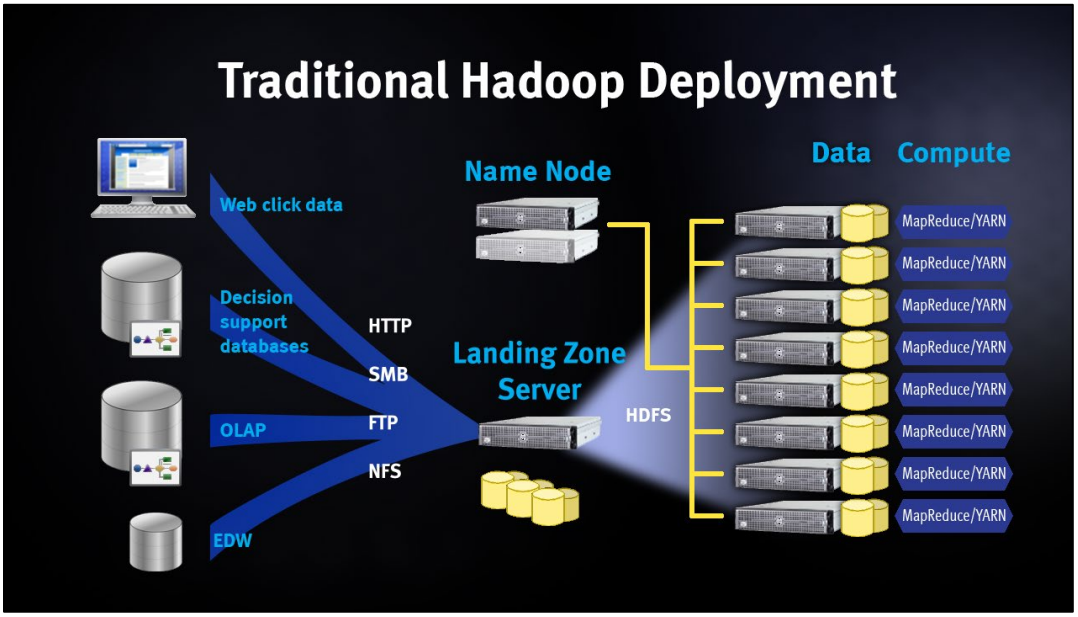
This guide is intended for systems administrators, IT program managers, IT architects, and IT managers who are installing PowerScale OneFS with a Cloudera distribution of Hadoop.

Overview

The PowerScale OneFS scale-out network-attached storage (NAS) platform provides Hadoop clients with direct access to big data through a Hadoop Distributed File System (HDFS) protocol interface. A PowerScale cluster powered by the OneFS operating system delivers a scalable pool of storage with a global namespace.

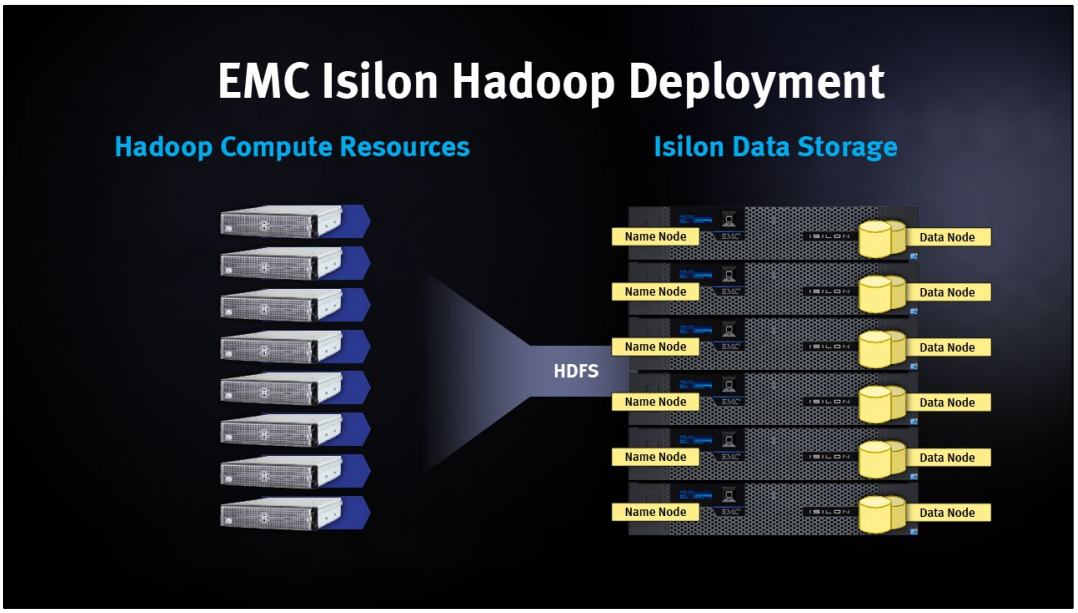
Hadoop compute clients can access the data that is stored on a PowerScale cluster by connecting to any node over the HDFS protocol. All nodes that are configured for HDFS provide NameNode and DataNode functionality. Each node boosts performance and expands the cluster capacity. For Hadoop analytics, the PowerScale scale-out distributed architecture minimizes bottlenecks, rapidly serves big data, and optimizes performance for MapReduce jobs.

In a traditional Hadoop deployment, the Hadoop compute nodes run analytics jobs against large sets of data. A NameNode directs the compute nodes to the data stored on a series of DataNodes. The NameNode is a separate server that holds metadata for every file that is stored on the DataNodes. Often data is stored in production environments and then copied to a landing zone server before it is loaded on to HDFS. This process is network intensive and exposes the NameNode as a potential single point of failure.



In a PowerScale OneFS cluster with Hadoop deployment, OneFS serves as the file system for Hadoop compute clients. On a OneFS cluster, every node in the cluster acts as a NameNode and DataNode, providing automated failover protection.

When a Hadoop client runs a job, the clients access the data that is stored on a OneFS cluster by connecting over HDFS. The HDFS protocol is native to the OneFS operating system, and no data migration is required.



The Cloudera distribution is stored on a separate compute cluster, and individual clients connect directly to the OneFS cluster to store and access Hadoop data. OneFS handles HDFS file data exchange as a protocol in order to store and retrieve the data to match the requirements of the client.

Updates and additional information about OneFS Hadoop installs

The rapid release of new features and versions of Hadoop projects can introduce new behaviors and requirements. It is recommended that you review the latest updates on the Using Hadoop with Isilon - Isilon Info Hub for updates and known issues while deploying OneFS and Hadoop.

Prerequisites

For supported versions, see Hadoop Distributions and Products Supported by OneFS.

Cloudera Distribution with Apache Hadoop (CDH)

Ensure that the following requirements are met:

- CDH 5 parcel: 5.7.1-1.cdh5.7.1.p0.11or later
- Familiarity with the Cloudera documentation and the installation instructions
    - To view the Cloudera documents, go to http://www.cloudera.com/documentation.html
    - Use the following table to record the components that you plan to install

Component	Version
Cloudera Manager version	
CDH parcel version	
Cloudera server (FQDN)	

OneFS cluster configuration

Ensure that the following requirements are met:

- A OneFS cluster running OneFS 8.1.2.0 or later.
- SmartConnect Advanced, a separately licensed module, is activated and SmartConnect is configured on your OneFS cluster.
- HDFS, a separately licensed module, is activated on your OneFS cluster. Contact your Dell EMC PowerScale sales representative for more information about receiving your license keys.
- A valid OneFS SmartConnect SSIP and Domain Name System (DNS) delegation is in place to provide name resolution services for a SmartConnect zone. For more information, see Isilon External Network Connectivity Guide.
- A dedicated OneFS Access Zone is in use; this is not the same as the System Zone.
- A OneFS HDFS root directory in the Access Zone.
- A simple access model between Hadoop and OneFS; UID and GUID, with parity.
- Use the following table to record the components that you have installed.

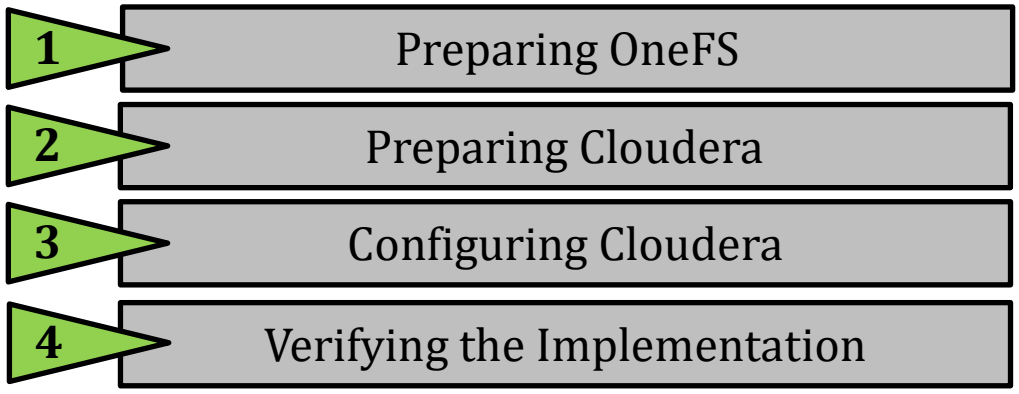
Component	Version or License
PowerScale OneFS	

Component	Version or License
SmartConnect module	
HDFS module	
OneFS cluster name	

Installing OneFS with Cloudera Manager

The installation of OneFS with Cloudera can be separated into four stages as represented in the following figure.

To complete each stage, you must perform tasks on both the Cloudera cluster and the OneFS cluster.



Preparing OneFS

Complete the following steps to configure your OneFS cluster for use with Cloudera Data Platform. Preparing OneFS requires you to configure DNS, SmartConnect, and Access Zones to allow for the Hadoop cluster to connect to the OneFS cluster. If these preparation steps are not successful, the subsequent configuration steps might fail.

Review the current Isilon OneFS and Hadoop Known Issues for any changes or updates to OneFS and Hadoop configuration.

Validate OneFS version and license activation

Validate your OneFS version, check your licenses, and confirm that they are activated. Other OneFS licenses may be needed for additional OneFS functionality to be interoperable with HDFS, they are not addressed in this installation guide.

1. From a node in your OneFS cluster, confirm that the cluster is running OneFS 8.1.2 or later by typing the following command:

   ```
   isi version
   ```

2. Add the licenses for HDFS using the following command:

   ```
   isi license add --evaluation=HDFS
   ```

3. Confirm that the license for HDFS is operational. If this license is not active and valid, some commands in this guide will not work.

   Run the following commands to confirm that HDFS is installed:

   ```
   isi license licenses list
   isi license licenses view HDFS
   ```

4. If your modules are not licensed, obtain a license key from your Dell EMC PowerScale sales representative. Type the following command to activate the license:

   ```
   isi license add --path <license file path>
   ```

5. Enable HDFS by running the following command:

   ```
   isi services hdfs enable
   ```

6. Install the latest rollup patches for your version of OneFS. See Current Isilon OneFS Patches for the latest rollup patches and run the following:

   ```
   isi upgrade patches list
   isi upgrade patches install patch-<patch-ID>.pkg --rolling=false
   ```

   Example:

   ```
   isi upgrade patches install patch-240163.pkg --rolling=false
   ```

Configure OneFS components

After you configure DNS for OneFS, set up and configure the following OneFS components.

- Create an access zone
- Create a SmartConnect zone
- Create and configure the HDFS root in the access zone
- Create users and groups
- Create a basic HDFS folder structure for use with HDFS

Use the following table to record the configuration information for the OneFS cluster with Cloudera integration:

Parameter	Value
Access zone name	
Access zone path	
SmartConnect zone name (FQDN)	
IP range for IP pool (ranges)	
SmartConnect pool name (subnet pool)	
Node and interfaces in the pool	
HDFS root path	

Create an Access Zone

On one of the OneFS nodes, you must define an access zone on the OneFS cluster and enable the Hadoop node to connect to it.

1. On a node in the OneFS cluster, create your Hadoop access zone:

```
isi zone zones create --name=zone1-cdh --path=/ifs/data/zone1/cdh --create-path
```

2. Verify that the access zones are set up correctly:

```
isi zone zones list --verbose
```

Output similar to the following displays:

```
                Name: System
                Path: /ifs
           Groupnet: groupnet0
      Map Untrusted: -
     Auth Providers: lsa-local-provider:System, lsa-file-provider:System
       NetBIOS Name: -
  User Mapping Rules: -
Home Directory Umask: 0077
  Skeleton Directory: /usr/share/skel
  Cache Entry Expiry: 4H
            Zone ID: 1
-------------------------------------------------------------------------------
                Name: zone1-cdh
                Path: /ifs/date/zone1/cdh
           Groupnet: groupnet0
      Map Untrusted: -
     Auth Providers: lsa-local-provider:zone1-cdh
       NetBIOS Name: -
  User Mapping Rules: -
Home Directory Umask: 0077
  Skeleton Directory: /usr/share/skel
  Cache Entry Expiry: 4H
            Zone ID: 2
```

3. Create the HDFS root directory within the access zone that you created:

```
mkdir -p /ifs/data/zone1/cdh/hadoop-root

isi hdfs settings modify --zone=zone1-hdp --root-
directory=/ifs/data/zone1/cdh/hadoop-root
```

4. List the contents of the Hadoop access zone root directory:

```
ls -al /ifs/data/zone1/cdh
```

Configure SmartConnect

On a node in the OneFS cluster, add a static IP address pool and associate it with the access zone you created earlier.

1. Modify your existing subnets and specify a service address:

```
isi network subnets modify groupnet0.subnet0 --sc-service-addr=x.x.x.x
```

2. Create an access network pool, run the following command, where:
   - `<groupnet>:<subnet>:<name>` is the new IP pool in subnet (for example, subnet0:pool1).
   - `<IP-IP>` is the IP range that is assigned to the IP pool
   - `<access-zone>` is the access zone that the pool is assigned to
   - `<interfaces>` are the node interfaces that are added to the pool
   - `<subnet>` is the SmartConnect service subnet that is responsible for this zone
   - `<smartconnectzone>` is the SmartConnect zone name

```
isi network pools create --id=<groupnet>:<subnet>:<name> --ranges=<IP-IP> --
access-zone=<access-zone> --alloc-method=static --ifaces=<interfaces> --sc-
subnet=<subnet> --sc-dns-zone=<smartconnectzone> --description=hadoop
```

For example:

```
isi network pools create groupnet0:subnet0:hadoop-pool-cdh --
ranges=10.120.130.30-10.120.140.40 --access-zone=zone1-cdh --alloc-
method=static --ifaces=1-4:40gige-1 --sc-subnet=subnet0 --sc-dns-
zone=cdh.zone1.emc.com --description=hadoop"
```

3. View the properties of the access network pool:

```
isi network pools view --id=groupnet0:subnet0:pool2
```

Output similar to the following displays:

```
                    ID: groupnet0.subnet0.hadoop-pool-cdh
             Groupnet: groupnet0
               Subnet: subnet0
                 Name: hadoop-pool-cdh
                Rules: -
          Access Zone: zone1-cdh
   Allocation Method: static
    Aggregation Mode: lacp
  SC Suspended Nodes: -
         Description: cdh_hadoop_access_zone
```

```
                Ifaces: 1:ext-1, 2:ext-1, 3:ext-1, 4:ext-1
             IP Ranges: 10.120.130.30-10.120.140.40
       Rebalance Policy: auto
SC Auto Unsuspend Delay: 0
      SC Connect Policy: round_robin
               SC Zone: cdh.zone1.emc.com
    SC DNS Zone Aliases: -
     SC Failover Policy: round_robin
              SC Subnet: subnet0
                 SC Ttl: 0
          Static Routes: -
```

Configure DNS for OneFS

Before you begin, the OneFS cluster must already be implemented according to Dell EMC PowerScale best practices. For more information, see the HDFS Setup section of the Dell EMC Isilon Best Practices Guide for Hadoop Data Storage.

Set up DNS records for a SmartConnect zone. Create the required DNS records that are used to access your OneFS cluster from the Hadoop cluster. All hosts in your Hadoop cluster must be configured for both forward and reverse DNS lookups Hadoop relies heavily on DNS and performs many DNS lookups during normal operation.

You can set up a SmartConnect zone for the connections from Hadoop compute clients. SmartConnect is a module that specifies how the OneFS cluster handles connection requests from clients. For additional information and best practices for SmartConnect, see the Isilon External Network Connectivity Guide.

Each SmartConnect zone represents a specific pool of IP addresses. When you associate a SmartConnect zone with an access zone, OneFS allows only clients that connect through the IP addresses in the SmartConnect zone to reach the HDFS data in the access zone. A root HDFS directory is specified for each access zone. This configuration isolates data within access zones and allows you to restrict client access to the data.

A SmartConnect zone distributes NameNode requests from Hadoop compute clients across the node interfaces in the IP pool. Each nodes NameNode process replies with the IP address of the HDFS DataNode where the client can access the data. When a Hadoop compute client makes an initial DNS request to connect to the SmartConnect zone FQDN, the Hadoop client requests are delegated to the SmartConnect Service IP, which responds with a valid node to connect to. The client connects to a OneFS node that serves as a NameNode. When a second Hadoop client makes a DNS request to connect to the SmartConnect zone, the SmartConnect Service routes the client connection to a different node than the node that is used by the previous Hadoop compute client.

When you create a SmartConnect zone, you must add a Name Server (NS) record as a delegated domain to the authoritative DNS zone that contains the OneFS cluster.

Verify the SmartConnect configuration

Validate that SmartConnect is set up correctly by pinging the SmartConnect zone FQDN several times from the Hadoop client.

```
ping cdh.zone1.emc.com
```

When you view the output of this command, different IP addresses are returned for each ping command, because with each DNS response, the IP addresses are returned through rotating round-robin DNS from the list of potential IP addresses. This validates that the SmartConnect zone name FQDN is operating correctly.

Create HDFS users and groups

For each Hadoop system account that will submit HDFS jobs or access the file system, you must create local users and groups on the OneFS cluster. You can add Hadoop users and groups to the OneFS cluster manually or by following the process at: https://github.com/Isilon/isilon_hadoop_tools

Important

Dell EMC PowerScale recommends that you maintain consistent names and numeric IDs for all users and groups on the OneFS cluster and your Hadoop clients. This consistency is important in multiprotocol environments because the HDFS protocol refers to users and groups by name, and NFS refers to users and groups by their numeric IDs (UIDs and GUIDs). Maintaining this parity is critical in the behavior of OneFS multiprotocol file access.

During installation of Hadoop with Cloudera Manager, the installer creates all the required system accounts on all the clients. For example, a Hadoop system account, *yarn,* is created with the UID of 502 and the GUID of 502 on the Hadoop cluster nodes, Cloudera will only create these accounts if they do not exist. You can ensure parity by precreating them on all nodes that will be installed in the Hadoop cluster. You can look to enforce parity by manually managing when and how these local system accounts get created. Since the Hadoop installer cannot create the local accounts directly on OneFS, they must be created manually. Create the OneFS *yarn* local account user in the OneFS access zone in which *yarn* accesses data. Create a local user *yarn* with the UID of 502 and the GUID of 502 to ensure consistency of access and permissions.

For guidance and more information about maintaining parity between OneFS and Hadoop local users and UIDs, see the following article: Isilon and Hadoop Local User UID Parity

There are many methods of achieving UID and GUID parity. You can leverage Tools for Using Hadoop with OneFS, perform manual matching, or create scripts that parse users and create the equivalent users. However you choose to achieve this, the sequence will depend on your deployment methodology and management practices. It is highly recommended that you maintain consistency between the Hadoop cluster and OneFS, for example, hdfs=hdfs, yarn=yarn, hbase=hbase, and so on from a UID and GUID consistency perspective.

Create users and directories on the OneFS cluster using Tools for Using Hadoop with OneFS

Go to Tools for Using Hadoop with OneFS to set up the users and directories on the cluster.

Create users on the OneFS cluster manually

You can add a user for each additional Hadoop user that submits MapReduce jobs in addition to the users that the OneFS script configures on the OneFS cluster. The following procedures show how to manually add a single user called *hduser1*.Warning

If your users and groups are defined by your directory service, such as Active Directory or LDAP, do NOT run these commands. This section addresses setting permissions of the HDFS root files or membership to run jobs. These steps create users but will likely fail when you run jobs with this configuration.

Manual steps to perform on the OneFS cluster

1. Add a group to the OneFS cluster.

```
isi auth groups create hduser1 --zone zone1 --provider local --gid <GUID>
```

2. Create the user and the user's Hadoop home directories on the OneFS cluster.

```
isi auth users create hduser1 --primary-group hduser1 --zone zone1 --provider local
--home-directory /ifs/data/zone1/hadoop/user/hduser1 --uid <UID> --enabled=true
```

3. Assign permissions to the user's home directory on the Hadoop cluster. The ID 2 in the example below is from when you previously ran the `isi zone zones view zone1` command.

```
isi_run -z2 chown hduser1:hduser1 /ifs/data/hadoop/user/hduser1

chmod 755 /ifs/data/hadoop/user/hduser1
```

Manual steps to perform on the Hadoop client

Since you created a user on OneFS to run jobs, you must create the same user with UID parity on any Linux hosts that the user accesses to run jobs.

1. Add the user to the Hadoop cluster.

```
adduser hduser1 -u <UID>
```

Create and configure HDFS for OneFS 8.1.2 and previous versions

In OneFS 8.1.2 and previous versions, the HDFS user must be mapped to root and you must modify the access control list (ACL).

On a node in the OneFS cluster, assign the Hadoop Distributed File System (HDFS) root directory.

1. View the HDFS service settings.

```
isi hdfs settings view --zone=zone1-cdh
```

2. Set the HDFS root directory for the access zone. **Note**: It is recommended that the directory for the access zone is not set to the root of /ifs.

```
isi hdfs settings modify --zone=zone1-cdh --root-
directory=/ifs/data/zone1/cdh/hadoop-root
```

3. Map the HDFS user to root. Create a user-mapping rule to map the HDFS user to the OneFS root account. This mapping enables the services from the Hadoop cluster to communicate with the OneFS cluster using the correct credentials.

```
isi zone zones modify --user-mapping-rules="hdfs=>root" --zone zone1-cdh
```

4. Create an indicator file in the Hadoop directory to view your OneFS cluster and access zone through HDFS.

```
touch /ifs/data/zone1/cdh/hadoop-root/THIS_IS_ISILON_zone1-cdh.txt
```

5. Modify the access control list (ACL) settings for OneFS.

   Run the following command to modify ACL settings BEFORE you create directories or files in the next section. This creates the correct permission behavior on the cluster for HDFS.

   Note

   ACL policies are cluster-wide, so you should understand this change before performing it on production clusters.

```
isi auth settings acls modify --group-owner-inheritance=parent

isi auth settings view
```

Configure HDFS user for OneFS 8.2 and later versions

In OneFS 8.2.0, the HDFS user no longer must be mapped to root. Instead a new role with backup and restore privileges must be assigned as follows:

On a node in the OneFS 8.2 cluster, create a role and configure the backup and restore privileges to the HDFS user.

1. View the HDFS service settings.

```
isi hdfs settings view --zone=zone1-cdh
```

2. Set the HDFS root directory for the access zone. **Note**: It is recommended that the directory for the access zone is not set to the root of /ifs.

```
isi hdfs settings modify --zone=zone1-cdh --root-
directory=/ifs/data/zone1/cdh/hadoop-root
```

3. Create a role for the Hadoop access zone.

```
isi auth roles create --name=<role_name> --description=<role_description> --
zone=<access_zone>
```

   For example:

```
isi auth roles create --name=HdfsAccess --description="Bypass FS permissions" --
zone=zone1-cdh
```

4. Add restore privileges to the new “HdfsAccess” role.

```
isi auth roles modify <role_name> --add-priv=ISI_PRIV_IFS_RESTORE --
zone=<access_zone>
```

   For example:

```
isi auth roles modify HdfsAccess --add-priv=ISI_PRIV_IFS_RESTORE --zone=zone1-cdh
```

5. Add backup privileges to the new “HdfsAccess” role.

```
isi auth roles modify <role_name> --add-priv=ISI_PRIV_IFS_BACKUP --
zone=<access_zone>
```

   For example:

```
isi auth roles modify HdfsAccess --add-priv=ISI_PRIV_IFS_BACKUP --zone=zone-cdh
```

6. Add user hdfs to the new “HdfsAccess” role.

```
isi auth roles modify <role_name> --add-user=hdfs --zone=<access_zone>
```

   For example:

```
isi auth roles modify HdfsAccess --add-user=hdfs --zone=zone1-cdh
```

7. Verify the role setup, backup and restore privileges, and HDFS user setup.

```
isi auth roles view <role_name> --zone=<access_zone>
```

   For example:

```
isi auth roles view HdfsAccess --zone=zone1-cdh
Name: HdfsAccess
Description: Bypass FS permissions
Members: - hdfs
Privileges
```

```
ID: ISI_PRIV_IFS_BACKUP
Read Only: True
ID: ISI_PRIV_IFS_RESTORE
Read Only: True
```

8. (Optional) Flush auth mapping and auth cache to make the HDFS user take immediate effect as the “HdfsAccess” role that you created above.

```
isi_for_array "isi auth mapping flush --all"
isi_for_array "isi auth cache flush --all"
```

Note

ACL Policies no longer must be modified for OneFS 8.2 and later as the HDFS protocols act the same as non-OneFS HDFS for File System Group Owner inheritance.

Preparing Cloudera

The steps in this stage occur on the Cloudera hosts that become your Hadoop servers and clients.

Hadoop clusters and services rely heavily on DNS. All client hosts in your system must be configured for both forward and reverse DNS lookups. Validate that all hosts can resolve the other hostnames and IP addresses. For additional information, see the Cloudera installation documentation.

Deploy Cloudera Manager

To prepare Cloudera Manager for configuration, follow the instructions for your version of Cloudera in the *Cloudera Installation and Upgrade Guide*. You can find a thorough overview of the installation procedure on the Cloudera site. This procedure begins with the download of the bits and installation of Cloudera Manager.

For an overview of Cloudera, see the following: Overview of Cloudera and the Cloudera Documentation Set.

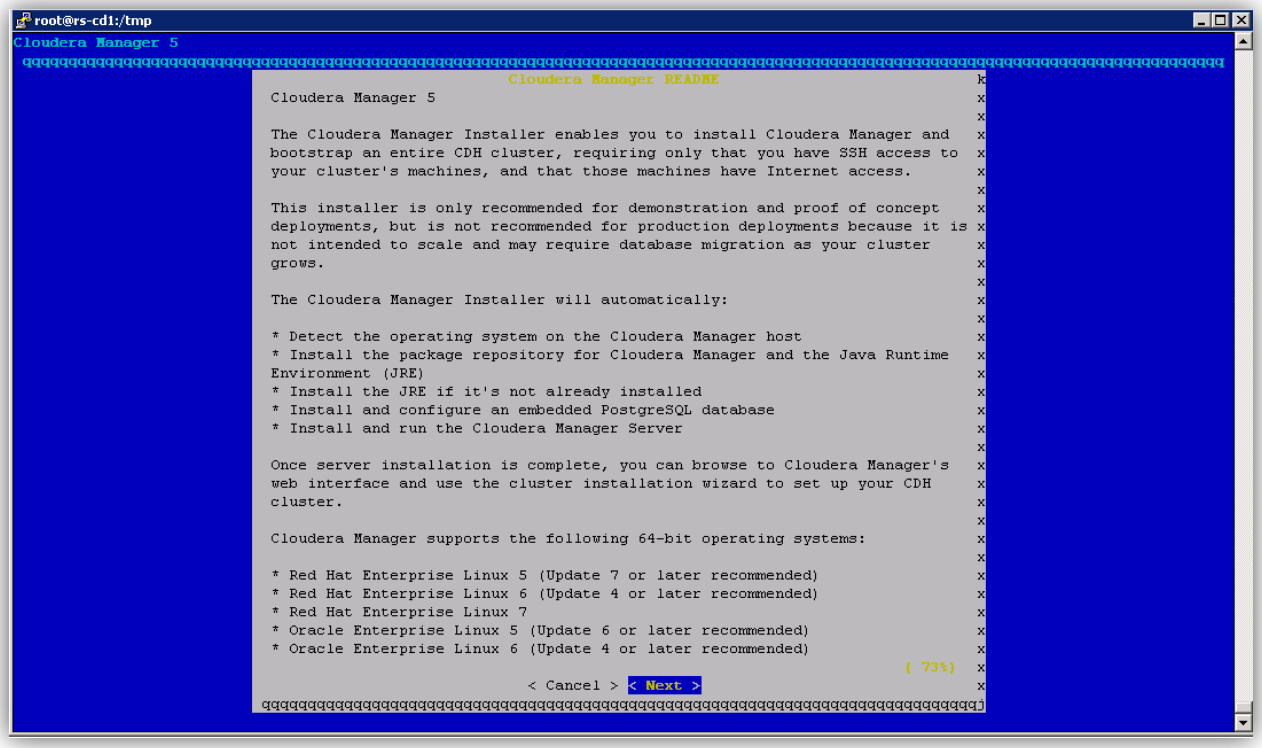
Install Cloudera Manager. Before you begin the installation of Cloudera, ensure that all your hosts meet the Cloudera requirements to complete a successful Hadoop cluster installation. For more information and these installation guides, go to the Overview of Cloudera and the Cloudera Documentation Set.

Before installing any Hadoop cluster, you should consult Hadoop Distributions and Products Supported by OneFS for Cloudera Manager and CDH compatibility.

1. Run the following to install Cloudera.

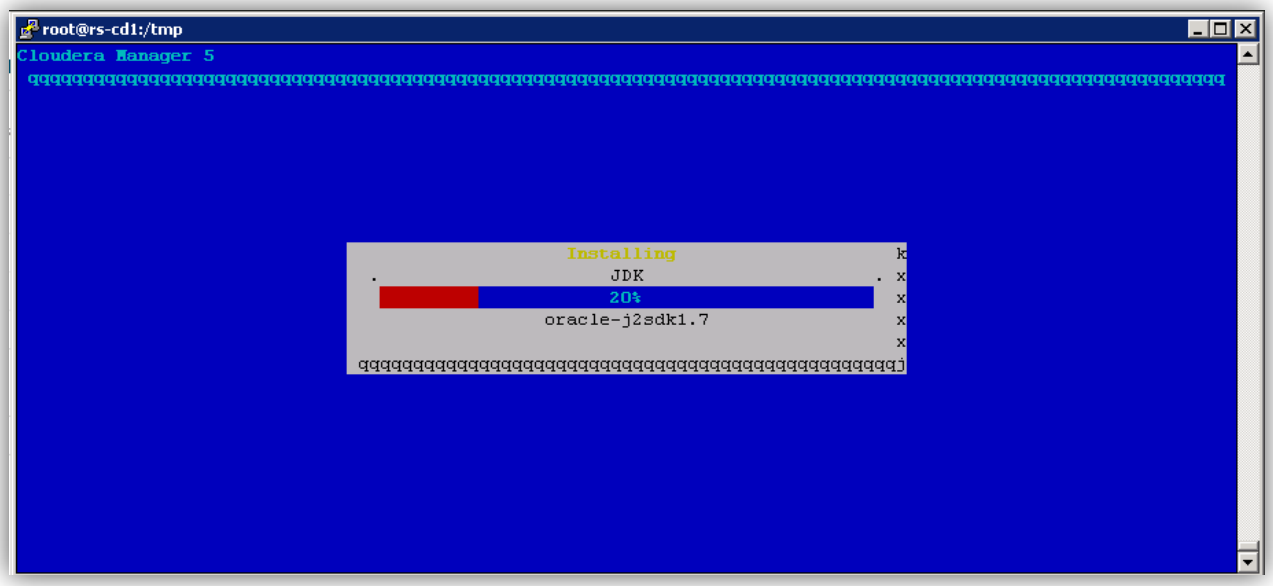
```
./cloudera-manager-installer.bin
```

2. The following displays while running the installer:

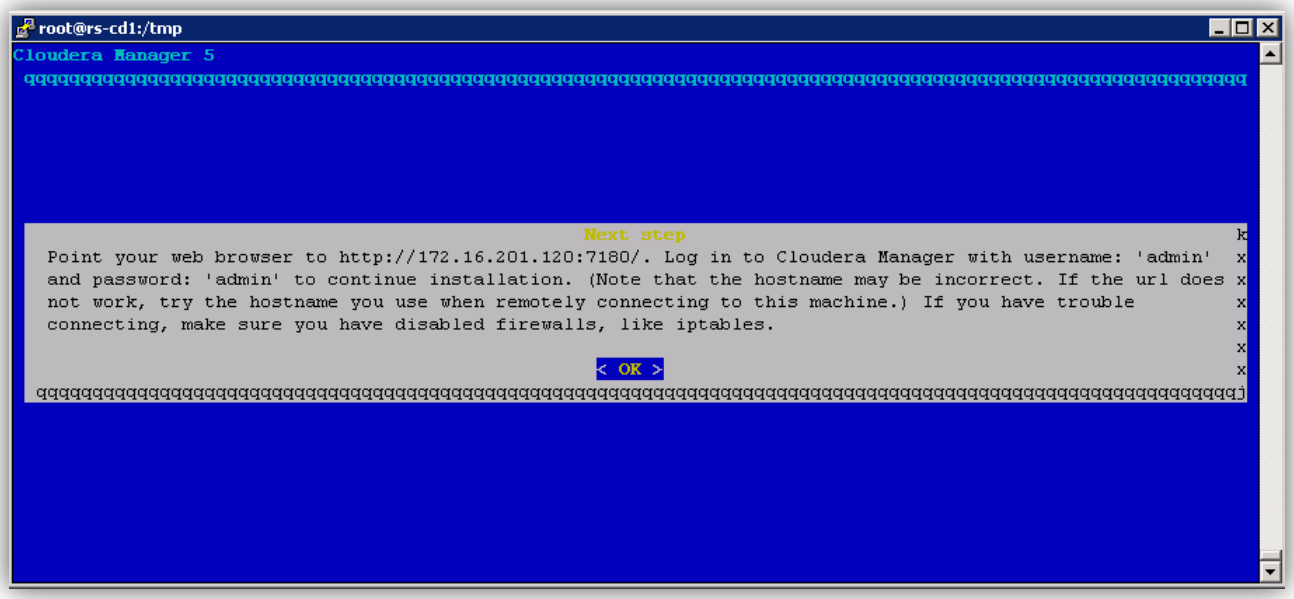


3. Review the Cloudera license and accept the license agreement.
4. Accept the Oracle Binary Code license agreement.
5. Review and accept the Oracle Binary Code license agreement.

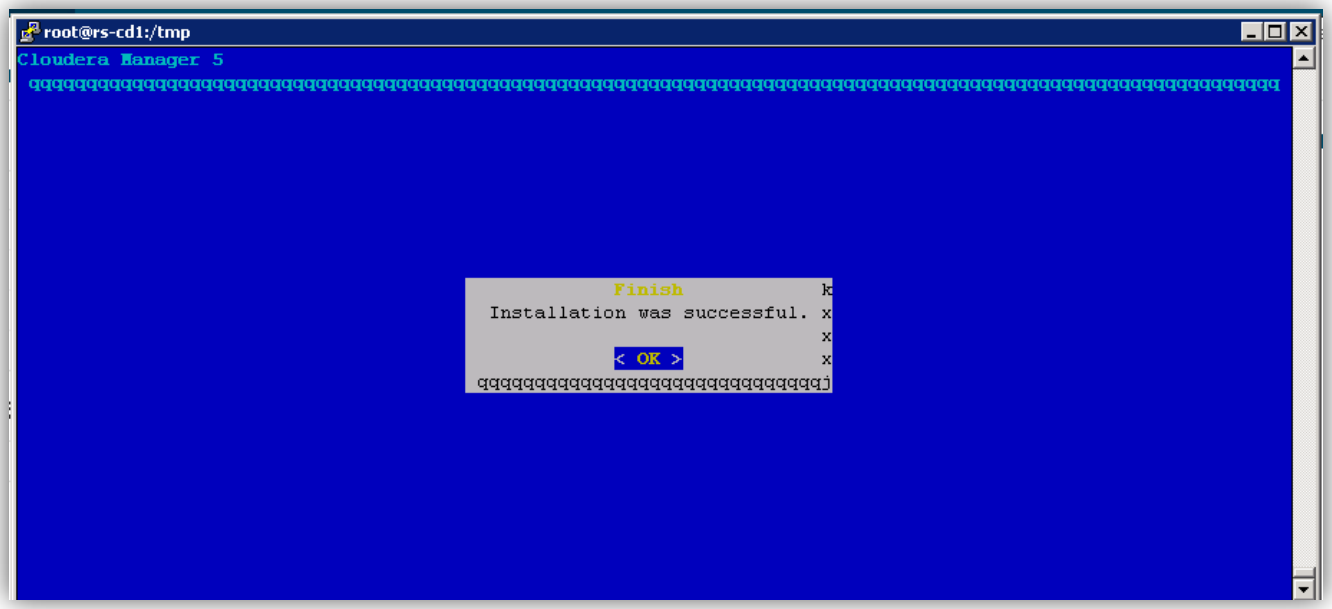
6. Install the Oracle JDK.



7. After you install the Oracle JDK, the following screen appears. Note the URL and the username and password for the Cloudera Manager WebUI.



8. Your installation should now be complete.



9. Validate that the Cloudera Manager Service is running.

```
service cloudera-scm-server status
```

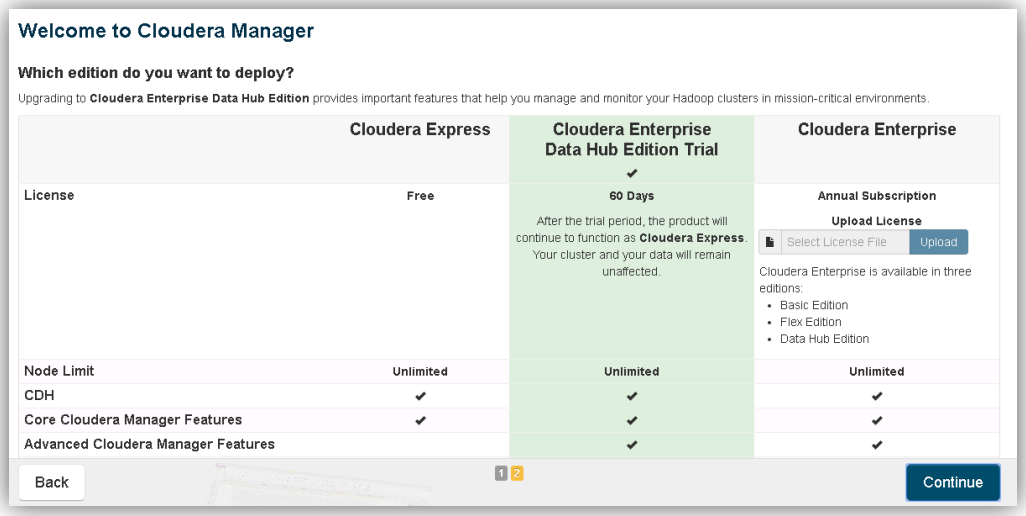
   A valid response is:

```
cloudera-scm-server (pid 10487) is running...
```

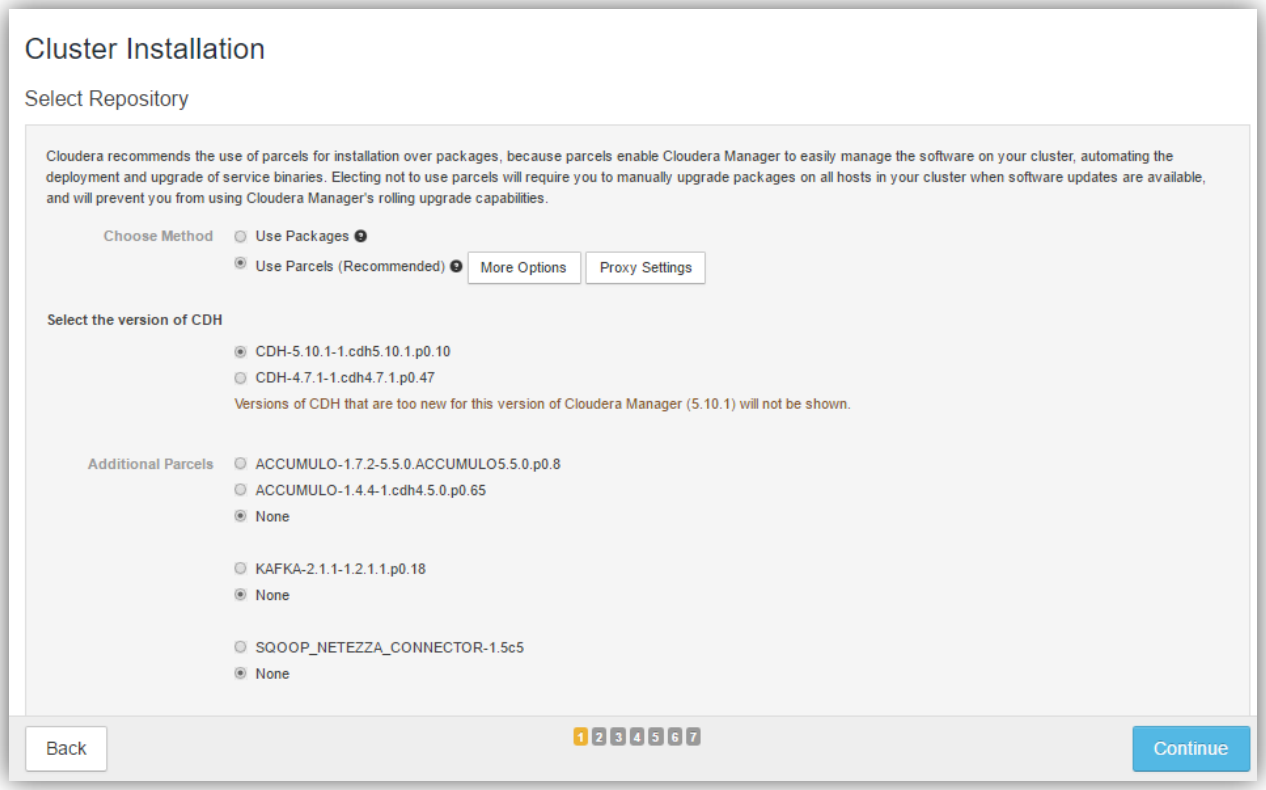
   Look for problems at the end of the `cloudera-scm-server.log` file.

```
tail -f /var/log/cloudera-scm-server/cloudera-scm-server.log
```

10. Log in to the Cloudera Manager WebUI (http://<IP of CM Server>:7180) with the following credentials: **user**: admin; **password**: admin and accept the EULA.
11. Select the version of Cloudera that you want to deploy.

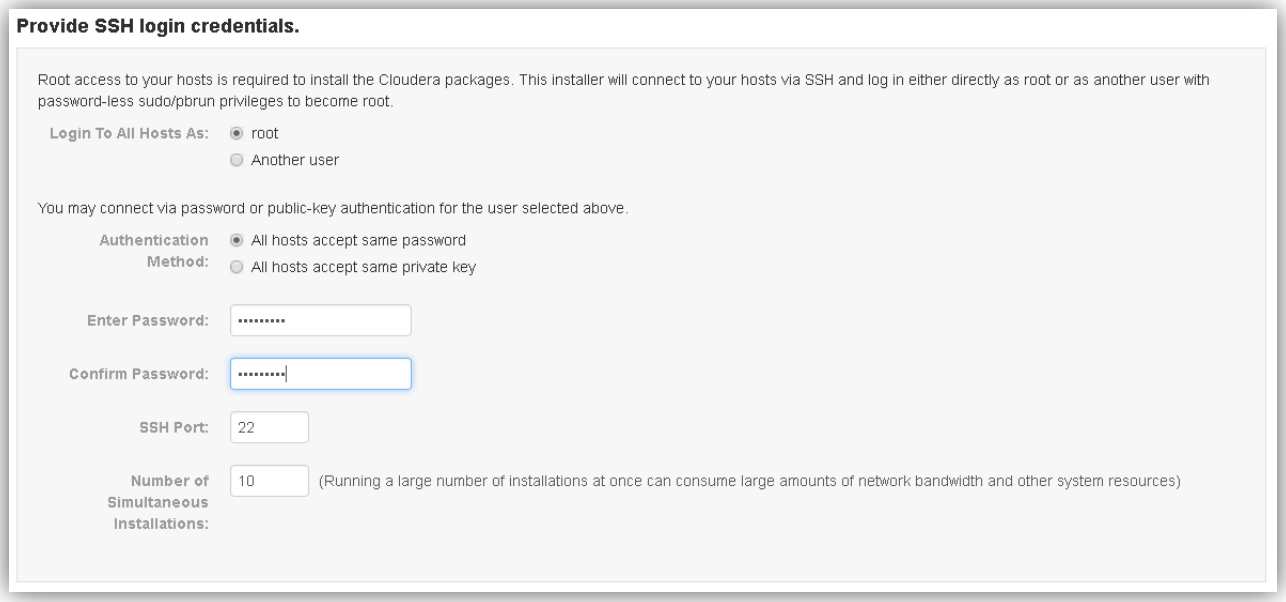


12. Specify the hosts for your CDH cluster installation. In this guide, you deploy to a single Linux host, but the process is the same when multiple hosts are used in the Hadoop cluster. Hosts should be specified using the same hostname (FQDN) that they identify themselves with.
    - Select **Use Parcels** (recommended), the CDH stack that you want to deploy, any additional parcels, and then click **Continue**.

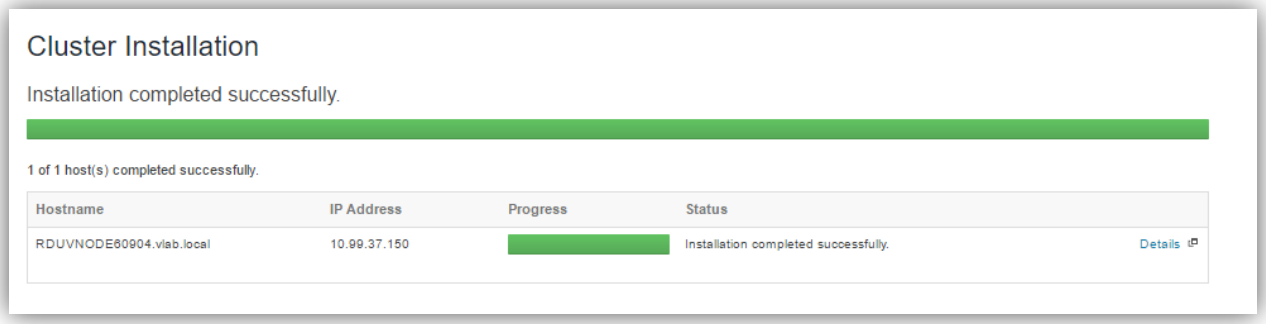


13. Install the Oracle Java Development Kit (JDK) and install the Java Unlimited Strength Encryption Policy (JUSEP) files to secure the cluster.
14. Do NOT select **Single User Mode**.

15. Provide the SSH credentials, either root password or SSH keys, depending on how you configure management of your Linux hosts. The installation of the Cloudera Management package initiates.



16. Wait for the installation to complete and then click **Continue**.



Note

The installer checks and validates hosts. If the validation check fails, follow the recommendations to resolve and then retry the validation. Common errors are the following, which are all related to Linux (not OneFS).

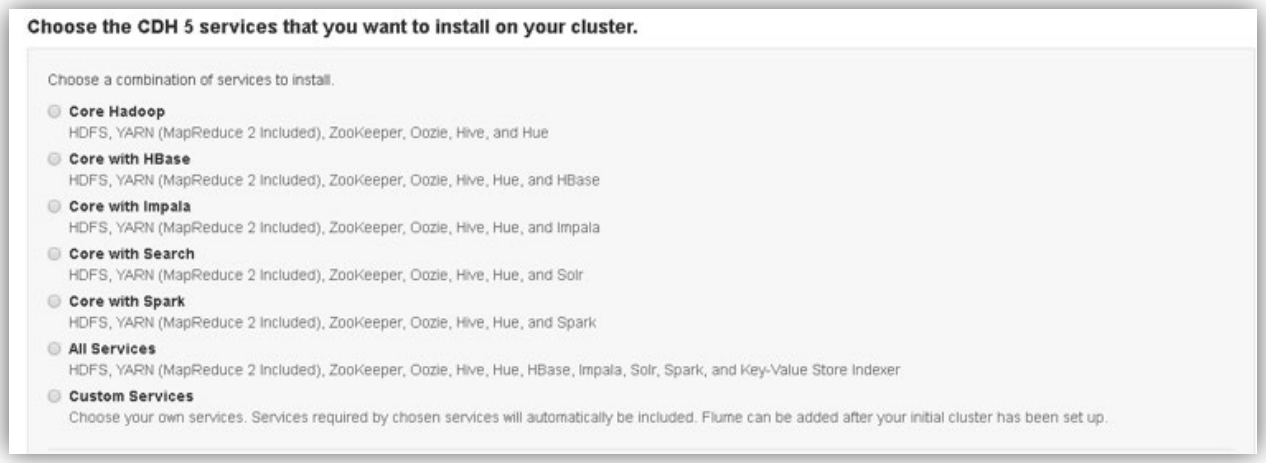
- Transparent_hugepage
- Swappiness

On successful completion of the host inspector, click **Continue**.

17. Click **Finish.**

Configure the Hadoop cluster

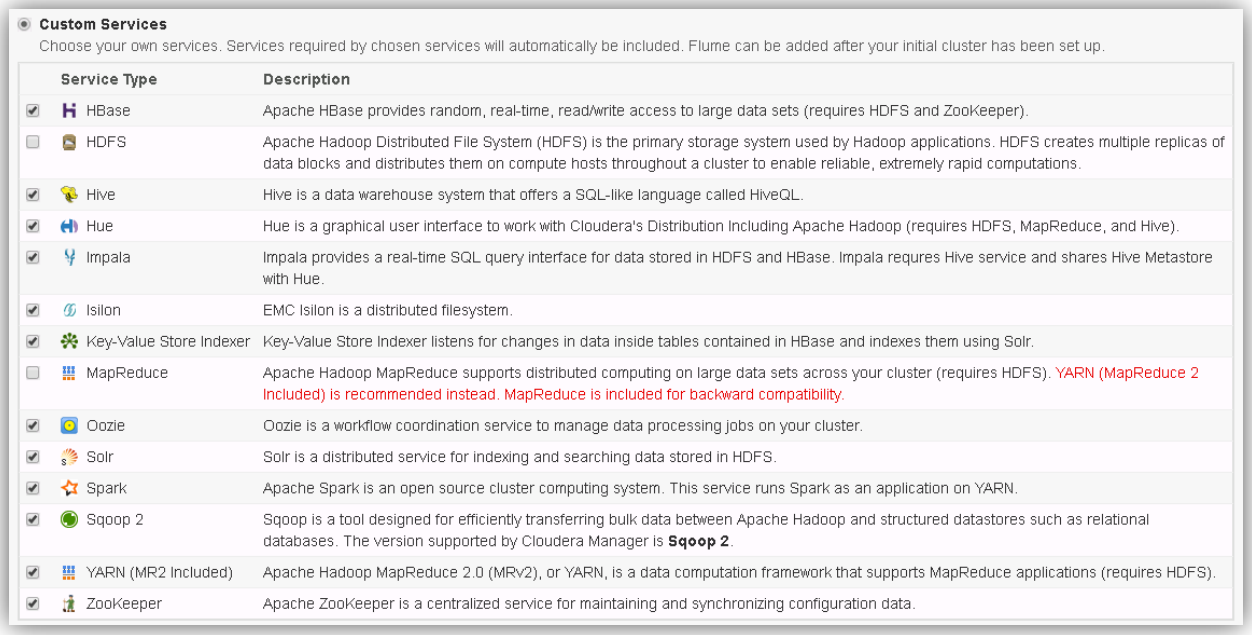
1. **IMPORTANT:** Select the **Custom Services** option on the **Choose the CDH 5 services that you want to install on your cluster** screen to deploy the Hadoop cluster with OneFS. This is key to the OneFS integration. If you select anything other than **Custom Services,** you cannot install OneFS.



2. Once you select **Custom Services**, select the Hadoop services that you want to deploy.

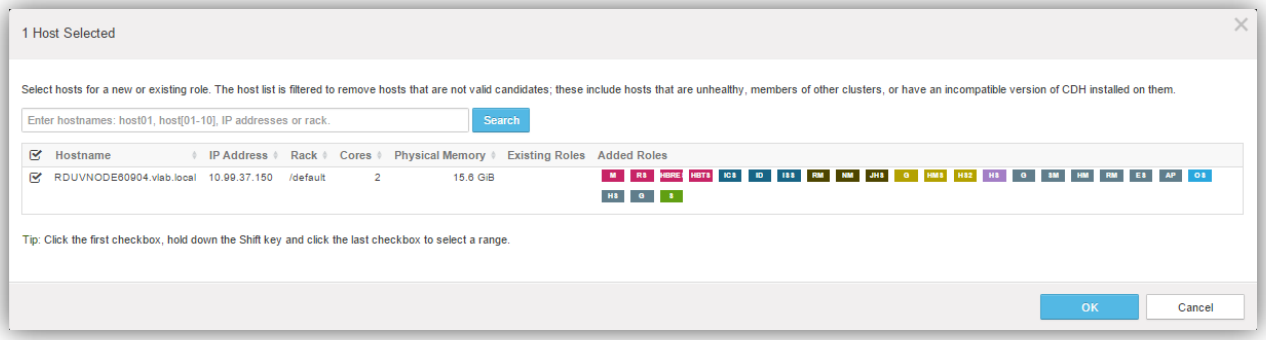
Important

- **Do not select HDFS**. The Cloudera HDFS service is not needed.
- Select **Isilon** as the storage.
- It is not recommended to select MapReduce, as MapReduce2 is depreciated and is in Yarn. If you have a legacy application and it is not written for Yarn, MapReducev1 can be enabled.

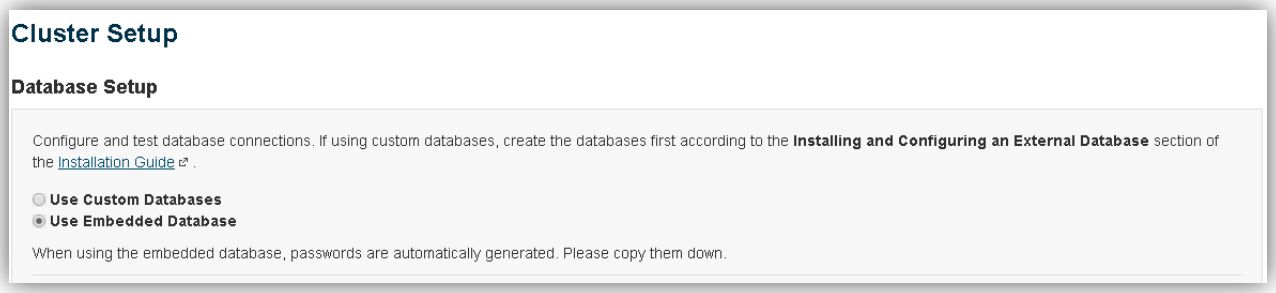


3. After you have selected the Hadoop services that you need, continue to assign roles on the **Customize Role Assignments** screen. Since this is a single host, all roles are deployed on the same host. Assign the hosts, per the Cloudera documentation and depending on your configuration.
4. Select the Isilon Gateway role on the same host that is running Cloudera Manager, and then click **Continue**.

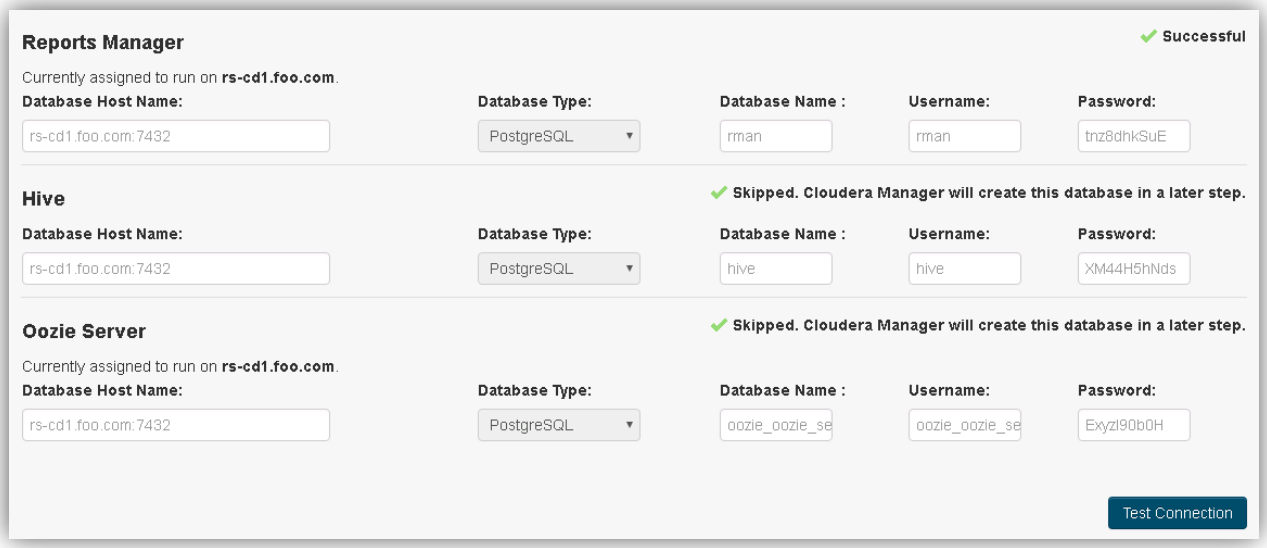
5. View the selected hosts.



6. Continue with the cluster database setup. Select **Use Embedded Database** and select the default settings. Consult the Cloudera documentation for best practices regarding database configuration.

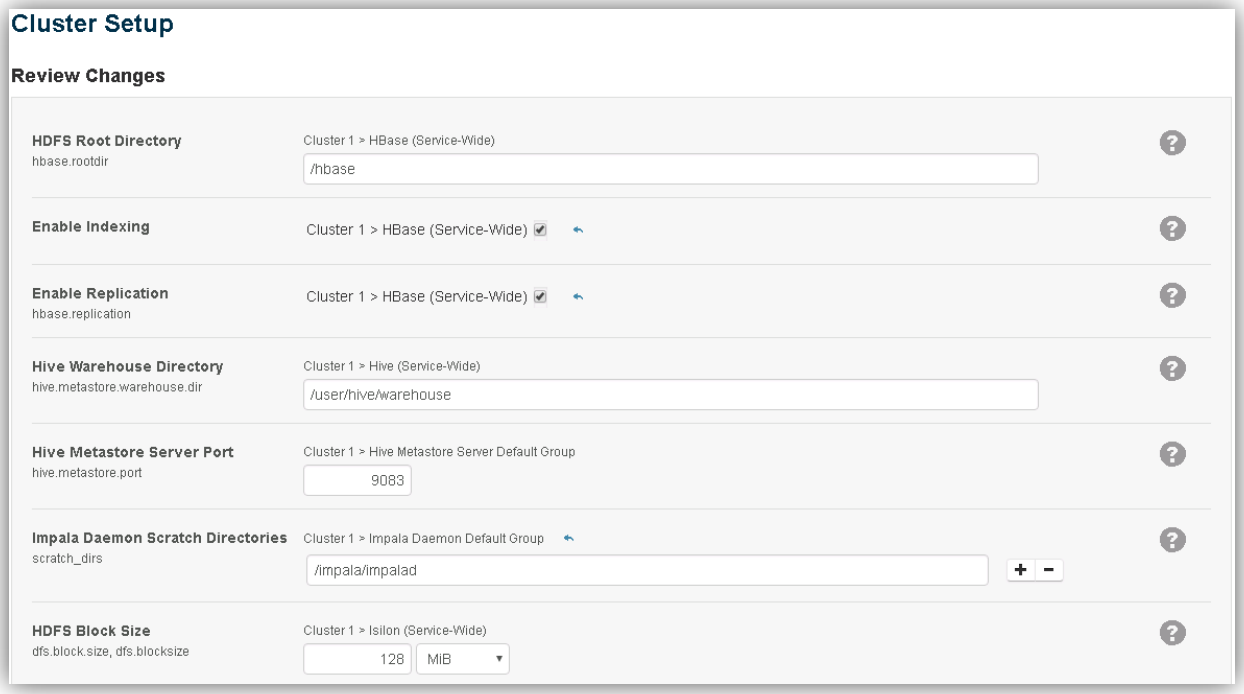


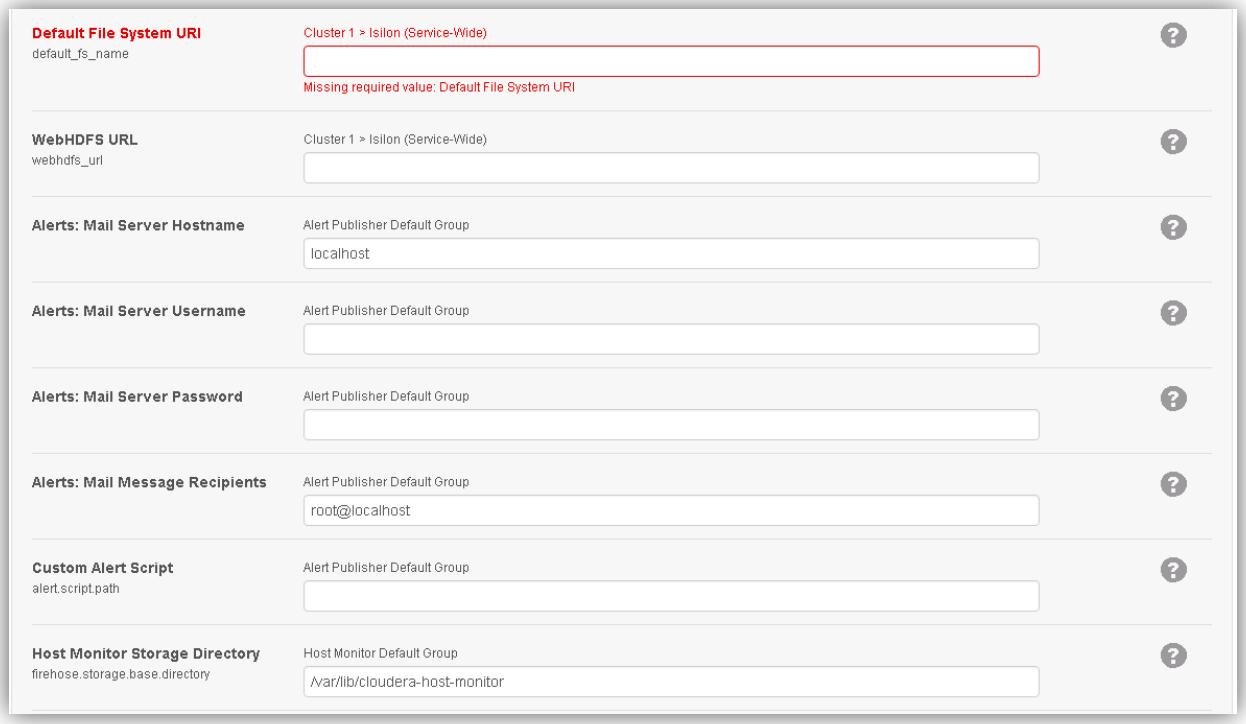
7. Test the database connections, and then click **Continue**.

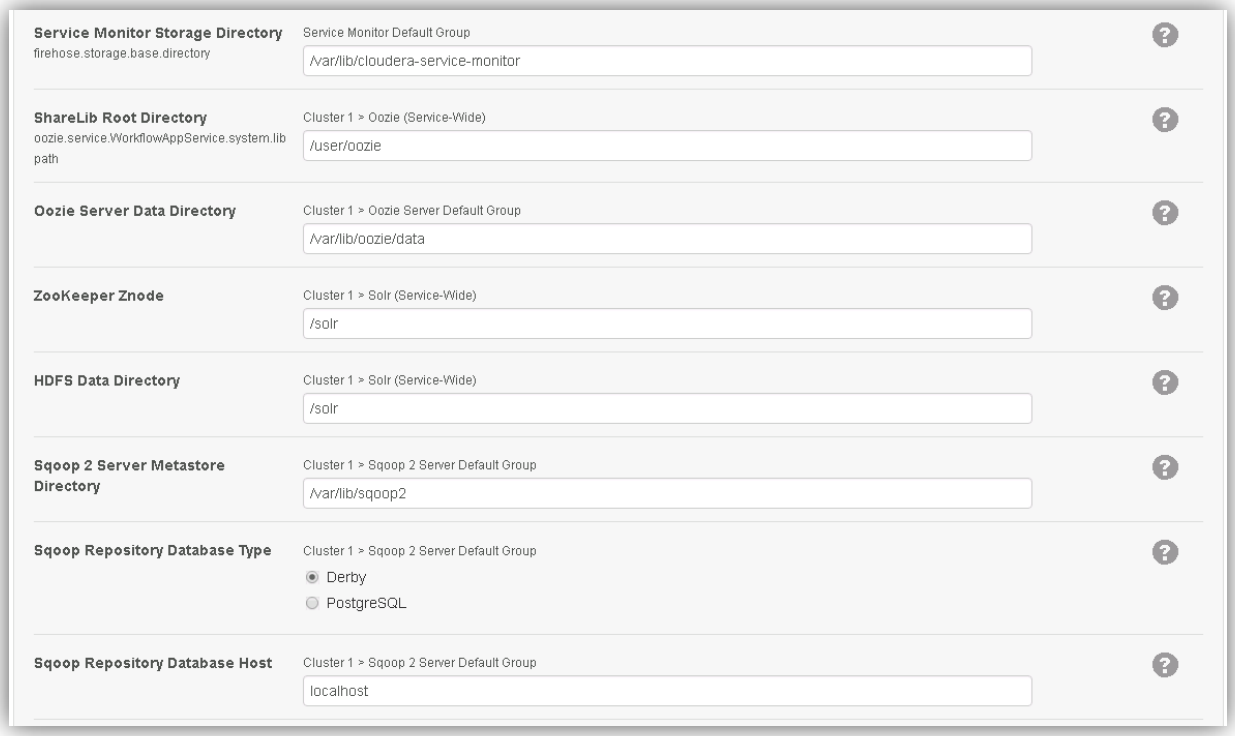


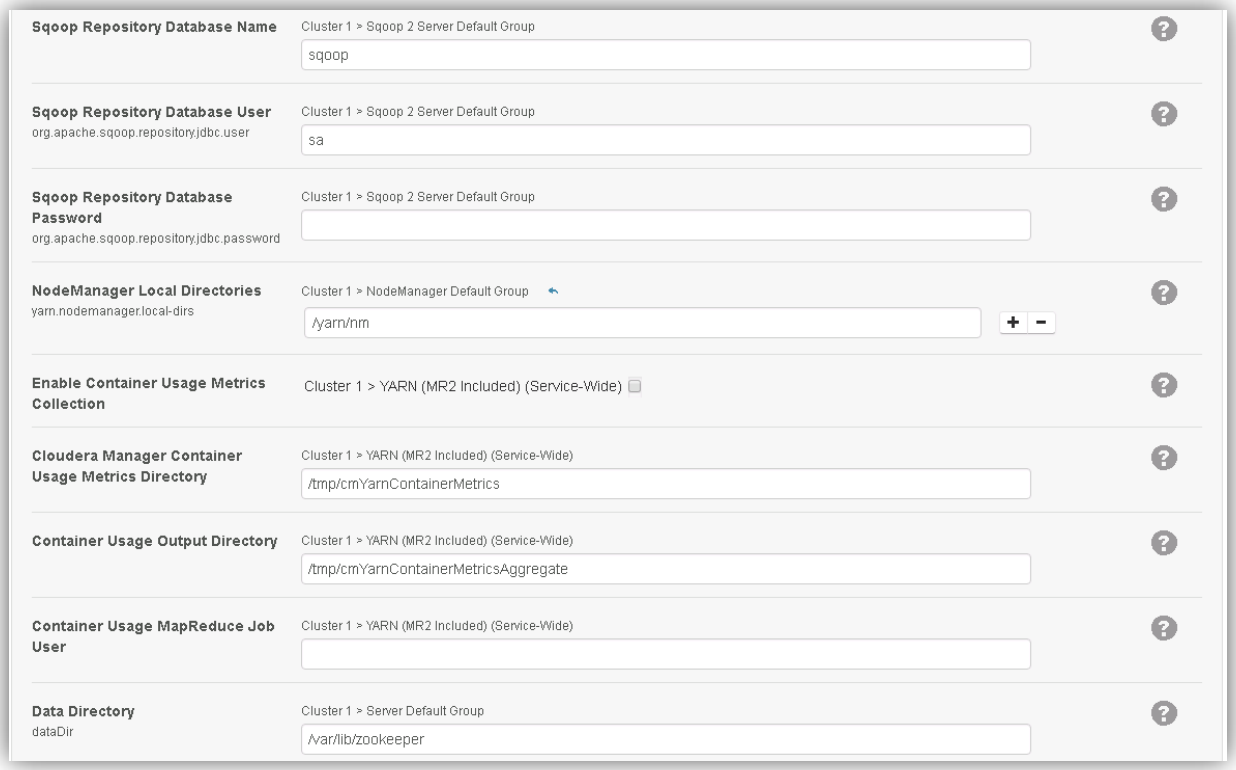
8. Continuing with the **Cluster Setup**, assign the OneFS cluster to the following parameters. The parameters for the SmartConnect zone ports are different.

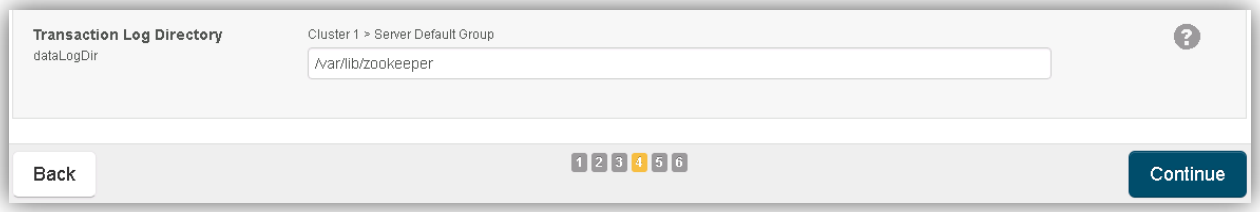
```
default_fs_name          hdfs://smartconnectzonename:8020
webhdfs_url              http://smartconnectzonename:8082/webhdfs/v1
```



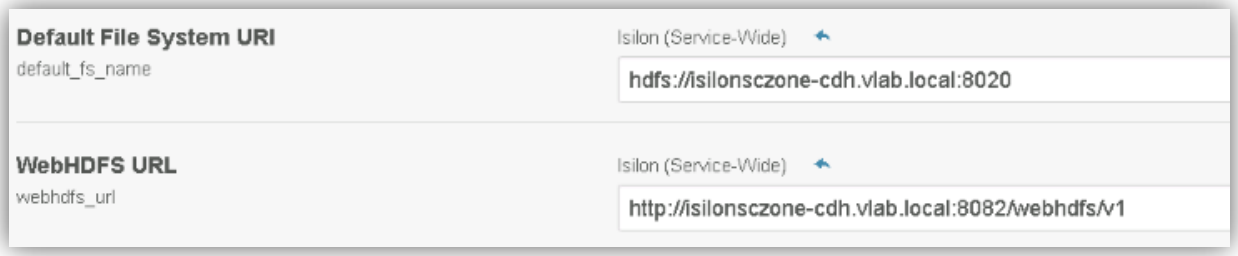



Service Monitor Storage Directory
firehose.storage.base.directory
Service Monitor Default Group
/var/lib/cloudera-service-monitor
ShareLib Root Directory
oozie.service.WorkflowAppService.system.libpath
Cluster 1 > Oozie (Service-Wide)
/user/oozie
Oozie Server Data Directory
Cluster 1 > Oozie Server Default Group
/var/lib/oozie/data
ZooKeeper Znode
Cluster 1 > Solr (Service-Wide)
/solr
HDFS Data Directory
Cluster 1 > Solr (Service-Wide)
/solr
Sqoop 2 Server Metastore Directory
Cluster 1 > Sqoop 2 Server Default Group
/var/lib/sqoop2
Sqoop Repository Database Type
Cluster 1 > Sqoop 2 Server Default Group
Derby
PostgreSQL
Sqoop Repository Database Host
Cluster 1 > Sqoop 2 Server Default Group
localhost

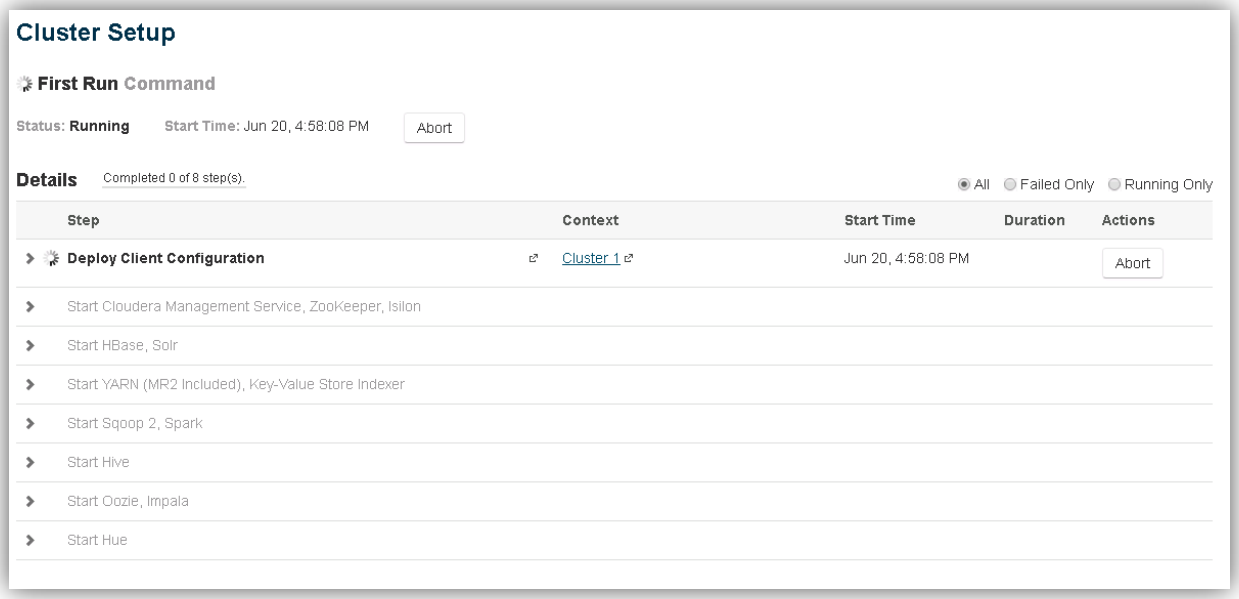

Sqoop Repository Database Name
Cluster 1 > Sqoop 2 Server Default Group
sqoop
Sqoop Repository Database User
org.apache.sqoop.repository.jdbc.user
Cluster 1 > Sqoop 2 Server Default Group
sa
Sqoop Repository Database Password
org.apache.sqoop.repository.jdbc.password
Cluster 1 > Sqoop 2 Server Default Group
NodeManager Local Directories
yarn.nodemanager.local-dirs
Cluster 1 > NodeManager Default Group
/yarn/nm
Enable Container Usage Metrics Collection
Cluster 1 > YARN (MR2 Included) (Service-Wide)
Cloudera Manager Container Usage Metrics Directory
Cluster 1 > YARN (MR2 Included) (Service-Wide)
/tmp/cmYarnContainerMetrics
Container Usage Output Directory
Cluster 1 > YARN (MR2 Included) (Service-Wide)
/tmp/cmYarnContainerMetricsAggregate
Container Usage MapReduce Job User
Cluster 1 > YARN (MR2 Included) (Service-Wide)
Data Directory
dataDir
Cluster 1 > Server Default Group
/var/lib/zookeeper



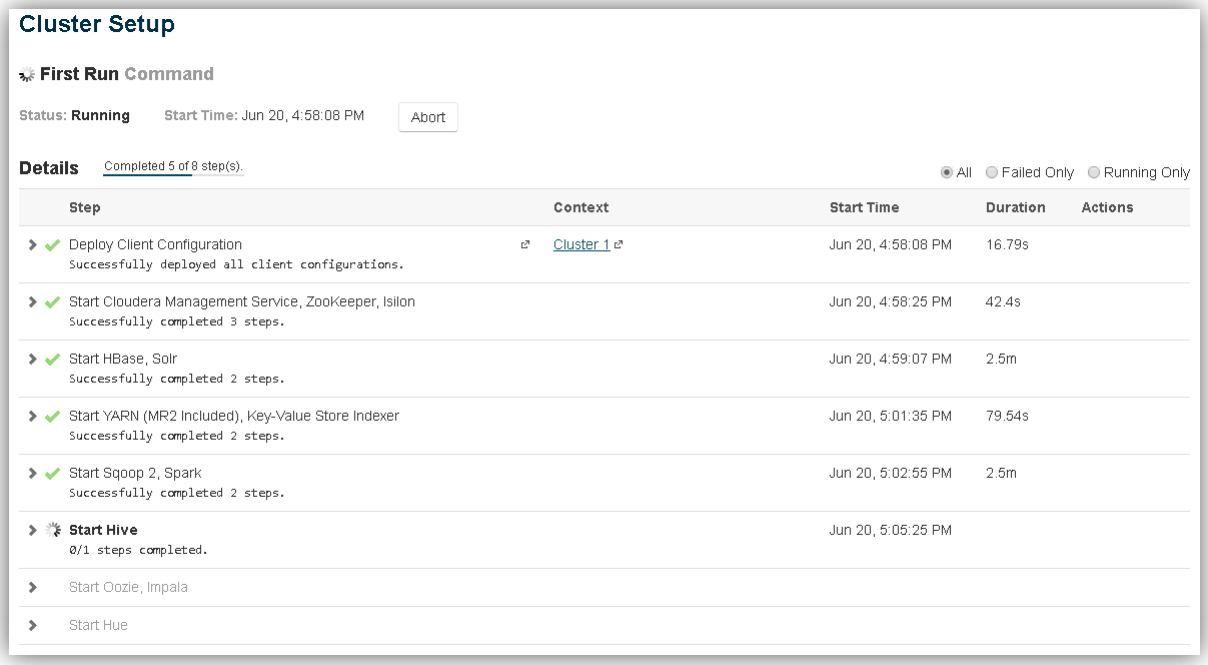
9. Assign the two Isilon parameters. Leave all the default settings, and then click **Continue**.

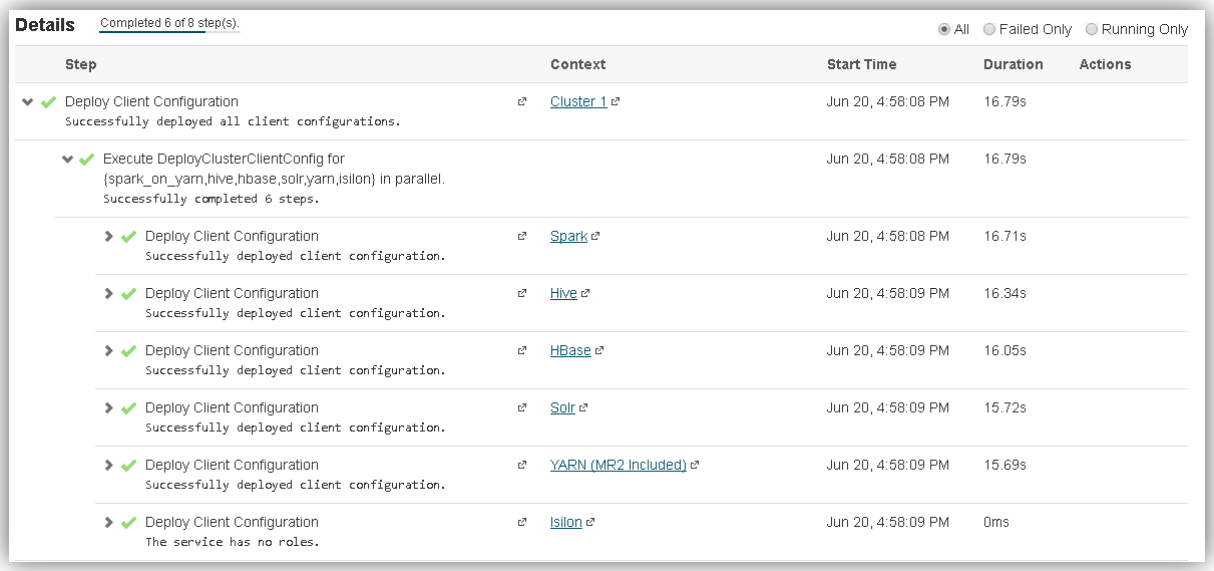


10. Continue the setup and monitor deployment.



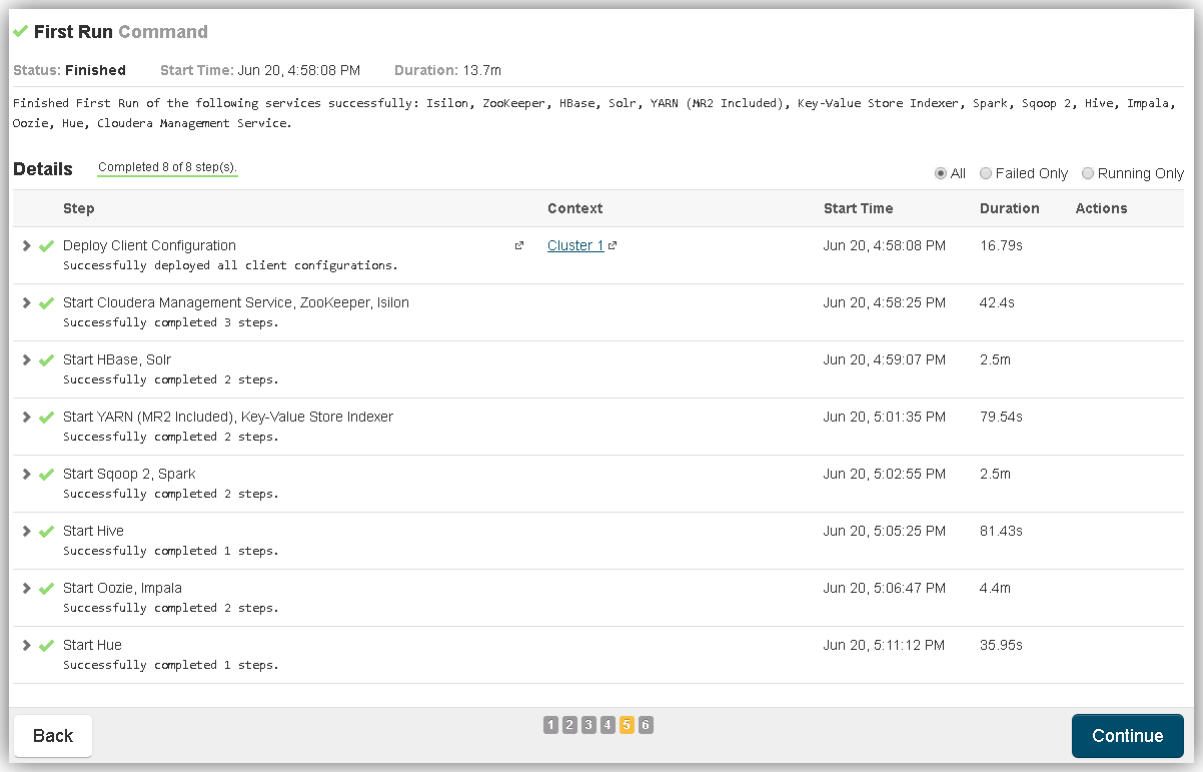
11. Review the cluster setup as it deploys.





12. The setup will complete. You can review additional details by opening specific services.

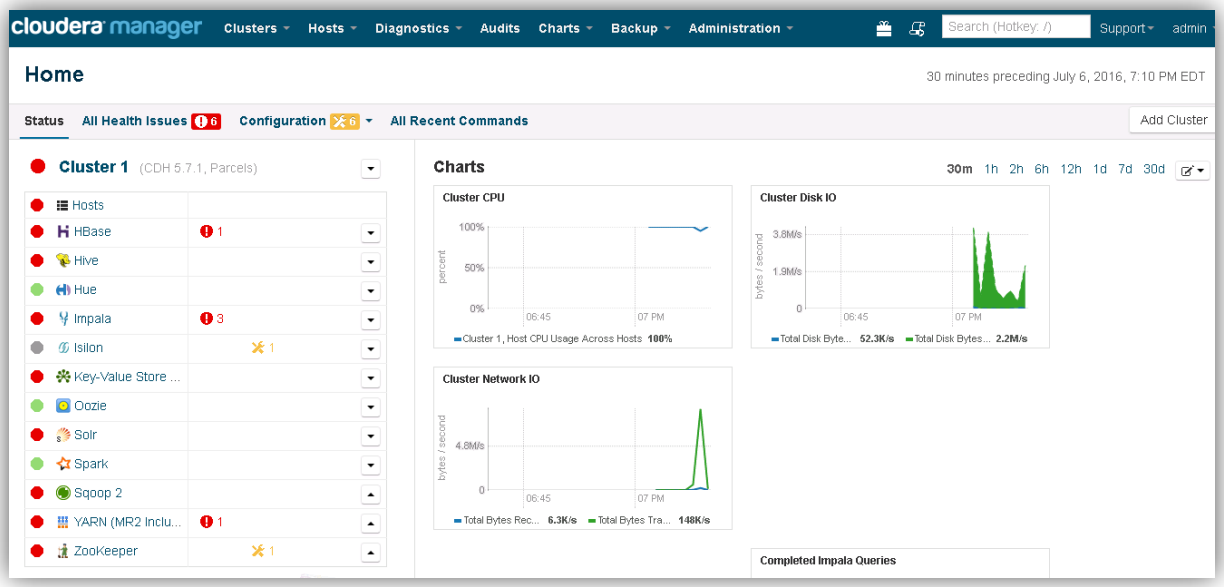
13. Click **Continue** to complete the Hadoop cluster deployment. The services are now installed, configured, and running on your cluster.



Troubleshoot Cloudera Manager

1. Return to the main Cloudera Manager dashboard to review the status. It is common to see alarms and down issues on the dashboard. Review the alarms and services and triage as needed. Some services might need restarting.

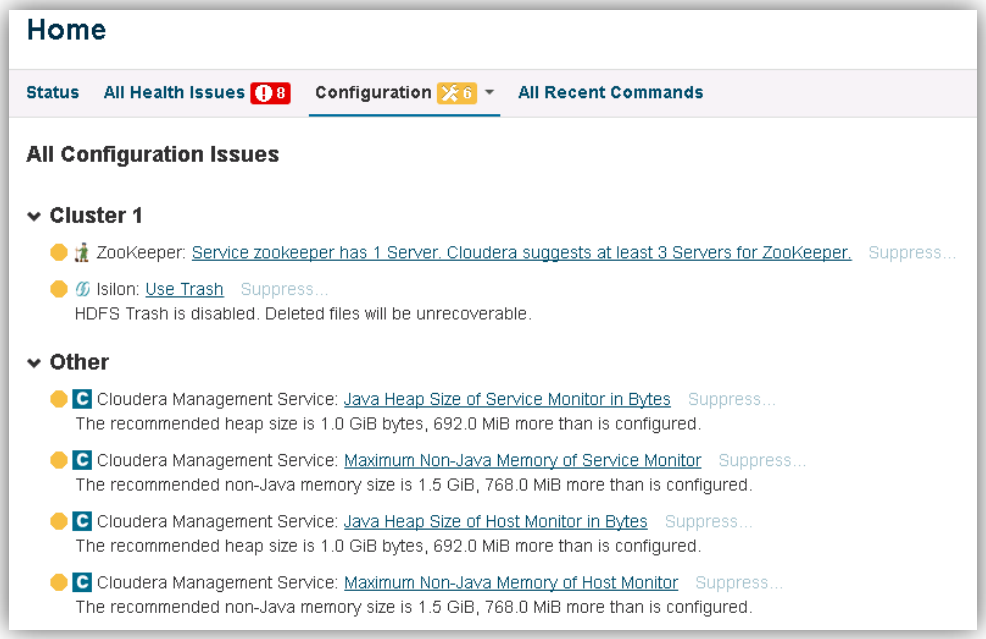
   Follow the standard protocols in starting these services, for example:
   - Start the service.
   - Monitor and review logs as needed.
   - Review the */var/log/hdfs.log* file on all nodes.
   - Restart services to resolve alarms following configuration changes.



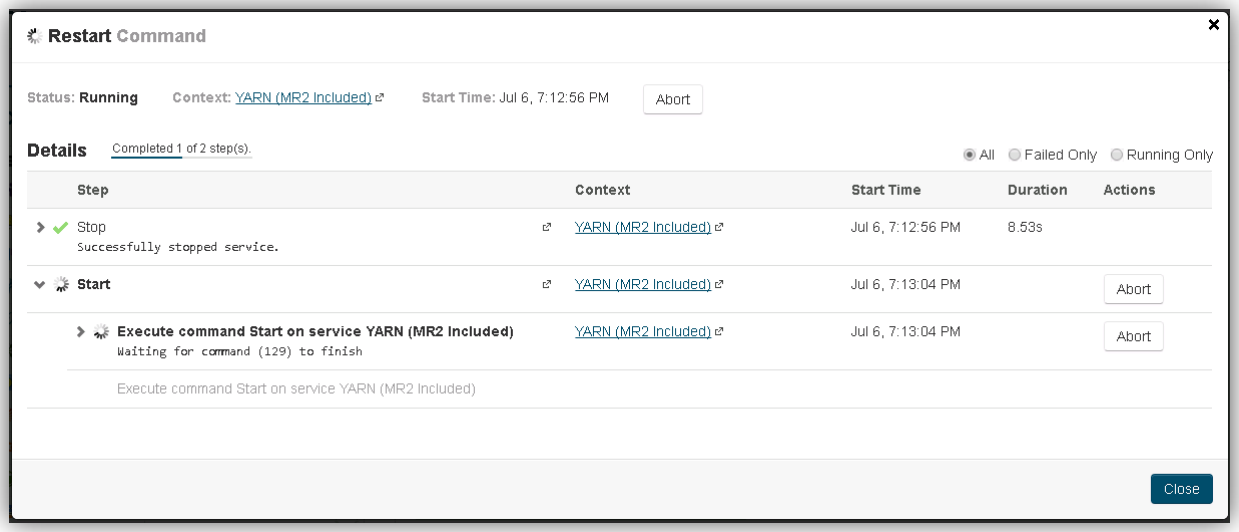
   It is common to see configuration issues. Address these issues and make the required changes as needed to resolve each issue.

Note

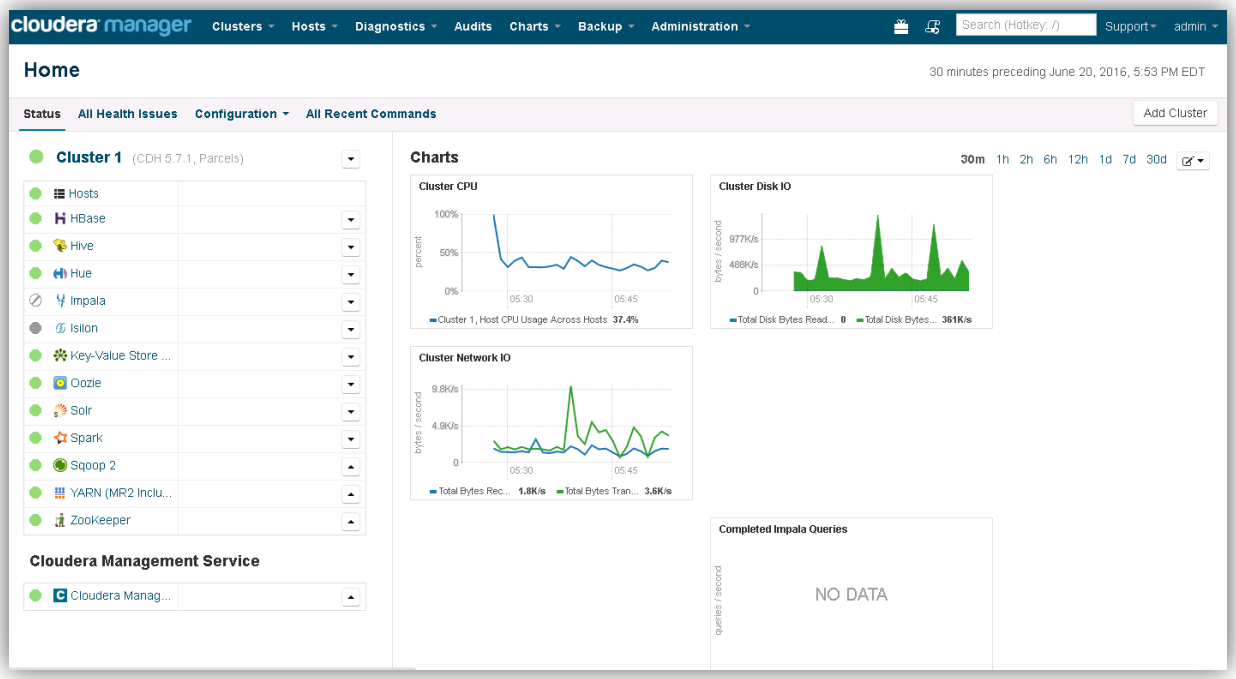
Since OneFS is the native file storage format and not HDFS, OneFS does not support HDFS Trash recovery so you can keep the HDFS Trash as disabled. OneFS does support other mechanisms to recover deleted files, as OneFS primarily supports Snapshots or SyncIQ.



Restart services to resolve alarms following configuration changes.



The Hadoop cluster and services are now fully operational and can be tested.



Verify the deployment

Test the basic functionality of the OneFS and Cloudera integration (without Kerberos) with the following steps.

1. Browse the HDFS root.

```
hadoop fs -ls /
```

Output similar to the following displays:

```
Found 5 items
-rw-r--r--   3 root  wheel                0 2017-04-21 13:55 /THIS_IS_ISILON_zone1-
cdh.txt
drwxr-xr-x   - hbase hbase                0 2017-04-21 14:04 /hbase
drwxrwxr-x   - solr  solr                 0 2017-04-21 13:56 /solr
drwxrwxrwt   - hdfs  supergroup           0 2017-04-21 14:04 /tmp
drwxr-xr-x   - hdfs  supergroup           0 2017-04-21 15:19 /user
```

2. Write to the HDFS root by creating a test directory, for example, “Made_from_Cloudera.”

Output similar to the following displays:

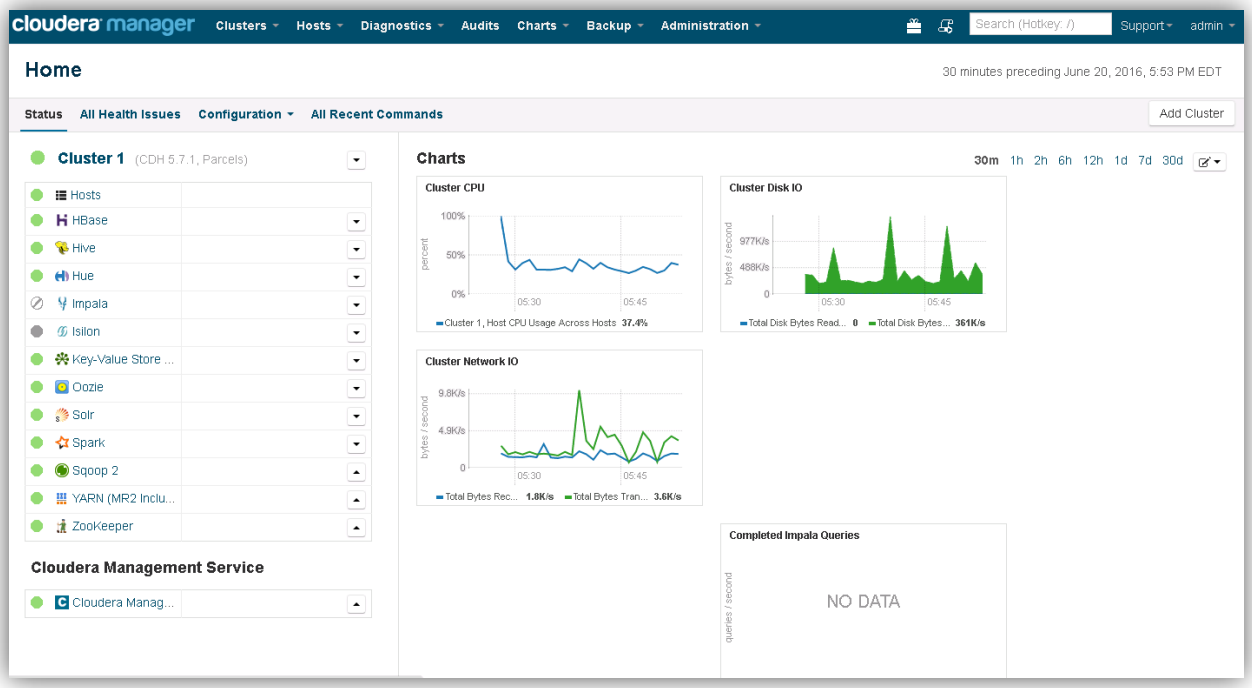
```
Found 6 items
drwxr-xr-x   - root  hadoop               0 2017-04-24 15:04 /Made_from_Cloudera
-rw-r--r--   3 root  wheel                0 2017-04-21 13:55 /THIS_IS_ISILON_zone1-
cdh.txt
drwxr-xr-x   - hbase hbase                0 2017-04-21 14:04 /hbase
drwxrwxr-x   - solr  solr                 0 2017-04-21 13:56 /solr
drwxrwxrwt   - hdfs  supergroup           0 2017-04-21 14:04 /tmp
drwxr-xr-x   - hdfs  supergroup           0 2017-04-21 15:19 /user
```

3. Run some basic smoke test jobs, for example PI or Teragen, Teravalidate, or Terasort to test MapReduce.

```
[root@rs-cd1 tmp]# yarn jar /opt/cloudera/parcels/CDH-5.7.1-1.cdh5.7.1.p0.11/lib/hadoop-mapreduce/hadoop-mapreduce-examples-2.6.0-cdh5.7.1.jar teragen 1000 /teragenOUT
16/07/07 17:25:34 INFO client.RMProxy: Connecting to ResourceManager at rs-cd1.foo.com/172.16.201.120:8032
16/07/07 17:25:35 INFO terasort.TeraSort: Generating 1000 using 2
16/07/07 17:25:35 INFO mapreduce.JobSubmitter: number of splits:2
16/07/07 17:25:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1467926691021_0001
16/07/07 17:25:37 INFO impl.YarnClientImpl: Submitted application application_1467926691021_0001
16/07/07 17:25:37 INFO mapreduce.Job: The url to track the job: http://rs-cd1.foo.com:8088/proxy/application_1467926691021_0001/
16/07/07 17:25:37 INFO mapreduce.Job: Running job: job_1467926691021_0001
16/07/07 17:26:05 INFO mapreduce.Job: Job job_1467926691021_0001 running in uber mode : false
16/07/07 17:26:05 INFO mapreduce.Job:  map 0% reduce 0%
16/07/07 17:26:36 INFO mapreduce.Job:  map 100% reduce 0%
16/07/07 17:26:37 INFO mapreduce.Job: Job job_1467926691021_0001 completed successfully
16/07/07 17:26:39 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=233210
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=164
                HDFS: Number of bytes written=100000
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=59843
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=59843
                Total vcore-seconds taken by all map tasks=59843
                Total megabyte-seconds taken by all map tasks=61279232
        Map-Reduce Framework
                Map input records=1000
                Map output records=1000
                Input split bytes=164
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=402
                CPU time spent (ms)=1480
                Physical memory (bytes) snapshot=361488384
                Virtual memory (bytes) snapshot=3003490304
                Total committed heap usage (bytes)=505413632
        org.apache.hadoop.examples.terasort.TeraGen$Counters
                CHECKSUM=2173251765740
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=100000
```

Note

With Cloudera 5.7, you may notice that Impala service is not started fully. Some additional configuration changes are needed to get this service started. The steps to install and configure Impala are in the following document: Start Cloudera 5.7 Impala with Isilon.



This completes the overview of deploying non-Kerberos Cloudera with OneFS.

Contacting Dell EMC PowerScale Technical Support

Online Support: https://support.dell.com/

Telephone Support:
United States: 800-782-4362 (800-SVC-4EMC)
Canada: 800-543-4782
Worldwide: +1-508-497-7901
Other worldwide access numbers