

Installation

Date published: 2020-11-30

Date modified: 2020-11-30

CLOUdera

Legal Notice

© Cloudera Inc. 2023. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

CDP Private Cloud Base Installation Guide.....	6
Version and Download Information.....	6
Cloudera Manager Version Information.....	6
Cloudera Manager Download Information.....	6
Cloudera Runtime Version Information.....	7
Cloudera Runtime Download Information.....	7
CDP Private Cloud Base Trial Download Information.....	8
CDP Private Cloud Base Requirements and Supported Versions.....	8
Hardware Requirements.....	8
Cloudera Manager.....	8
Cloudera Runtime.....	11
Operating System Requirements.....	18
Database Requirements.....	21
RDBMS High Availability Support.....	22
Java Requirements.....	22
Supported JDKs.....	23
Support Notes.....	23
JDK 8.....	23
Networking and Security Requirements.....	23
Data at Rest Encryption Requirements.....	25
Trial Installation.....	26
Before You Begin a Trial Installation.....	27
Installing a Trial Cluster.....	27
Step 1: Run the Cloudera Manager Server Installer.....	28
Step 2: Install Cloudera Runtime Using the Wizard.....	28
Step 3: Set Up a Cluster Using the Wizard.....	33
Stopping the Embedded PostgreSQL Database.....	35
Starting the Embedded PostgreSQL Database.....	35
Changing Embedded PostgreSQL Database Passwords.....	36
Migrating from the Cloudera Manager Embedded PostgreSQL Database Server to an External PostgreSQL Database.....	37
Prerequisites.....	37
Identify Roles that Use the Embedded Database Server.....	38
Migrate Databases from the Embedded Database Server to the External PostgreSQL Database Server.....	40
Production Installation.....	43
Before You Install.....	43
Storage Space Planning for Cloudera Manager.....	43
Configure Network Names.....	53
Disabling the Firewall.....	53
Setting SELinux Mode.....	54

Enable an NTP Service.....	55
Impala Requirements.....	56
Runtime Cluster Hosts and Role Assignments.....	58
Allocating Hosts for Key Trustee Server and Key Trustee KMS.....	63
Configuring Local Package and Parcel Repositories.....	64
Production Installation: Installing Cloudera Manager, Cloudera Runtime, and Managed Services.....	70
Step 1: Configure a Repository for Cloudera Manager.....	70
Step 2: Install Java Development Kit.....	71
Installing OpenJDK.....	72
Manually Installing OpenJDK.....	72
Manually Installing Oracle JDK.....	73
Tuning JVM Garbage Collection.....	73
Step 3: Install Cloudera Manager Server.....	76
Install Cloudera Manager Packages.....	76
Step 4: Install and Configure Databases.....	76
Required Databases.....	76
Install and Configure PostgreSQL for CDP.....	77
Install and Configure MySQL for Cloudera Software.....	82
Install and Configure MariaDB for Cloudera Software.....	87
Install and Configure Oracle Database for Cloudera Software.....	92
Configuring a database for Ranger.....	101
Configuring the Database for Streaming Components.....	105
Step 5: Set up the Cloudera Manager Database.....	107
Syntax for scm_prepare_database.sh.....	107
Step 6: Install Runtime and Other Software.....	109
Installation Wizard.....	110
Step 7: Set Up a Cluster Using the Wizard.....	114
Select Services.....	114
Assign Roles.....	115
Setup Database.....	115
Enter Required Parameters.....	116
Review Changes.....	116
Command Details.....	117
Summary.....	117
Additional Steps for Apache Ranger.....	117
Enable Plugins.....	117
Add Solr WebUI Users.....	118
Update the Time-to-live configuration for Ranger Audits.....	118
Installing Apache Knox.....	119
Apache Knox Install Role Parameters.....	120

Custom Installation Solutions.....122

Creating Virtual Images of Cluster Hosts.....	123
Creating a Pre-Deployed Cloudera Manager Host.....	123
Instantiating a Cloudera Manager Image.....	123
Creating a Pre-Deployed Worker Host.....	124
Instantiating a Worker Host.....	125
Configuring a Custom Java Home Location.....	126
Manually Install Cloudera Software Packages.....	126
Install Cloudera Manager Packages.....	126
Manually Install Cloudera Manager Agent Packages.....	127

Installation Reference.....127

Ports.....	128
------------	-----

Ports Used by Cloudera Manager.....	128
Ports Used by Cloudera Navigator Key Trustee Server.....	132
Ports Used by Cloudera Runtime Components.....	132
Ports Used by DistCp.....	138
Ports Used by Third-Party Components.....	138
Service Dependencies in Cloudera Manager.....	139
Cloudera Manager sudo command options.....	141
Introduction to Parcels.....	142
After You Install.....	142
Deploying Clients.....	143
Testing the Installation.....	143
Checking Host Heartbeats.....	143
Running a MapReduce Job.....	143
Testing with Hue.....	144
Secure Your Cluster.....	144
Troubleshooting Installation Problems.....	144
Uninstalling Cloudera Manager and Managed Software.....	147
Record User Data Paths.....	148
Stop all Services.....	148
Deactivate and Remove Parcels.....	148
Delete the Cluster.....	149
Uninstall the Cloudera Manager Server.....	149
Uninstall Cloudera Manager Agent and Managed Software.....	149
Remove Cloudera Manager, User Data, and Databases.....	150
Uninstalling a Runtime Component From a Single Host.....	151

CDP Private Cloud Base Installation Guide

Use this Installation Guide to learn how to install Cloudera software, including Cloudera Manager, Cloudera Runtime, and other managed services, in a production or trial environment.

Related Information

[Version and Download Information](#)

[CDP Private Cloud Base Requirements and Supported Versions](#)

[Trial Installation](#)

[Production Installation](#)

[Custom Installation Solutions](#)

[Installation Reference](#)

[After You Install](#)

[Troubleshooting Installation Problems](#)

[Uninstalling Cloudera Manager and Managed Software](#)

Version and Download Information

The following topics describe the available versions and download locations for Cloudera Manager and Cloudera Runtime.

Related Information

[CDP Private Cloud Base Installation Guide](#)

Cloudera Manager Version Information

Cloudera Manager is available in the following releases:

Cloudera Manager 7.2.3 is the current release of Cloudera Manager for CDP Private Cloud Base.

Release date: December 1, 2020

Previous releases:

Cloudera Manager Download Information

Important: Access to Cloudera Manager binaries for production purposes requires authentication. To access the binaries at the locations below, you must first have an active subscription agreement and obtain a license key file along with the required authentication credentials (username and password).

The license key file and authentication credentials are provided in an email sent to customer accounts from Cloudera when a new license is issued. If you have an existing license with a CDP Private Cloud Base Edition entitlement, you might not have received an email. In this instance you can identify the authentication credentials from the license key file. If you do not have access to the license key, contact your account representative to receive a copy.

To identify your authentication credentials using your license key file, complete the following steps:

- From cloudera.com, log into the cloudera.com account associated with the CDP Private Cloud Base license and subscription agreement.
- On the [CDP Private Cloud Base Download page](#), click Download Now and scroll down to the Credential Generator.
- In the Generate Credentials text box, copy and paste the text of the “PGP Signed Message” within your license key file and click Get Credentials. The credentials generator returns your username and password.



Important: Make a note of the authentication credentials. You might need them during installation to complete tasks such as configuring a remote parcel repository, or installing Cloudera Manager packages using a package manager such as YUM, APT, or other tools that you might be using in your environment.

When you obtain your authentication credentials, use them to form the URL where you can access the Cloudera Manager repository in the Cloudera Archive.

The repositories for Cloudera Manager 7.x are listed in the following tables:

Table 1: Cloudera Manager 7.2.3

Repository Type	Repository Location	Repository File
RHEL 7 Compatible (x86-compatible)	https://username:password@archive.cloudera.com/p/cm7-scale/7.2.3/redhat7/yum	https://username:password@archive.cloudera.com/p/cm7-scale/7.2.3/redhat7/yum/cloudera-manager.repo
RHEL 7 Compatible (IBM PowerPC-compatible)	https://username:password@archive.cloudera.com/p/cm7-scale/7.2.3/redhat7ppc/yum	https://username:password@archive.cloudera.com/p/cm7-scale/7.2.3/redhat7ppc/yum/cloudera-manager.repo

Cloudera Runtime Version Information

Version numbers for current and previous releases of Cloudera Runtime 7.x.

Cloudera Runtime 7.1.4 is based on Apache Hadoop 3. For more information, see *Cloudera Runtime Component Versions*.

Release date: November 30, 2020

Previous releases:

Cloudera Runtime Download Information

Important: Access to Cloudera Runtime parcels for production purposes requires authentication. To access the parcels at the locations below, you must first have an active subscription agreement and obtain a license key file along with the required authentication credentials (username and password).

The license key file and authentication credentials are provided in an email sent to customer accounts from Cloudera when a new license is issued. If you have an existing license with a CDP Private Cloud Base Edition entitlement, you might not have received an email. In this instance you can identify the authentication credentials from the license key file. If you do not have access to the license key, contact your account representative to receive a copy.

To identify your authentication credentials using your license key file, complete the following steps:

- From cloudera.com, log into the cloudera.com account associated with the CDP Private Cloud Base license and subscription agreement.
- On the [CDP Private Cloud Base Download page](#), click Download Now and scroll down to the Credential Generator.
- In the Generate Credentials text box, copy and paste the text of the “PGP Signed Message” within your license key file and click Get Credentials. The credentials generator returns your username and password.



Important: Make a note of the authentication credentials. You might need them during installation to complete tasks such as configuring a remote parcel repository.

When you obtain your authentication credentials, use them to form the URL where you can access the Runtime repository in the Cloudera Archive. Cloudera Manager can also download the Runtime parcels directly during the installation process.

The repositories for Cloudera Runtime 7.x are listed in the following tables:

Table 2: Cloudera Runtime 7.1.4.0:

Repository Type	Repository Location
Parcels	<code>https://[username]:[password]@archive.cloudera.com/p/cdh7/7.1.4.0/parcels</code>

CDP Private Cloud Base Trial Download Information

You can try the CDP Private Cloud Base Edition of Cloudera Data Platform for 60 days without obtaining a license key file.

To download CDP Private Cloud Base without obtaining a license key file, visit the [CDP Private Cloud Base Trial Download](#) page, click Try Now, and follow the download instructions. When you install CDP Private Cloud Base without a license key, you are performing a trial installation that includes an embedded PostgreSQL database and is not suitable for a production environment. For more information on trial installations, see the trial installation documentation.

A 60-day trial of CDP Private Cloud Base Edition can be enabled permanently with the appropriate license. To obtain a CDP Private Cloud Base Edition license, fill in the [Contact Us](#) form or call 866-843-7207

Related Information

[Trial Installation](#)

CDP Private Cloud Base Requirements and Supported Versions

Refer to the following topics for information about hardware, operating system, and database requirements, as well as product compatibility matrices.

Related Information

[CDP Private Cloud Base Installation Guide](#)

Hardware Requirements

As you create the architecture of your cluster, you will need to allocate Cloudera Manager and Runtime roles among the hosts in the cluster to maximize your use of resources. Cloudera provides some guidelines about how to assign roles to cluster hosts. See [Recommended Cluster Hosts and Role Distribution](#). When multiple roles are assigned to hosts, add together the total resource requirements (memory, CPUs, disk) for each role on a host to determine the required hardware.



Attention: All recommendations for the number of cores refer to logical cores, not physical cores.

For more information about sizing for a particular component, see the following minimum requirements:

Cloudera Manager

Hardware requirements for Cloudera Manager Server and related components.

Cloudera Manager Server

Table 3: Cloudera Manager Server Storage Requirements

Component	Storage	Notes
Partition hosting /usr	1 GB	
Partition hosting /var	5 GB to 1 TB	Scales according to number of nodes managed. See table below.
Partition hosting /opt	15 GB minimum	Usage grows as the number of parcels downloaded increases.
Cloudera Manager Database Server	5 GB	If the Cloudera Manager Database is shared with the Service Monitor and Host Monitor, more storage space is required to meet the requirements for those components.

Table 4: Host Based Cloudera Manager Server Requirements

Number of Cluster Hosts	Database Host Configuration	Cloudera Manager Server Heap Size	Logical Processors	Cloudera Manager Server /var Directory
Very small (#10)	Shared	4 GB	4	5 GB
Small (#20)	Shared	6 GB	6	20 GB minimum
Medium (#200)	Dedicated	10 GB	6	200 GB minimum
Large (#500)	Dedicated	12 GB	8	500 GB minimum
Extra Large (>500)	Dedicated	18 GB	16	1 TB minimum

Service Monitor Requirements

The requirements for the Service Monitor are based on the number of monitored entities. To see the number of monitored entities, perform the following steps:

1. Open the Cloudera Manager Admin Console and click **Clusters Cloudera Management Service**.
2. Find the Cloudera Management Service Monitored Entities chart. If the chart does not exist, add it from the Chart Library.

For more information about Cloudera Manager entities, see *Cloudera Manager Entity Types*.

Table 5: Clusters with HDFS, YARN, or Impala

Use the recommendations in this table for clusters where the only services with worker roles are HDFS, YARN, or Impala.

Number of Monitored Entities	Number of Hosts	Required Java Heap Size	Recommended Non-Java Heap Size
0-2,000	0-100	1 GB	6 GB
2,000-4,000	100-200	1.5 GB	6 GB
4,000-8,000	200-400	1.5 GB	12 GB
8,000-16,000	400-800	2.5 GB	12 GB
16,000-20,000	800-1,000	3.5 GB	12 GB

Table 6: Clusters with HBase, Solr, Kafka, or Kudu

Use these recommendations when services such as HBase, Solr, Kafka, or Kudu are deployed in the cluster. These services typically have larger quantities of monitored entities.

Number of Monitored Entities	Number of Hosts	Required Java Heap Size	Recommended Non-Java Heap Size
0-30,000	0-100	2 GB	12 GB
30,000-60,000	100-200	3 GB	12 GB
60,000-120,000	200-400	3.5 GB	12 GB
120,000-240,000	400-800	8 GB	20 GB

Related Information

Host Monitor and Service Monitor Memory Configuration

Host Monitor

The requirements for the Host Monitor are based on the number of monitored entities.

To see the number of monitored entities, perform the following steps:

1. Open the Cloudera Manager Admin Console and click **Clusters Cloudera Management Service**.
2. Find the Cloudera Management Service Monitored Entities chart. If the chart does not exist, add it from the Chart Library.

For more information about Cloudera Manager entities, see *Cloudera Manager Entity Types*.

Number of Hosts	Number of Monitored Entities	Heap Size	Non-Java Heap Size
0-200	<6k	1 GB	2 GB
200-800	6k-24k	2 GB	6 GB
800-1000	24k-30k	3 GB	6 GB

Ensure that you have at least 25 GB of disk space available for the Host Monitor, Service Monitor, Reports Manager, and Events Server databases.

Related Information

Host Monitor and Service Monitor Memory Configuration

Reports Manager

The Reports Manager fetches the fsimage from the NameNode at regular intervals. It reads the fsimage and creates a Lucene index for it. To improve the indexing performance, Cloudera recommends provisioning a host as powerful as possible and dedicating an SSD disk to the Reports Manager.

Table 7: Reports Manager

Component	Java Heap	CPU	Disk
Reports Manager	3-4 times the size of the fsimage.	<ul style="list-style-type: none"> Minimum: 8 cores Recommended: 16 cores (32 cores, with hyperthreading enabled.) 	1 dedicated disk that is at least 20 times the size of the fsimage. Cloudera strongly recommends using SSD disks.

Agent Hosts

An unpacked parcel requires approximately three times the space of the packed parcel that is stored on the Cloudera Manager Server.

Component	Storage	Notes
Partition hosting /opt	15 GB minimum	Usage grows as new parcels are downloaded to cluster hosts.
/var/log	2 GB per role	Each role running on the host will need at least 2 GB of disk space.

Event Server

The following table lists the minimum requirements for the Event Server:

CPU	RAM	Storage
1 core	256 MB	<ul style="list-style-type: none"> 5 GB for the Event Database 20 GB for the Event Server Index Directory. The location of this directory is set by the Event Server Index Directory Event Server configuration property.

Alert Publisher

The following table lists the minimum requirements for the Alert Publisher:

CPU	RAM	Storage
1 core	1 GB	Minimum of 1 disk for log files

Cloudera Runtime

Hardware requirements for Cloudera Runtime components.

Atlas

Memory	CPU	Disk	Additional Dependencies
Small: 4 GB Large: 32 GB	Minimum: 4 Medium: 8 Large: 16	No special requirement because HBase is used for storage.	Solr Shards: 4 (property: atlas_solr_shards) The shards for Atlas collections within Solr is determined by this number.

Data Analytics Studio (DAS)

DAS is a memory-heavy and a disk-light application. For optimum performance, consider profiling the CPU cores, memory allocation, and disk space depending upon the number of users, the total number of databases and tables, and the number of queries in the system.

If you are setting up a high-availability cluster, then add additional cores and memory for the load balancer.

The following table provides component-wise recommendation for provisioning CPU, memory, and disk space. These recommendations are approximated considering 10 users, 10,000 Hive tables, 100 parallel Event Processor threads, and 40,000 queries.

Table 8: Hardware requirements for DAS

DAS component	CPU	Memory	Local Disk
Webapp	<ul style="list-style-type: none"> Minimum: 2 cores Recommended: 2 cores *The number of cores that you allocate need to be proportional to U.	<ul style="list-style-type: none"> Minimum: 4 GB Recommended: 8 GB *The amount of memory that you allocate need to be proportional to U and T.	<ul style="list-style-type: none"> Minimum: 5 GB Recommended: 10 GB *The amount of disk space that you allocate need to be proportional to U.
Event Processor	<ul style="list-style-type: none"> Minimum: 2 cores Recommended: 4 cores *The number of cores that you allocate need to be proportional to P.	<ul style="list-style-type: none"> Minimum: 4 GB Recommended: 8 GB *The amount of memory that you allocate need to be proportional to P and T.	<ul style="list-style-type: none"> Minimum: 5 GB Recommended: 5 GB *The disk space is primarily used for logs, and can remain constant.

DAS component	CPU	Memory	Local Disk
Database	<ul style="list-style-type: none"> Minimum: 2 cores Recommended: 4 cores <p>*The number of cores that you allocate need to be proportional to (P + U).</p>	<ul style="list-style-type: none"> Minimum: 4 GB Recommended: 8 GB <p>*The amount of memory that you allocate need to be proportional to (T + Q).</p>	<ul style="list-style-type: none"> Minimum: 5 GB Recommended: 20 GB <p>*The amount of disk space that you allocate need to be proportional to (T + U + Q).</p>

Where,

U is the number of users concurrently accessing the DAS Webapp

T is the number of tables in Hive

P denotes the parallelism configured in the DAS Event Processor

Q is the total number of queries in the system

Table 9: DAS Port Specifications

Default Port Number	Description
30900	Event Processor server port
30901	Event Processor admin server port
30800	Webapp server port
30801	Webapp admin port

HDFS

Component	Memory	CPU	Disk
JournalNode	1 GB (default) Set this value using the Java Heap Size of JournalNode in Bytes HDFS configuration property.	1 core minimum	1 dedicated disk
NameNode	<ul style="list-style-type: none"> Minimum: 1 GB (for proof-of-concept deployments) Add an additional 1 GB for each additional 1,000,000 blocks <p>Snapshots and encryption can increase the required heap memory.</p> <p>See <i>Sizing NameNode Heap Memory</i>.</p> <p>Set this value using the Java Heap Size of NameNode in Bytes HDFS configuration property.</p>	Minimum of 4 dedicated cores; more may be required for larger clusters	<ul style="list-style-type: none"> Minimum of 2 dedicated disks for metadata 1 dedicated disk for log files (This disk may be shared with the operating system.) Maximum disks: 4

Component	Memory	CPU	Disk
DataNode	<p>Minimum: 4 GB Maximum: 8 GB</p> <p>Increase the memory for higher replica counts or a higher number of blocks per DataNode. When increasing the memory, Cloudera recommends an additional 1 GB of memory for every 1 million replicas above 4 million on the DataNodes. For example, 5 million replicas require 5 GB of memory.</p> <p>Set this value using the Java Heap Size of DataNode in Bytes HDFS configuration property.</p>	Minimum: 4 cores. Add more cores for highly active clusters.	<p>Minimum: 4 Maximum: 24</p> <p>The maximum acceptable size will vary depending upon how large average block size is. The DN's scalability limits are mostly a function of the number of replicas per DN, not the overall number of bytes stored. That said, having ultra-dense DNs will affect recovery times in the event of machine or rack failure. Cloudera does not support exceeding 100 TB per data node. You could use 12 x 8 TB spindles or 24 x 4TB spindles. Cloudera does not support drives larger than 8 TB.</p>



Warning: Running Runtime on storage platforms other than direct-attached physical disks can provide suboptimal performance. Cloudera Enterprise and the majority of the Hadoop platform are optimized to provide high performance by distributing work across a cluster that can utilize data locality and fast local I/O.

HBase

Component	Java Heap	CPU	Disk
Master	<ul style="list-style-type: none"> 100-10,000 regions: 4 GB 10,000 or more regions with 200 or more Region Servers: 8 GB 10,000 or more regions with 300 or more Region Servers: 12 GB <p>Set this value using the Java Heap Size of HBase Master in Bytes HBase configuration property.</p>	Minimum 4 dedicated cores. You can add more cores for larger clusters, when using replication, or for bulk loads.	1 disk for local logs, which can be shared with the operating system and/or other Hadoop logs
Region Server	<ul style="list-style-type: none"> Minimum: 8 GB Medium-scale production: 16 GB Heap larger than 16 GB requires special Garbage Collection tuning. See <i>Configuring the HBase BlockCache</i>. <p>Set this value using the Java Heap Size of HBase RegionServer in Bytes HBase configuration property.</p>	Minimum: 4 dedicated cores	<ul style="list-style-type: none"> 4 or more spindles for each HDFS DataNode 1 disk for local logs (this disk can be shared with the operating system and/or other Hadoop logs)
Thrift Server	<p>1 GB - 4 GB</p> <p>Set this value using the Java Heap Size of HBase Thrift Server in Bytes HBase configuration property.</p>	Minimum 2 dedicated cores.	1 disk for local logs, which can be shared with the operating system and other Hadoop logs.



Note: Consider adding more HBase Thrift Servers for production environments and deployments with a large number of Thrift client to scale horizontally.

Related Information

[Configuring HBase BlockCache](#)

Hive

Component	Java Heap		CPU	Disk
HiveServer 2	Single Connection	4 GB	Minimum 4 dedicated cores	Minimum 1 disk
	2-10 connections	4-6 GB		This disk is required for the following: <ul style="list-style-type: none">• HiveServer2 log files• stdout and stderr output files• Configuration files• Operation logs stored in the operation_logs_dir directory, which is configurable• Any temporary files that might be created by local map tasks under the /tmp directory
	11-20 connections	6-12 GB		
	21-40 connections	12-16 GB		
	41 to 80 connections	16-24 GB		
	Cloudera recommends splitting HiveServer2 into multiple instances and load balancing them once you start allocating more than 16 GB to HiveServer2. The objective is to adjust the size to reduce the impact of Java garbage collection on active processing by the service.			
Set this value using the Java Heap Size of HiveServer2 in Bytes Hive configuration property. For more information, see Apache Hive Performance Tuning .				
Hive Metastore	Single Connection	4 GB	Minimum 4 dedicated cores	Minimum 1 disk
	2-10 connections	4-10 GB		This disk is required so that the Hive metastore can store the following artifacts: <ul style="list-style-type: none">• Logs• Configuration files• Backend database that is used to store metadata if the database server is also hosted on the same node
	11-20 connections	10-12 GB		
	21-40 connections	12-16 GB		
	41 to 80 connections	16-24 GB		
	Set this value using the Java Heap Size of Hive Metastore Server in Bytes Hive configuration property. For more information, see Apache Hive Performance Tuning .			
Beeline CLI	Minimum: 2 GB		N/A	N/A

Hue

Component	Memory	CPU	Disk
Hue Server	<ul style="list-style-type: none"> • Minimum: 4 GB • Maximum 10 GB • If the cluster uses the Hue load balancer, add additional memory 	Minimum: 1 Core to run Django When Hue is configured for high availability, add additional cores	Minimum: 10 GB for the database, which grows proportionally according to the cluster size and workloads. When Hue is configured for high availability, add space is required for the /tmp (temporary) directory, approximately 5GB.

The term "cluster size" refers to the number of nodes in the cluster. "Workload" in Hue means the number of queries run and the number of concurrent unique users using the application in a given period of time.

A minimum of 10GB is needed for the database. The Hive MetaStore service largely uses the database. The database grows in size quickly because of the query history that it retains. To optimize performance, you must regularly cleanup old documents and queries.



Note: Hue is limited by cgroup settings. In Cloudera Manager, all memory soft/hard limits are set to -1.

Related Information

[Adding a Load Balancer for Hue](#)

Kafka

Kafka requires a fairly small amount of resources, especially with some configuration tuning. By default, Kafka, can run on as little as 1 core and 1GB memory with storage scaled based on requirements for data retention.



CPU is rarely a bottleneck because Kafka is I/O heavy, but a moderately-sized CPU with enough threads is still important to handle concurrent connections and background tasks.


Kafka brokers tend to have a similar hardware profile to HDFS data nodes. How you build them depends on what is important for your Kafka use cases.

Use the following guidelines:

To affect performance of these features:	Adjust these parameters:
Message Retention	Disk size
Client Throughput (Producer & Consumer)	Network capacity
Producer throughput	Disk I/O
Consumer throughput	Memory

A common choice for a Kafka node is as follows:

Component	Memory/Java Heap	CPU	Disk
Broker	<ul style="list-style-type: none"> RAM: 64 GB Recommended Java heap: 4 GB <p>Set this value using the Java Heap Size of Broker Kafka configuration property.</p>	12- 24 cores	<ul style="list-style-type: none"> 1 HDD For operating system 1 HDD for Zookeeper dataLogDir 10- HDDs, using Raid 10, for Kafka data
Cruise Control	1 GB	1 core  Note: A moderately-sized CPU with enough threads is important to handle metric fetching from Kafka and background tasks.	Because Cruise Control stores its data in Kafka the storage requirements will depend on the retention settings of the related Kafka topics.
Kafka Connect	0.5 - 4 GB heap size depending on the Connectors in use.	4 cores  Note: Depends on the Connectors in use.	
MirrorMaker	1 GB heap <p>Set this value using the Java Heap Size of MirrorMaker Kafka configuration property.</p>	1 core per 3-4 streams	No disk space needed on MirrorMaker instance. Destination brokers should have sufficient disk space to store the topics being copied over.
Schema Registry	1 GB heap	2 cores	1 MB <p>Serialization JAR files may be uploaded and may be of any size. The disk usage depends on the JAR files uploaded. The files may be stored locally on the same host where SchemaRegistry is running or in HDFS if available.</p>

Component	Memory/Java Heap	CPU	Disk
Streams Messaging Manager  Note: The hardware requirements for SMM depends on the number of Kafka partitions.	8 GB heap	8 cores	5 GB
Streams Replication Manager	<ul style="list-style-type: none"> 1 GB heap for SRM driver 1 GB heap for SRM Service 	The performance of the SRM driver is mostly impacted by network throughput and latency.	No resources required

Networking requirements: Gigabit Ethernet or 10 Gigabit Ethernet. Avoid clusters that span multiple data centers.

Kafka and Zookeeper: It is common to run ZooKeeper on 3 broker nodes that are dedicated for Kafka. However, for optimal performance Cloudera recommends the usage of dedicated Zookeeper hosts. This is especially true for larger, production environments.

Oozie

Component	Java Heap	CPU	Disk
Oozie	<ul style="list-style-type: none"> Minimum: 1 GB (this is the default set by Cloudera Manager). This is sufficient for less than 10 simultaneous workflows, without forking. If you notice excessive garbage collection, or out-of-memory errors, increase the heap size to 4 GB for medium-size production clusters or to 8 GB for large-size production clusters. Set this value using the Java Heap Size of Oozie Server in Bytes Oozie configuration property. 	No resources required	No resources required

Additional tuning:

For workloads with many coordinators that run with complex workflows (a max concurrency reached! warning appears in the log and the Oozie admin -queuedump command shows a large queue):

- Increase the value of the `oozie.service.CallableQueueService.callable.concurrency` property to 50.
- Increase the value of the `oozie.service.CallableQueueService.threads` property to 200.

Do not use a Derby database as a backend database for Oozie.

Ranger

Memory	CPU	Disk	Additional Dependencies
Ranger Admin: 1 GB minimum, then adjust heap as required (8 GB-16 GB)	1 core minimum	No special requirement.	
Ranger Usersync: 1 GB minimum	1 core minimum	No special requirement.	
Ranger Tagsync: 1 GB minimum	1 core minimum	No special requirement.	


Search

Component	Java Heap	CPU	Disk
Solr	<ul style="list-style-type: none"> Small workloads, or evaluations: 16 GB Smaller production environments: 32 GB Larger production environments: 96 GB is sufficient for most clusters. <p>Set this value using the Java Heap Size of Solr Server in Bytes Solr configuration property.</p> <p>See</p>	<ul style="list-style-type: none"> Minimum: 4 Recommended: 16 for production workloads 	No requirement. Solr uses HDFS for storage.

Note the following considerations for determining the optimal amount of heap memory:

- Size of searchable material: The more searchable material you have, the more memory you need. All things being equal, 10 TB of searchable data requires more memory than 1 TB of searchable data.
- Content indexed in the searchable material: Indexing all fields in a collection of logs, email messages, or Wikipedia entries requires more memory than indexing only the Date Created field.
- The level of performance required: If the system must be stable and respond quickly, more memory may help. If slow responses are acceptable, you may be able to use less memory.

Spark

Component	Java Heap	CPU	Disk
Spark History Server	<p>Minimum: 512 MB</p> <p>Set this value using the Java Heap Size of History Server in Bytes Spark configuration property.</p>	<p>1</p>  <p>Important: Cloudera recommends that you adjust the number of CPUs and memory for the Spark History Server based on your specific cluster usage patterns.</p>	Minimum 1 disk for log files.

YARN

Component	Java Heap	CPU	Other Recommendations
Job History Server	<ul style="list-style-type: none"> Minimum: 1 GB Increase memory by 1.6 GB for each 100,000 tasks kept in memory. For example: 5 jobs @ 100,000 mappers + 20,000 reducers = 600,000 total tasks requiring 9.6 GB of heap. <p>See the Other Recommendations column for additional tuning suggestions.</p> <p>Set this value using the Java Heap Size of JobHistory Server in Bytes YARN configuration property.</p>	Minimum: 1 core	<ul style="list-style-type: none"> Set the <code>mapreduce.jobhistory.loadedtasks.cache.size</code> property to a total loaded task count. Using the example in the Java Heap column to the left, of 650,000 total tasks, you can set it to 700,000 to allow for some safety margin. This should also prevent the JobHistoryServer from hanging during garbage collection, since the job count limit does not have a task limit.

Component	Java Heap	CPU	Other Recommendations
NodeManager	<p>Minimum: 1 GB.</p> <p>Configure additional heap memory for the following conditions:</p> <ul style="list-style-type: none"> Large number of containers Large shuffle sizes in Spark or MapReduce <p>Set this value using the Java Heap Size of NodeManager in Bytes YARN configuration property.</p>	<ul style="list-style-type: none"> Minimum: 8-16 cores Recommended: 32-64 cores 	<p>Disks:</p> <ul style="list-style-type: none"> Minimum: 8 disks Recommended: 12 or more disks <p>Networking:</p> <ul style="list-style-type: none"> Minimum: Dual 1Gbps or faster Recommended: Single/Dual 10 Gbps or faster
ResourceManager	<p>Minimum: 6 GB</p> <p>Configure additional heap memory for the following conditions:</p> <ul style="list-style-type: none"> More jobs Larger cluster size Number of retained finished applications (configured with the yarn.resourcemanager .max-completed-applications property. Scheduler configuration <p>Set this value using the Java Heap Size of ResourceManager in Bytes YARN configuration property.</p>	Minimum: 1 core	
Other Settings	<ul style="list-style-type: none"> Set the ApplicationMaster Memory YARN configuration property to 512 MB Set the Container Memory Minimum YARN configuration property to 1 GB. 	N/A	N/A

Related Information

[Tuning Apache Hadoop YARN](#)

ZooKeeper

Component	Java Heap	CPU	Disk
ZooKeeper Server	<ul style="list-style-type: none"> Minimum: 1 GB Increase heap size when watching 10,000 - 100,000 ephemeral znodes and are using 1,000 or more clients. <p>Set this value using the Java Heap Size of ZooKeeper Server in Bytes ZooKeeper configuration property.</p>	Minimum: 4 cores	<p>ZooKeeper was not designed to be a low-latency service and does not benefit from the use of SSD drives. The ZooKeeper access patterns – append-only writes and sequential reads – were designed with spinning disks in mind. Therefore Cloudera recommends using HDD drives.</p>

Related Information

[Add a ZooKeeper service](#)

Operating System Requirements

This topic describes the operating system requirements for Cloudera software.

CDP Private Cloud Base Supported Operating Systems

Table 10: Supported operating systems for CDP Private Cloud Base 7.1.4

Operating System	Version (bold=new)
IBM Spectrum Scale on RHEL/CentOS/Oracle	7.6, 7.7, 7.8

Runtime and Cloudera Manager Supported Operating Systems

Runtime provides parcels for select versions of RHEL-compatible operating systems.



Important:

In order to be covered by Cloudera Support:

- All Runtime hosts in a logical cluster must run on the same major OS release.
- Cloudera supports a temporarily mixed OS configuration during an OS upgrade project.
- Cloudera Manager must run on the same OS release as one of the clusters it manages.

Cloudera recommends running the same minor release on all cluster nodes. However, the risk caused by running different minor OS releases is considered lower than the risk of running different major OS releases.

Points to note:

- Cloudera does not support Runtime cluster deployments in Docker containers.
- Cloudera Enterprise is supported on platforms with Security-Enhanced Linux (SELinux) enabled and in enforcing mode. Cloudera is not responsible for policy support or policy enforcement. If you experience issues with SELinux, contact your OS provider.

Software Dependencies

- Python - CDP Private Cloud Base, with the exception of Hue, is supported on the Python version that is included in the operating system by default, as well as higher versions, but is not compatible with Python 3.0 or higher.

For example, CDP Private Cloud Base requires Python 2.7 or higher on RHEL 7 compatible operating systems.

Spark 2 requires Python 2.7 or higher. If the right level of Python is not picked up by default, set the `PYSPARK_PYTHON` and `PYSPARK_DRIVER_PYTHON` environment variables to point to the correct Python executable before running the `pyspark` command.

Python 3 is not supported.

- Perl - Cloudera Manager requires perl.
- python-psycpg2 - Cloudera Manager 7 has a dependency on the package python-psycpg2. Hue in Runtime 7 requires a higher version of psycpg2 than is required by the Cloudera Manager dependency. For more information, see *Installing the psycpg2 Python Package*.
- iproute package - CDP Private Cloud Base has a dependency on the iproute package. Any host that runs the Cloudera Manager Agent requires the package. The required version varies depending on the operating system:

Table 11: iproute package

Operating System	iproute version
RHEL 7 Compatible	iproute-3.10

Filesystem Requirements

Supported Filesystems

The Hadoop Distributed File System (HDFS) is designed to run on top of an underlying filesystem in an operating system. Cloudera recommends that you use either of the following filesystems tested on the supported operating systems:

- ext3: This is the most tested underlying filesystem for HDFS.
- ext4: This scalable extension of ext3 is supported in more recent Linux releases.



Important: Cloudera does not support in-place upgrades from ext3 to ext4. Cloudera recommends that you format disks as ext4 before using them as data directories.

- XFS: This is the default filesystem in RHEL 7.
- S3: Amazon Simple Storage Service

Kudu Filesystem Requirements - Kudu is supported on ext4 and XFS. Kudu requires a kernel version and filesystem that supports hole punching. Hole punching is the use of the `fallocate(2)` system call with the `FALLOC_FL_PUNCH_HOLE` option set.

File Access Time

Linux filesystems keep metadata that record when each file was accessed. This means that even reads result in a write to the disk. To speed up file reads, Cloudera recommends that you disable this option, called `atime`, using the `noatime` mount option in `/etc/fstab`:

```
/dev/sdb1 /data1 ext4 defaults,noatime 0
```

Apply the change without rebooting:

```
mount -o remount /data1
```

Filesystem Mount Options

The filesystem mount options have a `sync` option that allows you to write synchronously.

Using the `sync` filesystem mount option reduces performance for services that write data to disks, such as HDFS, YARN, Kafka and Kudu. In CDH, most writes are already replicated. Therefore, synchronous writes to disk are unnecessary, expensive, and do not measurably improve stability.

NFS and NAS options are not supported for use as DataNode Data Directory mounts, even when using Hierarchical Storage features.

Mounting `/tmp` as a filesystem with the `noexec` option is sometimes done as an enhanced security measure to prevent the execution of files stored there. However, this causes multiple problems with various parts of Cloudera Manager and CDH. Therefore, Cloudera does not support mounting `/tmp` with the `noexec` option.

Filesystem Requirements

Cloudera Manager automatically sets `nproc` configuration in `/etc/security/limits.conf`, but this configuration can be overridden by individual files in `/etc/security/limits.d/`. This can cause problems with Apache Impala and other components.

Make sure that the `nproc` limits are set sufficiently high, such as 65536 or 262144.

nscd for Kudu

Although not a strict requirement, it's highly recommended that you use `nscd` to cache both DNS name resolution and static name resolution for Kudu.

Database Requirements

Table 12: Database Support for CDP Private Cloud Base 7.1.5

Database Type	Supported Version
MySQL	5.7 Not supported for DAS.
MariaDB	10.2 Not supported for DAS.
PostgreSQL	10, 11.x For use with DAS only: 9.6 - 12
Oracle DB	12.2.0.1 (Oracle 12c Release 2) Not supported for DAS. 19.3.0.0 (Oracle 19c)



Important: When you restart processes, the configuration for each of the services is redeployed using information saved in the Cloudera Manager database. If this information is not available, your cluster cannot start or function correctly. You must schedule and maintain regular backups of the Cloudera Manager database to recover the cluster in the event of the loss of this database. For more information, see *Backing Up Databases*.

Cloudera Manager and Runtime come packaged with an embedded PostgreSQL database for use in non-production environments. The embedded PostgreSQL database is not supported in production environments. For production environments, you must configure your cluster to use dedicated external databases.

After installing a database, upgrade to the latest patch and apply appropriate updates. Available updates may be specific to the operating system on which it is installed.

Notes:

- Cloudera recommends that for most purposes you use the default versions of databases that correspond to the operating system of your cluster nodes. Refer to the operating system's documentation to verify support if you choose to use a database other than the default. Note that Hue requires the default MySQL/MariaDB version (if used) of the operating system on which it is installed.
- Data Analytics Studio requires PostgreSQL version 9.6, while RHEL 7.6 provides PostgreSQL 9.2.
- Cloudera does not support using Derby database with Oozie. You can use it for testing or debugging purposes, but Cloudera does not recommend using it in production environments. This could cause failures while upgrading from CDH to CDP.
- Use UTF8 encoding for all custom databases.

Oozie also supports UTF8MB4 character encoding out of box without any configuration change when the Oozie custom database is created with the encoding of UTF8MB4.

MySQL and MariaDB must use the MySQL utf8 encoding, not utf8mb4.

- For MySQL 5.7, you must install the MySQL-shared-compat or MySQL-shared package. This is required for the Cloudera Manager Agent installation.
- MySQL GTID-based replication is not supported.
- Both the Community and Enterprise versions of MySQL are supported, as well as MySQL configured by the AWS RDS service.

- Before upgrading from CDH 5 to CDH 6, check the value of the COMPATIBLE initialization parameter in the Oracle Database using the following SQL query:

```
SELECT name, value FROM v$parameter WHERE name = 'compatible'
```

The default value is 12.2.0. If the parameter has a different value, you can set it to the default as shown in the [Oracle Database Upgrade Guide](#).



Note: Before resetting the COMPATIBLE initialization parameter to its default value, make sure you consider the effects of this change can have on your system.

Related Information

[Required Databases](#)

RDBMS High Availability Support

Various Cloudera components rely on backing RDBMS services as critical infrastructure. You may require Cloudera components to support deployment in environments where RDBMS services are made highly-available. High availability (HA) solutions for RDBMS are implementation-specific, and can create constraints or behavioral changes in Cloudera components.

This section clarifies the support state and identifies known issues and limitations for HA deployments.

High Availability vs. Load Balancing

Understanding the difference between HA and load balancing is important for Cloudera components, which are designed to assume services are provided by a single RDBMS instance. Load balancing distributes operations across multiple RDBMS services in parallel, while HA focuses on service continuity. Load balanced deployments are often used as part of HA strategies to overcome demands of monitoring and failover management in an HA environment. While less easier to implement, load-balanced deployments require applications tailored to the behavior and limitations of the particular technology.

Support Statement: Cloudera components are not designed for and do not support load balanced deployments of any kind. Any HA strategy involving multiple active RDBMS services must ensure all connections are routed to a single RDBMS service at any given time, regardless of vendor or HA implementation/technology.

General High Availability Support

Cloudera supports various RDBMS options, each of which have multiple possible strategies to implement HA. Cloudera cannot reasonably test and certify on each strategy for each RDBMS. Cloudera expects HA solutions for RDBMS to be transparent to Cloudera software, and therefore are not supported and debugged by Cloudera. It is the responsibility of the customer to provision, configure, and manage the RDBMS HA deployment, so that Cloudera software behaves as it would when interfacing with a single, non-HA service. Cloudera will support and help customers troubleshoot issues when a cluster has HA enabled. While diagnosing database-related problems in Cloudera components, customers may be required to temporarily disable or bypass HA mechanisms for troubleshooting purposes. If an HA-related issue is found, it is the responsibility of the customer to engage with the database vendor so that a solution to that issue can be found.

Support Statement: Cloudera Support may require customers to temporarily bypass HA layers and connect directly to supported RDBMS back-ends to troubleshoot issues. Issues observed only when connected through HA layers are the responsibility of the customer DBA staff to resolve.

Java Requirements

Supported JDKs for CDP Private Cloud Base

Related Information

[Step 2: Install Java Development Kit](#)

Supported JDKs

CDP Private Cloud Base Version	Supported OpenJDK	Supported Oracle JDK
7.1	<ul style="list-style-type: none"> OpenJDK 1.8 	Oracle JDK 1.8

Support Notes



Note: A Java optimization called compressed oops (ordinary object pointers) enables a 64-bit JVM to address heap sizes up to about 32 GB using 4-byte pointers. For larger heap sizes, 8-byte pointers are required. This means that a heap size slightly less than 32 GB can hold more objects than a heap size slightly more than 32 GB.

If you do not need more than 32 GB heap, set your heap size to 31GB or less to avoid this issue. If you need 32 GB or more, set your heap size to 48 GB or higher to account for the larger pointers. In general, for heap sizes above 32 GB, multiply the amount of heap you need by 1.5.

Only 64 bit JDKs are supported.

Unless specifically excluded, Cloudera supports later updates to a major JDK release from the release that support was introduced. Cloudera excludes or removes support for select Java updates when security is jeopardized.

Running Runtime nodes within the same cluster on different JDK releases is not supported. All cluster hosts must use the same JDK update level.

JDK 8

Table 13: Oracle JDK 8 versions that are tested and recommended

Oracle JDK 8 Version	Notes
1.8u181	Recommended

Table 14: OpenJDK 8 versions that are tested and recommended

OpenJDK Version	Notes
1.8u232	Minimum required / Latest version tested
11.0.4+11	

Networking and Security Requirements

Cloudera Runtime and Cloudera Manager Supported Transport Layer Security Versions

The following components are supported by the indicated versions of Transport Layer Security (TLS):

Cloudera Runtime and Cloudera Manager Networking and Security Requirements

The hosts in a Cloudera Manager deployment must satisfy the following networking and security requirements:

- Networking Protocols Support

CDH requires IPv4. IPv6 is not supported and must be disabled.



Note: Contact your OS vendor for help disabling IPv6.

See also *Configure Network Names*.

- Multihoming Support

Multihoming Cloudera Runtime or Cloudera Manager is not supported outside specifically certified Cloudera partner appliances. Cloudera finds that current Hadoop architectures combined with modern network infrastructures and security practices remove the need for multihoming. Multihoming, however, is beneficial internally in appliance form factors to take advantage of high-bandwidth InfiniBand interconnects.

Although some subareas of the product may work with unsupported custom multihoming configurations, there are known issues with multihoming. In addition, unknown issues may arise because multihoming is not covered by our test matrix outside the Cloudera-certified partner appliances.

- Entropy

Data at rest encryption requires sufficient entropy to ensure randomness.

See entropy requirements in *Data at Rest Encryption Requirements*.

- Cluster hosts must have a working network name resolution system and correctly formatted `/etc/hosts` file. All cluster hosts must have properly configured forward and reverse host resolution through DNS. The `/etc/hosts` files must:

- Contain consistent information about hostnames and IP addresses across all hosts
- Not contain uppercase hostnames
- Not contain duplicate IP addresses

Cluster hosts must not use aliases, either in `/etc/hosts` or in configuring DNS. A properly formatted `/etc/hosts` file should be similar to the following example:

```
127.0.0.1 localhost.localdomain localhost
192.168.1.1 cluster-01.example.com cluster-01
192.168.1.2 cluster-02.example.com cluster-02
192.168.1.3 cluster-03.example.com cluster-03
```

- In most cases, the Cloudera Manager Server must have SSH access to the cluster hosts when you run the installation or upgrade wizard. You must log in using a root account or an account that has password-less sudo permission. For authentication during the installation and upgrade procedures, you must either enter the password or upload a public and private key pair for the root or sudo user account. If you want to use a public and private key pair, the public key must be installed on the cluster hosts before you use Cloudera Manager.

Cloudera Manager uses SSH only during the initial install or upgrade. Once the cluster is set up, you can disable root SSH access or change the root password. Cloudera Manager does not save SSH credentials, and all credential information is discarded when the installation is complete.

- The Cloudera Manager Agent runs as root so that it can make sure that the required directories are created and that processes and files are owned by the appropriate user (for example, the hdfs and mapred users).
- Security-Enhanced Linux (SELinux) must not block Cloudera Manager or Runtime operations.



Note: Cloudera Enterprise is supported on platforms with Security-Enhanced Linux (SELinux) enabled and in enforcing mode. Cloudera is not responsible for SELinux policy development, support, or enforcement. If you experience issues running Cloudera software with SELinux enabled, contact your OS provider for assistance.

If you are using SELinux in enforcing mode, Cloudera Support can request that you disable SELinux or change the mode to permissive to rule out SELinux as a factor when investigating reported issues.

- Firewalls (such as iptables and firewalld) must be disabled or configured to allow access to ports used by Cloudera Manager, Runtime, and related services.
- For RHEL and CentOS, the `/etc/sysconfig/network` file on each host must contain the correct hostname.
- Cloudera Manager and Runtime use several user accounts and groups to complete their tasks. The set of user accounts and groups varies according to the components you choose to install. Do not delete these accounts or groups and do not modify their permissions and rights. Ensure that no existing systems prevent these accounts and groups from functioning. For example, if you have scripts that delete user accounts not in a whitelist, add these accounts to the list of permitted accounts. Cloudera Manager, Runtime, and managed services create and use the following accounts and groups:

Data at Rest Encryption Requirements

Encryption comprises several components, each with its own requirements.

Data at rest encryption protection can be applied at a number of levels within Hadoop:

- OS filesystem-level
- Network-level
- HDFS-level (protects both data at rest and in transit)

This section contains the various hardware and software requirements for all encryption products used for Data at Rest Encryption.

For more information on supported operating systems, see..

For more information on the components, concepts, and architecture for encrypting data at rest, see *Encrypting Data at Rest*.

Entropy Requirements

Cryptographic operations require entropy to ensure randomness.

You can check the available entropy on a Linux system by running the following command:

```
cat /proc/sys/kernel/random/entropy_avail
```

The output displays the entropy currently available. Check the entropy several times to determine the state of the entropy pool on the system. If the entropy is consistently low (500 or less), you must increase it by installing rng-tools and starting the rngd service.

For RHEL 7, run the following commands:

```
sudo yum install rng-tools
cp /usr/lib/systemd/system/rngd.service /etc/systemd/system/
sed -i -e 's/ExecStart=\/sbin\/rngd -f/ExecStart=\/sbin\/rngd -f -r \/dev\/u
random/' /etc/systemd/system/rngd.service
systemctl daemon-reload
systemctl start rngd
systemctl enable rngd
```

Make sure that the hosts running Key Trustee Server and Key Trustee KMS have sufficient entropy to perform cryptographic operations.

Cloudera Manager Requirements

Installing and managing Key Trustee Server using Cloudera Manager requires Cloudera Manager 5.4.0 and higher.

umask Requirements

Key Trustee Server installation requires the default umask of 0022.

Network Requirements

For new Key Trustee Server installations (5.4.0 and higher) and migrated upgrades, Key Trustee Server requires the following TCP ports to be opened for inbound traffic:

- 11371

Clients connect to this port over HTTPS.

- 11381 (PostgreSQL)

The passive Key Trustee Server connects to this port for database replication.

For upgrades that are not migrated to the CherryPy web server, the pre-upgrade port settings are preserved:

- 80

Clients connect to this port over HTTP to obtain the Key Trustee Server public key.

- 443 (HTTPS)

Clients connect to this port over HTTPS.

- 5432 (PostgreSQL)

The passive Key Trustee Server connects to this port for database replication.

TLS Certificate Requirements

To ensure secure network traffic, Cloudera recommends obtaining Transport Layer Security (TLS) certificates specific to the hostname of your Key Trustee Server. To obtain the certificate, generate a Certificate Signing Request (CSR) for the fully qualified domain name (FQDN) of the Key Trustee Server host. The CSR must be signed by a trusted Certificate Authority (CA). After the certificate has been verified and signed by the CA, the Key Trustee Server TLS configuration requires:

- The CA-signed certificate
- The private key used to generate the original CSR
- The intermediate certificate/chain file (provided by the CA)

Cloudera recommends not using self-signed certificates. If you use self-signed certificates, you must use the `--skip-ssl-check` parameter when registering Navigator Encrypt with the Key Trustee Server. This skips TLS hostname validation, which safeguards against certain network-level attacks. For more information regarding insecure mode, see *Registration Options*.

Trial Installation

These topics provide instructions for installing the trial version of CDP Private Cloud Base in a non-production environment for demonstration and proof-of-concept use cases.

In these procedures, Cloudera Manager automates the installation of the JDK, Cloudera Manager Server, an embedded PostgreSQL database, Cloudera Manager Agent, Cloudera Runtime, and other managed services on cluster hosts. Cloudera Manager also configures databases for the Cloudera Manager Server and Hive Metastore, Ranger, DAS, and for Cloudera Management Service roles.

This installation method is recommended for trial deployments, but is not supported for production deployments because it is not designed to scale. To use this method, server and cluster hosts must satisfy the following requirements:

- All hosts must have a [supported operating system](#) installed.
- You must be able to log in to the Cloudera Manager Server host using the root user account or an account that has passwordless sudo privileges.
- The Cloudera Manager Server host must have uniform SSH access on the same port to all hosts. For more information, see [Runtime and Cloudera Manager Networking and Security Requirements](#).
- All hosts must have access to standard package repositories for the operating system and either `archive.cloudera.com` or a local repository with the required installation files.
- SELinux must be disabled or set to permissive mode before running the installer.

Related Information

[CDP Private Cloud Base Trial Download Information](#)

[CDP Private Cloud Base Installation Guide](#)

Before You Begin a Trial Installation

Before you begin a trial installation, you must disable SELinux if you want the Cloudera Manager installer to run. You can also optionally configure an HTTP proxy.

(Optional) Configure an HTTP Proxy

The Cloudera Manager installer accesses archive.cloudera.com by using yum on RHEL systems, zypper on SLES systems, or apt-get on Ubuntu systems. If your hosts access the Internet through an HTTP proxy, you can configure yum system-wide, to access archive.cloudera.com through a proxy.

To do so, modify the system configuration on every cluster host as follows:

OS	File	Property
RHEL-compatible	/etc/yum.conf	proxy=http://server:port/

Disable SELinux



Note: CDP Private Cloud Base is supported on platforms with Security-Enhanced Linux (SELinux) enabled and in enforcing mode. Cloudera is not responsible for SELinux policy development, support, or enforcement. If you experience issues running Cloudera software with SELinux enabled, contact your OS provider for assistance.

If you are using SELinux in enforcing mode, Cloudera Support can request that you disable SELinux or change the mode to permissive to rule out SELinux as a factor when investigating reported issues.

Although Cloudera supports running Cloudera software with SELinux enabled, the Cloudera Manager installer will not proceed if SELinux is enabled. Disable SELinux or set it to permissive mode before running the installer.

After you have installed and deployed Cloudera Manager and Runtime, you can re-enable SELinux by changing SELINUX=permissive back to SELINUX=enforcing in /etc/selinux/config (or /etc/sysconfig/selinux), and then running the following command to immediately switch to enforcing mode:

```
setenforce 1
```

If you are having trouble getting Cloudera Software working with SELinux, contact your OS vendor for support. Cloudera is not responsible for developing or supporting SELinux policies.

Installing a Trial Cluster

In this procedure, Cloudera Manager automates the installation of the Oracle JDK, Cloudera Manager Server, embedded PostgreSQL database, Cloudera Manager Agent, Runtime, and managed service software on cluster hosts. Cloudera Manager also configures databases for the Cloudera Manager Server and Hive Metastore and optionally for Cloudera Management Service roles.



Important: This procedure is intended for trial and proof-of-concept deployments only. It is not supported for production deployments because it is not designed to scale.

Cluster Host Requirements:

The hosts you intend to use must satisfy the following requirements:

- You must be able to log in to the Cloudera Manager Server host using the root user account or an account that has passwordless sudo privileges.
- The Cloudera Manager Server host must have uniform SSH access on the same port to all hosts. For more information, see *Runtime and Cloudera Manager Networking and Security Requirements*.
- All hosts must have access to standard package repositories for the operating system and either archive.cloudera.com or a local repository with the required installation files.

- SELinux must be disabled or set to permissive mode before running the installer.

Refer to the following topics for the steps required to install a trial cluster.

Step 1: Run the Cloudera Manager Server Installer

Download the Cloudera Manager installer to the cluster host to which you are installing the Cloudera Manager Server. By default, the automated installer binary (cloudera-manager-installer.bin) installs the highest version of Cloudera Manager.

Before you begin

For information on downloading the CDP Private Cloud Base Trial installer, see [CDP Private Cloud Base Trial Download Information](#) on page 8.

Procedure

1. Run the Cloudera Manager installer:

- a) Change cloudera-manager-installer.bin to have execute permissions:

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

- b) Run the Cloudera Manager Server installer:

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

- c) For clusters without Internet access: Install Cloudera Manager packages from a [local repository](#):

```
sudo ./cloudera-manager-installer.bin --skip_repo_package=1
```

2. Read and accept the associated license agreements. After you accept the licenses, the installer does the following:

- a. Installs the Cloudera Manager repository files.
- b. Installs the Oracle JDK.
- c. Installs the Cloudera Manager Server and embedded PostgreSQL packages.
- d. Starts the embedded PostgreSQL database and Cloudera Manager Server.



Note: If the installation is interrupted, run the following command on the Cloudera Manager Server host before you retry the installation:

```
sudo /usr/share/cmfd/uninstall-cloudera-manager.sh
```

Log files for the installer are stored in /var/log/cloudera-manager-installer/.

3. Exit the installer:

- a) When the installation completes, the complete URL for the Cloudera Manager Admin Console displays, including the port number (7180 by default). Make a note of this URL.
- b) Press Enter to choose OK to exit the installer, and then again to acknowledge the successful installation.
- c) Wait several minutes for the Cloudera Manager Server to start. To observe the startup process, run `sudo tail -f /var/log/cloudera-scm-server/cloudera-scm-server.log` on the Cloudera Manager Server host. When you see the following log entry, the Cloudera Manager Admin Console is ready:

```
INFO WebServerImpl:com.cloudera.server.cmf.WebServerImpl: Started Jetty server.
```

If the Cloudera Manager Server does not start, see *Troubleshooting Installation Problems*.

Step 2: Install Cloudera Runtime Using the Wizard

Proceed through the installation wizard to specify hosts, install and configure Cloudera Runtime, and more.

Log Into the Cloudera Manager Admin Console

1. 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.
2. Log into Cloudera Manager Admin Console. The default credentials are:

Username: admin

Password: admin



Note: Cloudera Manager does not support changing the admin username for the installed account. You can change the password using Cloudera Manager after you run the installation wizard. Although you cannot change the admin username, you can add a new user, assign administrative privileges to the new user, and then delete the default admin account.

Upload License File

On the Upload License File page, you can select either the trial version of CDP Data Center or upload a license file:

1. Choose one of the following options:
 - Upload Cloudera Data Platform License
 - Try Cloudera Data Platform for 60 days. The CDP Data Center trial does not require a license file, but the trial expires after 60 days.
2. If you choose the CDP Data Center Edition Trial, you can upload a license file at a later time. Read the license agreement and click the checkbox labeled Yes, I accept the Cloudera Standard License Terms and Conditions if you accept the terms and conditions of the license agreement. Then click Continue.
3. If you have a license file for CDP Data Center, upload the license file:
 - a. Select Upload Cloudera Data Platform License.
 - b. Click Upload License File.
 - c. Browse to the location of the license file, select the file, and click Open.
 - d. Click Upload.
 - e. Click Continue.
4. Click Continue to proceed with the installation.

The Welcome page displays.

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.

Cluster Basics

The Cluster Basics page allows you to specify the Cluster Name

For new installations, a Regular Cluster (also called a base cluster) is the only option. You can add a compute cluster after you finish installing the base cluster.

For more information on regular and compute clusters, and data contexts, see [Virtual Private Clusters and Cloudera SDX](#).

Enter a cluster name and click Continue.

Specify Hosts

Choose which hosts will run Runtime and other managed services.



Note: If you have enabled Auto-TLS, you must include the Cloudera Manager server host when you specify hosts.

1. To enable Cloudera Manager to automatically discover hosts on which to install Runtime and managed services, enter the cluster hostnames or IP addresses in the Hostnames field. You can specify hostname and IP address ranges as follows:

Expansion Range	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].example.com	host1.example.com, host2.example.com, host3.example.com
host[07-10].example.com	host07.example.com, host08.example.com, host09.example.com, host10.example.com



Important: Unqualified hostnames (short names) must be unique in a Cloudera Manager instance. For example, you cannot have both *host01.example.com* and *host01.standby.example.com* managed by the same Cloudera Manager Server.

You can specify multiple addresses and address ranges by separating them with commas, semicolons, tabs, or blank spaces, or by placing them on separate lines. Use this technique to make more specific searches instead of searching overly wide ranges. Only scans that reach hosts running SSH will be selected for inclusion in your cluster by default. You can enter an address range that spans over unused addresses and then clear the nonexistent hosts later in the procedure, but wider ranges require more time to scan.

2. Click Search. If there are a large number of hosts on your cluster, wait a few moments to allow them to be discovered and shown in the wizard. If the search is taking too long, you can stop the scan by clicking Abort Scan. You can modify the search pattern and repeat the search as many times as you need until you see all of the expected hosts.



Note: Cloudera Manager scans hosts by checking for network connectivity. If there are some hosts where you want to install services that are not shown in the list, make sure you have network connectivity between the Cloudera Manager Server host and those hosts, and that firewalls and SELinux are not blocking access.

3. Verify that the number of hosts shown matches the number of hosts where you want to install services. Clear host entries that do not exist or where you do not want to install services.
4. Click Continue.

The Select Repository screen displays.

Select Repository



Important: You cannot install software using both parcels and packages in the same cluster.

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*.

In the CDH and other software section:

1. Select the repository type to use for the installation. In the Install Method section select one of the following:
 - Use Parcels (Recommended)

A parcel is a binary distribution format containing the program files, along with additional metadata used by Cloudera Manager. Parcels are required for rolling upgrades. For more information, see *Parcels*.

2. Select the version of Cloudera Runtime or CDH to install. If you do not see the version you want to install:

- **Parcels** – Click the Parcel Repository & Network Settings link to add the repository URL for your version. If you are using a local Parcel repository, enter its URL as the repository URL.

Repository URLs for CDH 6 parcels are documented in [CDH 6 Download Information](#)

Repository URLs for the Cloudera Runtime 7 parcels are documented in [Cloudera Runtime Download Information](#)



Important: If you are installing Cloudera Runtime 7.1.5.0 and you have selected to use a 60-day trial license, use the following Parcel Repository URL:

```
https://archive.cloudera.com/cdh7/7.1.5.0/parcels/
```

After adding the repository, click Save Changes and wait a few seconds for the version to appear. If your Cloudera Manager host uses an HTTP proxy, click the Proxy Settings button to configure your proxy.

Note that if you have a Cloudera Enterprise license and are using Cloudera Manager 6.3.3 or higher to install a CDH version 6.3.3 or higher, or a Cloudera Runtime version 7.0 or higher using parcels, you do not need to add a username and password or "@" to the parcel repository URL. Cloudera Manager will authenticate to the Cloudera archive using the information in your license key file. Use a link to the repository in the following format:

```
https://archive.cloudera.com/p/cdh6/6.x.x/parcels/
```

If you are using a version of CM older than 6.3.3 to install CDH 6.3.3 or higher parcels, you must include the username/password and "@" in the repository URL during installation or when you configure a CDH 6.3.3 or higher parcel repository. After you add the repository, click Save Changes and wait a few seconds for the version to appear. If your Cloudera Manager host uses an HTTP proxy, click the Proxy Settings button to configure your proxy.



Note: Cloudera Manager only displays CDH versions it can support. If an available CDH version is too new for your Cloudera Manager version, it is not displayed. If the parcels do not appear on the Parcels page, ensure that the Parcel URL you entered is correct.



Note: Cloudera Manager only displays Cloudera Runtime versions it can support. If an available CDH version is too new for your Cloudera Manager version, it is not displayed.

3. If you selected Use Parcels, specify any Additional Parcels you want to install.

4. Click Continue.

Select JDK



Note: CDP Data Center is no longer bundled with Oracle JDK software. Cloudera provides a supported version of OpenJDK.

If you installed your own JDK version, such as Oracle JDK 8, in *Step 2: Install Java Development Kit*, select Manually manage JDK.

To allow Cloudera Manager to automatically install the OpenJDK on cluster hosts, select Install a Cloudera-provided version of OpenJDK.

To install the default OpenJDK that is provided by your operating system, select Install a system-provided version of OpenJDK.

After checking the applicable boxes, click Continue.

Enter Login Credentials

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

2. Select an authentication method:

- If you choose password authentication, enter and confirm the password.
- If you choose public-key authentication, provide a passphrase and path to the required key files.

You can modify the default SSH port if necessary.

3. Specify the maximum number of host installations to run at once. The default and recommended value is 10. You can adjust this based on your network capacity.
4. Click Continue.

The Install Agents page displays.

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.

If you selected the option to manually install agents, see *Manually Install Cloudera Manager Agent Packages* for the procedure and then continue with the next steps on this page.

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

If you are using parcels, the Install Parcels page displays. If you chose to install using packages, the Inspect Cluster page displays.

Install Parcels

If you selected parcels for the installation method, the Install Parcels page reports the installation progress of the parcels you selected earlier. 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.

After the installation is complete, click Continue.

The Inspect Cluster page displays.

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. You can click Advanced Options to customize some ping parameters.
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, and you want to abandon the cluster creation wizard to address them, select the radio button labeled Quit the wizard and Cloudera Manager will delete the temporarily created cluster and then click Continue.

Otherwise, after addressing any identified problems, 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.

Continue to *Step 7: Set Up a Cluster Using the Wizard*.

Step 3: Set Up a Cluster Using the Wizard

After completing the Cluster Installation wizard, the Cluster Setup wizard automatically starts. The following sections guide you through each page of the wizard.

Select Services

The Select Services page allows you to select the services you want to install and configure. Make sure that you have the appropriate license key for the services you want to use.

You can choose from:

Regular (Base) Clusters

Data Engineering

Process develop, and serve predictive models.

Services included: HDFS, YARN, YARN Queue Manager, Ranger, Atlas, Hive, Hive on Tez, Spark, Oozie, Hue, and Data Analytics Studio

Data Mart

Browse, query, and explore your data in an interactive way.

Services included: HDFS, Ranger, Atlas, Hive, and Hue

Operational Database

Real-time insights for modern data-driven business.

Services included: HDFS, Ranger, Atlas, and HBase

Custom Services

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

Compute Clusters

Data Engineering

Process develop, and serve predictive models.

Services included: Spark, Oozie, Hive on Tez, Data Analytics Studio, HDFS, YARN, and YARN Queue Manager

Spark

Spark for Compute

Services included: Core Configuration, Spark, Oozie, YARN, and YARN Queue Manager

Data Mart

Impala for Compute

Services included: Core Configuration, Impala, and Hue

Streams Messaging (Simple)

Simple Kafka cluster for streams messaging

Services included: Kafka, Schema Registry, and Zookeeper

Streams Messaging (Full)

Advanced Kafka cluster with monitoring and replication services for streams messaging

Services included: Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager, Cruise Control, and Zookeeper

Custom Services

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

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

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.

To review the recommended role assignments, see *Recommended Cluster Hosts and Role Distribution*.

After assigning all of the roles for your services, click Continue. The Setup Database page displays.

Setup Database

When using the Cloudera Manager installer with the embedded database, the Setup Database page is pre-populated with the database names and passwords. 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. Some databases will be created in a future step. For these, the words Skipped. Cloudera Manager will create this database in a later step. appear next to the green checkmark.

After verifying that each connection is successful, click Continue. The Review Changes page displays.

Enter Required Parameters

The **Enter Required Parameters** page lists required parameters for DAS, the Cloudera Manager API client, and Ranger.

The DAS database hostname, database name, database username, and database password were configured when you created the required DAS database. The default database name is “das” and the default database user is “das”.

If you do not have an existing user for the Cloudera Manager API client, use the default username and password “admin” for both the The Existing Cloudera Manager API Client Username and The Existing Cloudera Manager API Client Password.

The Ranger Admin user, Usersync user, Tagsync User, and KMS Keyadmin User are created during cluster deployment. In this page you must give a password for each of these users.



Note: Passwords for the Ranger Admin, Usersync, Tagsync, and KMS Keyadmin users must be a minimum of 8 characters long, with at least one alphabetic and one numeric character. The following characters are not valid: " ' \ ` ^ .

The Ranger database host, name, user, and user password were configured when you created the required Ranger database. If you ran the `gen_embedded_ranger_db.sh` script to create the Ranger database, the output of the script contained the host and database user password. Enter those here. The default database name is “ranger” and the default database user is “rangeradmin.”

Review Changes

The Review Changes page lists default and suggested settings for several configuration parameters, including data directories.



Warning: Do not place DataNode data directories on NAS devices. When resizing an NAS, block replicas can be deleted, which results in missing blocks.

Review and make any necessary changes, and then click Continue. The Command Details page displays.

Command Details

The Command Details page lists the details of the First Run command. You can expand the running commands to view the details of any step, including log files and command output. You can filter the view by selecting Show All Steps, Show Only Failed Steps, or Show Running Steps.

After the First Run command completes, click Continue to go to the Summary page.

Summary

The Summary page reports the success or failure of the setup wizard. Click Finish to complete the wizard. The installation is complete.

Cloudera recommends that you change the default password as soon as possible by clicking the logged-in username at the top right of the home screen and clicking Change Password.

Stopping the Embedded PostgreSQL Database

To stop the embedded PostgreSQL database, stop the services and servers in the order listed below.

Procedure

1. Log into the Cloudera Manager user interface and stop the services that have a dependency on the Hive metastore (Hue, Impala, and Hive) in the following order:
 - Stop the Hue and Impala services.
 - Stop the Hive service.
2. On the Cloudera Manager **Home** page, click the 3 vertical dots next to Cloudera Management Service and select Stop to stop the Cloudera Management Service.
3. Stop the Cloudera Manager Server.

RHEL 7:

```
sudo systemctl stop cloudera-scm-server.service
```

4. Stop the Cloudera Manager Server database.

RHEL 7:

```
sudo systemctl stop cloudera-scm-server-db.service
```

Starting the Embedded PostgreSQL Database

To start the embedded PostgreSQL database, start the servers and services in the order listed below.

Procedure

1. Start the Cloudera Manager Server database.

RHEL 7:

```
sudo systemctl start cloudera-scm-server-db.service
```

2. Start the Cloudera Manager Server.

RHEL 7:

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

3. Log into Cloudera Manager and start the Cloudera Manager Service. On the Cloudera Manager **Home** page, click the 3 vertical dots next to Cloudera Management Service and select Start.
4. In the Cloudera Manager user interface, start the services that have a dependency on the Hive metastore (Hue, Impala, and Hive) in the following order:
 - Start the Hive service.
 - Start the Hue and Impala services.

Changing Embedded PostgreSQL Database Passwords

The embedded PostgreSQL database has generated user accounts and passwords. You can change a password associated PostgreSQL database account.

About this task

You can see the generated accounts and passwords during the installation process and you should record them at that time.

To find information about the PostgreSQL database account that the Cloudera Manager Server uses, read the `/etc/cloudera-scm-server/db.properties` file:

```
# cat /etc/cloudera-scm-server/db.properties

Auto-generated by scm_prepare_database.sh
#
Sat Oct 1 12:19:15 PDT 201
#
com.cloudera.cmf.db.type=postgresql
com.cloudera.cmf.db.host=localhost:7432
com.cloudera.cmf.db.name=scm
com.cloudera.cmf.db.user=scm
com.cloudera.cmf.db.password=TXqEESuhj5
```

To change a password associated with a PostgreSQL database account:

Procedure

1. Obtain the root password from the `/var/lib/cloudera-scm-server-db/data/generated_password.txt` file:

```
# cat /var/lib/cloudera-scm-server-db/data/generated_password.txt

MnPwGeWaip

The password above was generated by /usr/share/cmf/bin/initialize_embedded_db.sh (part of the cloudera-scm-server-db package)
and is the password for the user 'cloudera-scm' for the database in the
current directory.

Generated at Fri Jun 29 16:25:43 PDT 2012.
```

2. On the host on which the Cloudera Manager Server is running, log into PostgreSQL as the root user:

```
psql -U cloudera-scm -p 7432 -h localhost -d postgres
Password for user cloudera-scm: MnPwGeWaip
psql (8.4.18)
Type "help" for help.

postgres=#
```

3. Determine the database and owner names:

```
postgres=# \l
```

List of databases					
Name	Owner	Encoding	Collation	Ctype	Access privileges
amon	amon	UTF8	en_US.UTF8	en_US.UTF8	
hive	hive	UTF8	en_US.UTF8	en_US.UTF8	

```

nav          nav          UTF8          en_US.UTF8    en_US.UTF8
navms        navms        UTF8          en_US.UTF8    en_US.UTF8
postgres     cloudera-scm  UTF8          en_US.UTF8    en_US.UTF8
rman         rman         UTF8          en_US.UTF8    en_US.UTF8
scm          scm          UTF8          en_US.UTF8    en_US.UTF8
template0    cloudera-scm  UTF8          en_US.UTF8    en_US.UTF8
cloudera-scm"                                =c/"clou
                                                : "cloudera
-scm"=CTc/"cloudera-scm"
template1    cloudera-scm  UTF8          en_US.UTF8    en_US.UTF8    =c/"cl
cloudera-scm"                                : "cloude
ra-scm"=CTc/"cloudera-scm"
(9 rows)

```

4. Set the password for an owner using the `\password` command. For example, to set the password for the `amon` owner, do the following:

```

postgres=# \password amon
Enter new password:
Enter it again:

```

5. Configure the role with the new password:
 - a) In the Cloudera Manager Admin Console, select ClustersCloudera Management Service.
 - b) Click the Configuration tab.
 - c) In the Scope section, select the role where you are configuring the database.
 - d) Select CategoryDatabase category.
 - e) Set the *Role Name Database Password* property.
 - f) Enter a Reason for change, and then click Save Changes to commit the changes.

Migrating from the Cloudera Manager Embedded PostgreSQL Database Server to an External PostgreSQL Database

If you have already used the embedded PostgreSQL database and you are unable to redeploy a fresh cluster, you must migrate the embedded PostgreSQL database server to an external PostgreSQL database.

Cloudera Manager provides an embedded PostgreSQL database server for trial and proof of concept deployments when creating a cluster. To remind users that this embedded database is not suitable for production, Cloudera Manager displays the banner text: "You are running Cloudera Manager in non-production mode, which uses an embedded PostgreSQL database. Switch to using a supported external database before moving into production."

If, however, you have already used the embedded database, and you are unable to redeploy a fresh cluster, then you must migrate to an external PostgreSQL database.



Note: This procedure does not describe how to migrate to a database server other than PostgreSQL. Moving databases from one database server to a different type of database server is a complex process that requires modification of the schema and matching the data in the database tables to the new schema. It is strongly recommended that you engage with Cloudera Professional Services if you wish to perform a migration to an external database server other than PostgreSQL.

Prerequisites

Before migrating the Cloudera Manager embedded PostgreSQL database to an external PostgreSQL database, ensure that your setup meets the following conditions:

- The external PostgreSQL database server is running.
- The database server is configured to accept remote connections.
- The database server is configured to accept user logins using `md5`.

- No one has manually created any databases in the external database server for roles that will be migrated.



Note: To view a list of databases in the external database server (requires default superuser permission):

```
sudo -u postgres psql -l
```

- All health issues with your cluster have been resolved.

For details about configuring the database server, see *Configuring and Starting the PostgreSQL Server*.



Important: Only perform the steps in *Configuring and Starting the PostgreSQL Server*. Do not proceed with the creation of databases as described in the subsequent section.

For large clusters, Cloudera recommends running your database server on a dedicated host. Engage Cloudera Professional Services or a certified database administrator to correctly tune your external database server.

Identify Roles that Use the Embedded Database Server

Before you can migrate to another database server, you must first identify the databases using the embedded database server.

About this task

When the Cloudera Manager Embedded Database server is initialized, it creates the Cloudera Manager database and databases for roles in the Management Services. The Installation Wizard (which runs automatically the first time you log in to Cloudera Manager) or Add Service action for a cluster creates additional databases for roles when run. It is in this context that you identify which roles are used in the embedded database server.

To identify which roles are using the Cloudera Manager embedded database server:

Procedure

1. Obtain and save the cloudera-scm superuser password from the embedded database server. You will need this password in subsequent steps:

```
head -1 /var/lib/cloudera-scm-server-db/data/generated_password.txt
```

2. Make a list of all services that are using the embedded database server. Then, after determining which services are not using the embedded database server, remove those services from the list. The scm database must remain in your list. Use the following table as a guide:

Table 15: Cloudera Manager Embedded Database Server Databases

Service	Role	Default Database Name	Default Username
Cloudera Manager Server		scm	scm
Cloudera Management Service	Activity Monitor	amon	amon
Hive	Hive Metastore Server	hive	hive
Hue	Hue Server	hue	7uu7uu7uhue
Oozie	Oozie Server	oozie_oozie_server	oozie_oozie_server
Cloudera Management Service	Reports Manager	rman	rman
DAS		das	das
Ranger		ranger	rangeradmin

3. Verify which roles are using the embedded database. Roles using the embedded database server always use port 7432 (the default port for the embedded database) on the Cloudera Manager Server host.
 - a. Verify which roles are using the embedded database. Roles using the embedded database server always use port 7432 (the default port for the embedded database) on the Cloudera Manager Server host.

For Cloudera Management Services:

1. Select Cloudera Management Service > Configuration, and type "7432" in the Search field.
2. Confirm that the hostname for the services being used is the same hostname used by the Cloudera Manager Server.



Note:

If any of the following fields contain the value "7432", then the service is using the embedded database:

- Activity Monitor
- Reports Manager

For the Oozie Service:

1. Select Oozie service > Configuration, and type "7432" in the Search field.
2. Confirm that the hostname is the Cloudera Manager Server.

For Hive and Hue Services:

1. Select the specific service > Configuration, and type "database host" in the Search field.
 2. Confirm that the hostname is the Cloudera Manager Server.
 3. In the Search field, type "database port" and confirm that the port is 7432.
 4. Repeat these steps for each of the services (Hive and Hue).
4. Verify the database names in the embedded database server match the database names on your list (Step 2). Databases that exist on the database server and not used by their roles do not need to be migrated. This step is to confirm that your list is correct.



Note: Do not add the postgres, template0, or template1 databases to your list. These are used only by the PostgreSQL server.

```
psql -h localhost -p 7432 -U cloudera-scm -l
```

```
Password for user cloudera-scm: <password>
```

		List of databases			
Name	Access	Owner	Encoding	Collate	Ctype
amon		amon	UTF8	en_US.UTF8	en_US.U
TF8					
hive		hive	UTF8	en_US.UTF8	en_US.UT
F8					
hue		hue	UTF8	en_US.UTF8	en_US
.UTF8					
navms		navms	UTF8	en_US.UTF8	en_US.
UTF8					
oozie_oozie_server		oozie_oozie_server	UTF8	en_US.UTF8	en_US.U
TF8					
postgres		cloudera-scm	UTF8	en_US.UTF8	en_US.UT
F8					
rman		rman	UTF8	en_US.UTF8	en_US
.UTF8					

```

scm | scm | UTF8 | en_US.UTF8 | en_US.
UTF8 |
template0 | cloudera-scm | UTF8 | en_US.UTF8 | en_US.U
TF8 | =c/"cloudera-scm"
template1 | cloudera-scm | UTF8 | en_US.UTF8 | en_US.
UTF8 | =c/"cloudera-scm"
(12 rows)

```

Results

You should now have a list of all roles and database names that use the embedded database server, and are ready to proceed with the migration of databases from the embedded database server to the external PostgreSQL database server.

Migrate Databases from the Embedded Database Server to the External PostgreSQL Database Server

After you identify the roles that use the embedded database, you are ready to migrate from the embedded database server to an external PostgreSQL database server.

About this task

While performing this procedure, ensure that the Cloudera Manager Agents remain running on all hosts. Unless otherwise specified, when prompted for a password use the cloudera-scm password.



Note: After completing this migration, you cannot delete the cloudera-scm postgres superuser unless you remove the access privileges for the migrated databases. Minimally, you should change the cloudera-scm postgres superuser password.

Procedure

1. In Cloudera Manager, stop the cluster services identified as using the embedded database server. Be sure to stop the Cloudera Management Service as well. Also be sure to stop any services with dependencies on these services. The remaining Runtime services will continue to run without downtime.



Note: If you do not stop the services from within Cloudera Manager before stopping Cloudera Manager Server from the command line, they will continue to run and maintain a network connection to the embedded database server. If this occurs, then the embedded database server will ignore any command line stop commands (Step 2) and require that you manually kill the process, which in turn causes the services to crash instead of stopping cleanly.

2. Navigate to Hosts > All Hosts, and make note of the number of roles assigned to hosts. Also take note whether or not they are in a commissioned state. You will need this information later to validate that your scm database was migrated correctly.
3. Stop the Cloudera Manager Server. To stop the server:

```
sudo service cloudera-scm-server stop
```

4. Obtain and save the embedded database superuser password (you will need this password in subsequent steps) from the generated_password.txt file:

```
head -1 /var/lib/cloudera-scm-server-db/data/generated_password.txt
```

5. Export the PostgreSQL user roles from the embedded database server to ensure the correct users, permissions, and passwords are preserved for database access. Passwords are exported as an md5sum and are not visible in plain text. To export the database user roles (you will need the cloudera-scm user password):

```
pg_dumpall -h localhost -p 7432 -U cloudera-scm -v --roles-only -f "/var/tmp/cloudera_user_roles.sql"
```


6. Edit `/var/tmp/cloudera_user_roles.sql` to remove any `CREATE ROLE` and `ALTER ROLE` commands for databases not in your list. Leave the entries for `cloudera-scm` untouched, because this user role is used during the database import.
7. Export the data from each of the databases on your list you created in *Identify Roles that Use the Embedded Database Server*:

```
pg_dump -F c -h localhost -p 7432 -U cloudera-scm [database_name] > /var/
tmp/[database_name]_db_backup-$(date +%m-%d-%Y).dump
```

The following is a sample data export command for the `scm` database:

```
pg_dump -F c -h localhost -p 7432 -U cloudera-scm scm > /var/tmp/scm_db_
backup-$(date +%m-%d-%Y).dump
```

Password:

8. Stop and disable the embedded database server:

```
service cloudera-scm-server-db stop
chkconfig cloudera-scm-server-db off
```

Confirm that the embedded database server is stopped:

```
netstat -at | grep 7432
```

9. Back up the Cloudera Manager Server database configuration file:

```
cp /etc/cloudera-scm-server/db.properties /etc/cloudera-scm-server/db.pr
operties.embedded
```

10. Copy the file `/var/tmp/cloudera_user_roles.sql` and the database dump files from the embedded database server host to `/var/tmp` on the external database server host:

```
cd /var/tmp
scp cloudera_user_roles.sql *.dump <user>@<postgres-server>:/var/tmp
```

11. Import the PostgreSQL user roles into the external database server.

The external PostgreSQL database server superuser password is required to import the user roles. If the superuser role has been changed, you will be prompted for the username and password.



Note: Only run the command that applies to your context; do not execute both commands.

- To import users when using the default PostgreSQL superuser role:

```
sudo -u postgres psql -f /var/tmp/cloudera_user_roles.sql
```

- To import users when the superuser role has been changed:

```
psql -h <database-hostname> -p <database-port> -U <superuser> -f /var/tm
p/cloudera_user_roles.sql
```

For example:

```
psql -h pg-server.example.com -p 5432 -U postgres -f /var/tmp/cloudera_u
ser_roles.sql
```

Password for user postgres

12. Import the Cloudera Manager database on the external server. First copy the database dump files from the Cloudera Manager Server host to your external PostgreSQL database server, and then import the database data:



Note: To successfully run the `pg_restore` command, there must be an existing database on the database server to complete the connection; the existing database will not be modified. If the `-d <existing-database>` option is not included, then the `pg_restore` command will fail.

```
pg_restore -C -h <database-hostname> -p <database-port> -d <existing-database> -U cloudera-scm -v <data-file>
```

Repeat this import for each database.

The following example is for the scm database:

```
pg_restore -C -h pg-server.example.com -p 5432 -d postgres -U cloudera-scm -v /var/tmp/scm_server_db_backup-20180312.dump
```

```
pg_restore: connecting to database for restore
Password:
```

13. Update the Cloudera Manager Server database configuration file to use the external database server. Edit the `/etc/cloudera-scm-server/db.properties` file as follows:
- Update the `com.cloudera.cmf.db.host` value with the hostname and port number of the external database server.
 - Change the `com.cloudera.cmf.db.setupType` value from "EMBEDDED" to "EXTERNAL".
14. Start the Cloudera Manager Server and confirm it is working:

```
service cloudera-scm-server start
```

Note that if you start the Cloudera Manager GUI at this point, it may take up to five minutes after executing the start command before it becomes available.

In Cloudera Manager Server, navigate to **Hosts > All Hosts** and confirm the number of roles assigned to hosts (this number should match what you found in Step 2); also confirm that they are in a commissioned state that matches what you observed in Step 2.

15. Update the role configurations to use the external database hostname and port number. Only perform this task for services where the database has been migrated.

For Cloudera Management Services:

- Select **Cloudera Management Service > Configuration**, and type "7432" in the Search field.
- Change any database hostname properties from the embedded database to the external database hostname and port number.
- Click **Save Changes**.

For the Oozie Service:

- Select **Oozie service > Configuration**, and type "7432" in the Search field.
- Change any database hostname properties from the embedded database to the external database hostname and port number.
- Click **Save Changes**.

For Hive and Hue Services:

- Select the specific service **> Configuration**, and type "database host" in the Search field.
- Change the hostname from the embedded database name to the external database hostname.
- Click **Save Changes**.

16. Start the Cloudera Management Service and confirm that all management services are up and no health tests are failing.

17. Start all Services via the Cloudera Manager web UI. This should start all services that were stopped for the database migration. Confirm that all services are up and no health tests are failing.
18. On the embedded database server host, remove the embedded PostgreSQL database server:
 - a) Make a backup of the /var/lib/cloudera-scm-server-db/data directory:

```
tar czvf /var/tmp/embedded_db_data_backup-$(date +%m-%d-%Y).tgz /var/lib/cloudera-scm-server-db/data
```

- b) Remove the embedded database package:

For RHEL/SLES:

```
rpm --erase cloudera-manager-server-db-2
```

For Ubuntu:

```
apt-get remove cloudera-manager-server-db-2
```

- c) Delete the /var/lib/cloudera-scm-server-db/data directory.

Production Installation

These topics provide procedures for installing CDP Private Cloud Base in a production environment.

Related Information

[CDP Private Cloud Base Installation Guide](#)

Before You Install

Before you begin a production installation of Cloudera Manager, Cloudera Runtime, and other managed services, review the Cloudera Data Platform 7 Requirements and Supported Versions, in addition to the Cloudera Data Platform Release Notes.

For planning, best practices, and recommendations, review the reference architecture for your environment. For example, for on-premises deployments, review the Cloudera Enterprise Reference Architecture for Bare Metal Deployments (PDF). In addition, the importance of security in a production environment cannot be understated. TLS and Kerberos form the baseline for secure operations of your CDP Runtime environment; Cloudera supports security services such as Ranger and Atlas only when they are run on clusters where Kerberos is enabled to authenticate users.

The following topics describe additional considerations you should be aware of before beginning an installation:

Storage Space Planning for Cloudera Manager

This topic helps you plan for the storage needs and data storage locations used by the Cloudera Manager Server and the Cloudera Management Service to store metrics and data.

Minimum Required Role: [Full Administrator](#). This feature is not available when using Cloudera Manager to manage Data Hub clusters.

Cloudera Manager tracks metrics of services, jobs, and applications in many background processes. All of these metrics require storage. Depending on the size of your organization, this storage can be local or remote, disk-based or in a database, managed by you or by another team in another location.

Most system administrators are aware of common locations like /var/log/ and the need for these locations to have adequate space. Failing to plan for the storage needs of all components of the Cloudera Manager Server and the Cloudera Management Service can negatively impact your cluster in the following ways:

- The cluster might not be able to retain historical operational data to meet internal requirements.

- The cluster might miss critical audit information that was not gathered or retained for the required length of time.
- Administrators might be unable to research past events or health status.
- Administrators might not have historical MR1, YARN, or Impala usage data when they need to reference or report on them later.
- There might be gaps in metrics collection and charts.
- The cluster might experience data loss due to filling storage locations to 100% of capacity. The effects of such an event can impact many other components.

The main theme here is that you must architect your data storage needs well in advance. You must inform your operations staff about your critical data storage locations for each host so that they can provision your infrastructure adequately and back it up appropriately. Make sure to document the discovered requirements in your internal build documentation and run books.

This topic describes both local disk storage and RDBMS storage. This distinction is made both for storage planning and also to inform migration of roles from one host to another, preparing backups, and other lifecycle management events.

The following tables provide details about each individual Cloudera Management service to enable Cloudera Manager administrators to make appropriate storage and lifecycle planning decisions.

Table 16: Cloudera Manager Server

Configuration Topic	Cloudera Manager Server Configuration
Default Storage Location	<p>RDBMS:</p> <p>Any Supported RDBMS.</p> <p>Disk:</p> <p>Cloudera Manager Server Local Data Storage Directory (command_storage_path) on the host where the Cloudera Manager Server is configured to run. This local path is used by Cloudera Manager for storing data, including command result files. Critical configurations are not stored in this location.</p> <p>Default setting: /var/lib/cloudera-scm-server/</p>
Storage Configuration Defaults, Minimum, or Maximum	There are no direct storage defaults relevant to this entity.
Where to Control Data Retention or Size	<p>The size of the Cloudera Manager Server database varies depending on the number of managed hosts and the number of discrete commands that have been run in the cluster. To configure the size of the retained command results in the Cloudera Manager Administration Console, select AdministrationSettings and edit the following property:</p> <p>Command Eviction Age</p> <p>Length of time after which inactive commands are evicted from the database.</p> <p>Default is two years.</p>
Sizing, Planning & Best Practices	<p>The Cloudera Manager Server database is the most vital configuration store in a Cloudera Manager deployment. This database holds the configuration for clusters, services, roles, and other necessary information that defines a deployment of Cloudera Manager and its managed hosts.</p> <p>Make sure that you perform regular, verified, remotely-stored backups of the Cloudera Manager Server database.</p>

Table 17: Cloudera Management Service - Activity Monitor Configuration

Configuration Topic	Activity Monitor
Default Storage Location	Any Supported RDBMS.
Storage Configuration Defaults / Minimum / Maximum	Default: 14 Days worth of MapReduce (MRv1) jobs/tasks

Configuration Topic	Activity Monitor
Where to Control Data Retention or Size	<p>You control Activity Monitor storage usage by configuring the number of days or hours of data to retain. Older data is purged.</p> <p>To configure data retention in the Cloudera Manager Administration Console:</p> <ol style="list-style-type: none"> 1. Go the Cloudera Management Service. 2. Click the Configuration tab. 3. Select ScopeActivity Monitor or Cloudera Management Service (Service-Wide). 4. Select CategoryMain. 5. Locate the following properties or search for them by typing the property name in the Search box: <p>Purge Activities Data at This Age</p> <p>In Activity Monitor, purge data about MapReduce jobs and aggregate activities when the data reaches this age in hours. By default, Activity Monitor keeps data about activities for 336 hours (14 days).</p> <p>Purge Attempts Data at This Age</p> <p>In the Activity Monitor, purge data about MapReduce attempts when the data reaches this age in hours. Because attempt data can consume large amounts of database space, you might want to purge it more frequently than activity data. By default, Activity Monitor keeps data about attempts for 336 hours (14 days).</p> <p>Purge MapReduce Service Data at This Age</p> <p>The number of hours of past service-level data to keep in the Activity Monitor database, such as total slots running. The default is to keep data for 336 hours (14 days).</p> 6. Enter a Reason for change, and then click Save Changes to commit the changes.
Sizing, Planning, and Best Practices	<p>The Activity Monitor only monitors MapReduce jobs, and does not monitor YARN applications.</p> <p>If you no longer use MapReduce (MRv1) in your cluster, the Activity Monitor is not required.</p> <p>The amount of storage space needed for 14 days worth of MapReduce activities can vary greatly and directly depends on the size of your cluster and the level of activity that uses MapReduce. It might be necessary to adjust and readjust the amount of storage as you determine the "stable state" and "burst state" of the MapReduce activity in your cluster.</p> <p>For example, consider the following test cluster and usage:</p> <ul style="list-style-type: none"> • A simulated 1000-host cluster, each host with 32 slots • MapReduce jobs with 200 attempts (tasks) per activity (job) <p>Sizing observations for this cluster:</p> <ul style="list-style-type: none"> • Each attempt takes 10 minutes to complete. • This usage results in roughly 20 thousand jobs a day with approximately 5 million total attempts. • For a retention period of 7 days, this Activity Monitor database required 200 GB.

Table 18: Cloudera Management Service - Service Monitor Configuration

Configuration Topic	Service Monitor Configuration
Default Storage Location	/var/lib/cloudera-service-monitor/ on the host where the Service Monitor role is configured to run.
Storage Configuration Defaults / Minimum / Maximum	<ul style="list-style-type: none"> • 10 GiB Services Time Series Storage • 1 GiB Impala Query Storage • 1 GiB YARN Application Storage <p>Total: ~12 GiB Minimum (No Maximum)</p>

Configuration Topic	Service Monitor Configuration
Where to Control Data Retention or Size	<p>Service Monitor data growth is controlled by configuring the maximum amount of storage space it can use.</p> <p>To configure data retention in Cloudera Manager Administration Console:</p> <ol style="list-style-type: none"> 1. Go the Cloudera Management Service. 2. Click the Configuration tab. 3. Select Scope Service Monitor or Cloudera Management Service (Service-Wide) . 4. Select Category Main . 5. Locate the <i>propertyName</i> property or search for it by typing its name in the Search box. <p>Time-Series Storage</p> <p>The approximate amount of disk space dedicated to storing time series and health data. When the store has reached its maximum size, it deletes older data to make room for newer data. The disk usage is approximate because the store only begins deleting data when it reaches the limit.</p> <p>Note that Cloudera Manager stores time-series data at a number of different data granularities, and these granularities have different effective retention periods. The Service Monitor stores metric data not only as raw data points but also as ten-minute, hourly, six-hourly, daily, and weekly summary data points. Raw data consumes the bulk of the allocated storage space and weekly summaries consume the least. Raw data is retained for the shortest amount of time while weekly summary points are unlikely to ever be deleted.</p> <p>Select Cloudera Management ServiceCharts Library tab in Cloudera Manager for information about how space is consumed within the Service Monitor. These pre-built charts also show information about the amount of data retained and time window covered by each data granularity.</p> <p>Impala Storage</p> <p>The approximate amount of disk space dedicated to storing Impala query data. When the store reaches its maximum size, it deletes older data to make room for newer queries. The disk usage is approximate because the store only begins deleting data when it reaches the limit.</p> <p>YARN Storage</p> <p>The approximate amount of disk space dedicated to storing YARN application data. When the store reaches its maximum size, it deletes older data to make room for newer applications. The disk usage is approximate because Cloudera Manager only begins deleting data when it reaches the limit.</p> <ol style="list-style-type: none"> 6. Enter a Reason for change, and then click Save Changes to commit the changes.
Sizing, Planning, and Best Practices	<p>The Service Monitor gathers metrics about configured roles and services in your cluster and also runs active health tests. These health tests run regardless of idle and use periods, because they are always relevant. The Service Monitor gathers metrics and health test results regardless of the level of activity in the cluster. This data continues to grow, even in an idle cluster.</p>

Table 19: Cloudera Management Service - Host Monitor

Configuration Topic	Host Monitor Configuration
Default Storage Location	/var/lib/cloudera-host-monitor/ on the host where the Host Monitor role is configured to run.
Storage Configuration Defaults / Minimum/ Maximum	Default (and minimum): 10 GiB Host Time Series Storage

Configuration Topic	Host Monitor Configuration
Where to Control Data Retention or Size	<p>Host Monitor data growth is controlled by configuring the maximum amount of storage space it can use.</p> <p>See <i>Data Storage for Monitoring Data</i>.</p> <p>To configure these data retention configuration properties in the Cloudera Manager Administration Console:</p> <ol style="list-style-type: none"> 1. Go the Cloudera Management Service. 2. Click the Configuration tab. 3. Select Scope Host Monitor or Cloudera Management Service (Service-Wide). 4. Select Category Main . 5. Locate each property or search for it by typing its name in the Search box. <p>Time-Series Storage</p> <p>The approximate amount of disk space dedicated to storing time series and health data. When the store reaches its maximum size, it deletes older data to make room for newer data. The disk usage is approximate because the store only begins deleting data when it reaches the limit.</p> <p>Note that Cloudera Manager stores time-series data at a number of different data granularities, and these granularities have different effective retention periods. Host Monitor stores metric data not only as raw data points but also as summaries of ten minute, one hour, six hour, one day, and one week increments. Raw data consumes the bulk of the allocated storage space and weekly summaries consume the least. Raw data is retained for the shortest amount of time, while weekly summary points are unlikely to ever be deleted.</p> <p>See the Cloudera Management Service Charts Library tab in Cloudera Manager for information on how space is consumed within the Host Monitor. These pre-built charts also show information about the amount of data retained and the time window covered by each data granularity.</p> <ol style="list-style-type: none"> 6. Enter a Reason for change, and then click Save Changes to commit the changes.
Sizing, Planning and Best Practices	<p>The Host Monitor gathers metrics about host-level items of interest (for example: disk space usage, RAM, CPU usage, swapping, etc) and also informs host health tests. The Host Monitor gathers metrics and health test results regardless of the level of activity in the cluster. This data continues to grow fairly linearly, even in an idle cluster.</p>

Table 20: Cloudera Management Service - Event Server

Configuration Topic	Event Server Configuration
Default Storage Location	/var/lib/cloudera-scm-eventserver/ on the host where the Event Server role is configured to run.
Storage Configuration Defaults	5,000,000 events retained
Where to Control Data Retention or Minimum /Maximum	<p>The amount of storage space the Event Server uses is influenced by configuring how many discrete events it can retain.</p> <p>To configure data retention in Cloudera Manager Administration Console,</p> <ol style="list-style-type: none"> 1. Go the Cloudera Management Service. 2. Click the Configuration tab. 3. Select Scope Event Server or Cloudera Management Service (Service-Wide). 4. Select CategoryMain. 5. Edit the following property: Maximum Number of Events in the Event Server Store <p>The maximum size of the Event Server store, in events. When this size is exceeded, events are deleted starting with the oldest first until the size of the store is below this threshold</p> <ol style="list-style-type: none"> 6. Enter a Reason for change, and then click Save Changes to commit the changes.


Configuration Topic	Event Server Configuration
Sizing, Planning, and Best Practices	<p>The Event Server is a managed Lucene index that collects relevant events that happen within your cluster, such as results of health tests, log events that are created when a log entry matches a set of rules for identifying messages of interest and makes them available for searching, filtering and additional action. You can view and filter events on the Diagnostics Events tab of the Cloudera Manager Administration Console. You can also poll this data using the Cloudera Manager API.</p> <p> Note: The Cloudera Management Service role Alert Publisher sources all the content for its work by regularly polling the Event Server for entries that are marked to be sent out using SNMP or SMTP(S). The Alert Publisher is not discussed because it has no noteworthy storage requirements of its own.</p>

Table 21: Cloudera Management Service - Reports Manager

Configuration Topic	Reports Manager Configuration
Default Storage Location	<p>RDBMS:</p> <p>Any Supported RDBMS.</p> <p>Disk:</p> <p>/var/lib/cloudera-scm-headlamp/ on the host where the Reports Manager role is configured to run.</p>
Storage Configuration Defaults	<p>RDBMS:</p> <p>There are no configurable parameters to directly control the size of this data set.</p> <p>Disk:</p> <p>There are no configurable parameters to directly control the size of this data set. The storage utilization depends not only on the size of the HDFS fsimage, but also on the HDFS file path complexity. Longer file paths contribute to more space utilization.</p>
Where to Control Data Retention or Minimum / Maximum	<p>The Reports Manager uses space in two main locations: on the Reports Manager host and on its supporting database. Cloudera recommends that the database be on a separate host from the Reports Manager host for process isolation and performance.</p>
Sizing, Planning, and Best Practices	<p>Reports Manager downloads the fsimage from the NameNode (every 60 minutes by default) and stores it locally to perform operations against, including indexing the HDFS filesystem structure. More files and directories results in a larger fsimage, which consumes more disk space.</p> <p>Reports Manager has no control over the size of the fsimage. If your total HDFS usage trends upward notably or you add excessively long paths in HDFS, it might be necessary to revisit and adjust the amount of local storage allocated to the Reports Manager. Periodically monitor, review, and adjust the local storage allocation.</p>

Table 22: Cloudera Navigator - Navigator Audit Server

Configuration Topic	Navigator Audit Server Configuration
Default Storage Location	Any Supported RDBMS.
Storage Configuration Defaults	Default: 90 Days retention

Configuration Topic	Navigator Audit Server Configuration
Where to Control Data Retention or Min/Max	<p>Navigator Audit Server storage usage is controlled by configuring how many days of data it can retain. Any older data is purged.</p> <p>To configure data retention in the Cloudera Manager Administration Console:</p> <ol style="list-style-type: none"> 1. Go the Cloudera Management Service. 2. Click the Configuration tab. 3. Select Scope Navigator Audit Server or Cloudera Management Service (Service-Wide). 4. Select CategoryMain. 5. Locate the Navigator Audit Server Data Expiration Period property or search for it by typing its name in the Search box. <p>Navigator Audit Server Data Expiration Period</p> <p>In Navigator Audit Server, purge audit data of various auditable services when the data reaches this age in days. By default, Navigator Audit Server keeps data about audits for 90 days.</p> <ol style="list-style-type: none"> 6. Click Save Changes to commit the changes.
Sizing, Planning, and Best Practices	<p>The size of the Navigator Audit Server database directly depends on the number of audit events the cluster's audited services generate. Normally the volume of HDFS audits exceeds the volume of other audits (all other components like MRv1, Hive and Impala read from HDFS, which generates additional audit events).</p> <p>The average size of a discrete HDFS audit event is ~1 KB. For a busy cluster of 50 hosts with ~100K audit events generated per hour, the Navigator Audit Server database would consume ~2.5 GB per day. To retain 90 days of audits at that level, plan for a database size of around 250 GB. If other configured cluster services generate roughly the same amount of data as the HDFS audits, plan for the Navigator Audit Server database to require around 500 GB of storage for 90 days of data.</p> <p>Notes:</p> <ul style="list-style-type: none"> • Individual Hive and Impala queries themselves can be very large. Since the query itself is part of an audit event, such audit events consume space in proportion to the length of the query. • The amount of space required increases as activity on the cluster increases. In some cases, Navigator Audit Server databases can exceed 1 TB for 90 days of audit events. Benchmark your cluster periodically and adjust accordingly. <p>To map Cloudera Navigator versions to Cloudera Manager versions, see <i>Product Compatibility Matrix for Cloudera Navigator</i>.</p>

Table 23: Cloudera Navigator - Navigator Metadata Server

Configuration Topic	Navigator Metadata Server Configuration
Default Storage Location	<p>RDBMS:</p> <p>Any Supported RDBMS.</p> <p>Disk:</p> <p>/var/lib/cloudera-scm-navigator/ on the host where the Navigator Metadata Server role is configured to run.</p>
Storage Configuration Defaults	<p>RDBMS:</p> <p>There are no exposed defaults or configurations to directly cull or purge the size of this data set.</p> <p>Disk:</p> <p>There are no configuration defaults to influence the size of this location. You can change the location itself with the Navigator Metadata Server Storage Dir property. The size of the data in this location depends on the amount of metadata in the system (HDFS fsimage size, Hive Metastore size) and activity on the system (the number of MapReduce Jobs run, Hive queries executed, etc).</p>

Configuration Topic	Navigator Metadata Server Configuration
Where to Control Data Retention or Min/Max	<p>RDBMS:</p> <p>The Navigator Metadata Server database should be carefully tuned to support large volumes of metadata.</p> <p>Disk:</p> <p>The Navigator Metadata Server index (an embedded Solr instance) can consume lots of disk space at the location specified for the Navigator Metadata Server Storage Dir property. Ongoing maintenance tasks include purging metadata from the system.</p>
Sizing, Planning, and Best Practices	<p>Memory:</p> <p><i>See Navigator Metadata Server Tuning.</i></p> <p>RDBMS:</p> <p>The database is used to store policies and authorization data. The dataset is small, but this database is also used during a Solr schema upgrade, where Solr documents are extracted and inserted again in Solr. This has same space requirements as above use case, but the space is only used temporarily during product upgrades.</p> <p>Use the product compatibility matrix to map Cloudera Navigator and Cloudera Manager versions.</p> <p>Disk:</p> <p>This filesystem location contains all the metadata that is extracted from managed clusters. The data is stored in Solr, so this is the location where Solr stores its index and documents. Depending on the size of the cluster, this data can occupy tens of gigabytes. A guideline is to look at the size of HDFS fsimage and allocate two to three times that size as the initial size. The data here is incremental and continues to grow as activity is performed on the cluster. The rate of growth can be on order of tens of megabytes per day.</p>

General Performance Notes

When possible:

- For entities that use an RDBMS, install the database on a separate host from the service, and consolidate roles that use databases on as few servers as possible.
- Provide a dedicated spindle to the RDBMS or datastore data directory to avoid disk contention with other read/write activity.

Cluster Lifecycle Management with Cloudera Manager

Cloudera Manager clusters that use parcels to provide Cloudera Runtime and other components require adequate disk space in the following locations:

Table 24: Parcel Lifecycle Management

Parcel Lifecycle Path (default)	Notes
Local Parcel Repository Path (/opt/cloudera/parcel-repo)	<p>This path exists only on the host where Cloudera Manager Server (cloudera-scm-server) runs. The Cloudera Manager Server stages all new parcels in this location as it fetches them from any external repositories. Cloudera Manager Agents are then instructed to fetch the parcels from this location when the administrator distributes the parcel using the Cloudera Manager Administration Console or the Cloudera Manager API.</p> <p>Sizing and Planning</p> <p>The default location is /opt/cloudera/parcel-repo but you can configure another local filesystem location on the host where Cloudera Manager Server runs.</p> <p>Provide sufficient space to hold all the parcels you download from all configured Remote Parcel Repository URLs. Cloudera Manager deployments that manage multiple clusters store all applicable parcels for all clusters.</p> <p>Parcels are provided for each operating system, so be aware that heterogeneous clusters (distinct operating systems represented in the cluster) require more space than clusters with homogeneous operating systems.</p> <p>For example, a cluster with both RHEL6.x and 7.x hosts must hold -el6 and -el7 parcels in the Local Parcel Repository Path, which requires twice the amount of space.</p> <p>Lifecycle Management and Best Practices</p> <p>Delete any parcels that are no longer in use from the Cloudera Manager Administration Console, (never delete them manually from the command line) to recover disk space in the Local Parcel Repository Path and simultaneously across all managed cluster hosts which hold the parcel.</p> <p>Backup Considerations</p> <p>Perform regular backups of this path, and consider it a non-optional accessory to backing up Cloudera Manager Server. If you migrate Cloudera Manager Server to a new host or restore it from a backup (for example, after a hardware failure), recover the full content of this path to the new host, in the /opt/cloudera/parcel-repo directory before starting any cloudera-scm-agent or cloudera-scm-server processes.</p>
Parcel Cache (/opt/cloudera/parcel-cache)	<p>Managed Hosts running a Cloudera Manager Agent stage distributed parcels into this path (as .parcel files, unextracted). Do not manually manipulate this directory or its files.</p> <p>Sizing and Planning</p> <p>Provide sufficient space per-host to hold all the parcels you distribute to each host.</p> <p>You can configure Cloudera Manager to remove these cached .parcel files after they are extracted and placed in /opt/cloudera/parcels/. It is not mandatory to keep these temporary files but keeping them avoids the need to transfer the .parcel file from the Cloudera Manager Server repository should you need to extract the parcel again for any reason.</p> <p>To configure this behavior in the Cloudera Manager Administration Console, select AdministrationSettingsParcelsRetain Downloaded Parcel Files</p>

Parcel Lifecycle Path (default)	Notes
Host Parcel Directory (/opt/cloudera/parcels)	<p>Managed cluster hosts running a Cloudera Manager Agent extract parcels from the /opt/cloudera/parcel-cache directory into this path upon parcel activation. Many critical system symlinks point to files in this path and you should never manually manipulate its contents.</p> <p>Sizing and Planning</p> <p>Provide sufficient space on each host to hold all the parcels you distribute to each host. Be aware that the typical Runtime or CDH parcel size is approximately 2 GB per parcel, and some third party parcels can exceed 3 GB. If you maintain various versions of parcels staged before and after upgrading, be aware of the disk space implications.</p> <p>You can configure Cloudera Manager to automatically remove older parcels when they are no longer in use. As an administrator you can always manually delete parcel versions not in use, but configuring these settings can handle the deletion automatically, in case you forget.</p> <p>To configure this behavior in the Cloudera Manager Administration Console, select AdministrationSettingsParcels and configure the following property:</p> <p>Automatically Remove Old Parcels</p> <p>This parameter controls whether parcels for old versions of an activated product should be removed from a cluster when they are no longer in use.</p> <p>The default value is Disabled.</p> <p>Number of Old Parcel Versions to Retain</p> <p>If you enable Automatically Remove Old Parcels, this setting specifies the number of old parcels to keep. Any old parcels beyond this value are removed. If this property is set to zero, no old parcels are retained.</p> <p>The default value is 3.</p>

Table 25: Management Service Lifecycle - Space Reclamation Tasks

Task	Description
Activity Monitor (One-time)	<p>The Activity Monitor only works against a MapReduce (MR1) service, not YARN. So if your deployment has fully migrated to YARN and no longer uses a MapReduce (MR1) service, your Activity Monitor database is no longer growing. If you have waited longer than the default Activity Monitor retention period (14 days) to address this point, then the Activity Monitor has already purged it all for you and your database is mostly empty. If your deployment meets these conditions, consider cleaning up by dropping the Activity Monitor database (again, only when you are satisfied that you no longer need the data or have confirmed that it is no longer in use) and the Activity Monitor role.</p>
Service Monitor and Host Monitor (One-time)	<p>For those who used Cloudera Manager version 4.x and have now upgraded to version 5.x: The Service Monitor and Host Monitor were migrated from their previously-configured RDBMS into a dedicated time series store used solely by each of these roles respectively. After this happens, there is still legacy database connection information in the configuration for these roles. This was used to allow for the initial migration but is no longer being used for any active work.</p> <p>After the above migration has taken place, the RDBMS databases previously used by the Service Monitor and Host Monitor are no longer used. Space occupied by these databases is now recoverable. If appropriate in your environment (and you are satisfied that you have long-term backups or do not need the data on disk any longer), you can drop those databases.</p>
Ongoing Space Reclamation	<p>Cloudera Management Services are automatically rolling up, purging or otherwise consolidating aged data for you in the background. Configure retention and purging limits per-role to control how and when this occurs. These configurations are discussed per-entity above. Adjust the default configurations to meet your space limitations or retention needs.</p>

Log File Storage Space

All cluster hosts write out separate log files for each role instance assigned to the host. Cluster administrators can monitor and manage the disk space used by these roles and configure log rotation to prevent log files from consuming too much disk space.

Configure Network Names

You must configure each host in the cluster to ensure that all members can communicate with each other.

About this task



Important: Cloudera Runtime requires IPv4. IPv6 is not supported.



Tip: When bonding, use the bond0 IP address as it represents all aggregated links.

Procedure

1. Set the hostname to a unique name (not localhost).

```
sudo hostnamectl set-hostname foo-1.example.com
```

2. Edit /etc/hosts with the IP address and fully qualified domain name (FQDN) of each host in the cluster. You can add the unqualified name as well.

```
1.1.1.1  foo-1.example.com  foo-1
2.2.2.2  foo-2.example.com  foo-2
3.3.3.3  foo-3.example.com  foo-3
4.4.4.4  foo-4.example.com  foo-4
```



Important:

- The canonical name of each host in /etc/hosts must be the FQDN (for example myhost-1.example.com), not the unqualified hostname (for example myhost-1). The canonical name is the first entry after the IP address.
- Do not use aliases, either in /etc/hosts or in configuring DNS.
- Unqualified hostnames (short names) must be unique in a Cloudera Manager instance. For example, you cannot have both *host01.example.com* and *host01.standby.example.com* managed by the same Cloudera Manager Server.

3. Edit /etc/sysconfig/network with the FQDN of this host only:

```
HOSTNAME=foo-1.example.com
```

4. Verify that each host consistently identifies to the network:

- a) Run `uname -a` and check that the hostname matches the output of the `hostname` command.
- b) Run `/sbin/ifconfig` and note the value of `inet addr` in the `eth0` (or `bond0`) entry, for example:

```
eth0      Link encap:Ethernet  HWaddr 00:0C:29:A4:E8:97
          inet addr:172.29.82.176  Bcast:172.29.87.255  Mask:255.255.2
48.0
...
```

- c) Run `host -v -t A $(hostname)` and verify that the output matches the `hostname` command. The IP address should be the same as reported by `ifconfig` for `eth0` (or `bond0`):

```
Trying "foo-1.example.com"
...
;; ANSWER SECTION:
foo-1.example.com. 60 IN A 172.29.82.176
```

Disabling the Firewall

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

Procedure

1. For iptables, save the existing rule set:

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

2. Disable the firewall:

- RHEL 7 compatible:

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

- SLES:

```
sudo chkconfig SuSEfirewall2_setup off
sudo chkconfig SuSEfirewall2_init off
sudo rcSuSEfirewall2 stop
```

- Ubuntu:

```
sudo service ufw stop
```

Setting SELinux Mode

Security-Enhanced Linux (SELinux) allows you to set access control through policies. If you are having trouble deploying Runtime or CDH with your policies, set SELinux in permissive mode on each host before you deploy Runtime or CDH on your cluster.

About this task



Note: Cloudera Enterprise, with the exception of Cloudera Navigator Encrypt, is supported on platforms with Security-Enhanced Linux (SELinux) enabled and in enforcing mode. Cloudera is not responsible for SELinux policy development, support, or enforcement. If you experience issues running Cloudera software with SELinux enabled, contact your OS provider for assistance.

If you are using SELinux in enforcing mode, Cloudera Support can request that you disable SELinux or change the mode to permissive to rule out SELinux as a factor when investigating reported issues.

Procedure

1. Check the SELinux state:

```
getenforce
```

2. If the output is either Permissive or Disabled, you can skip this task and continue on to disabling the firewall. If the output is enforcing, continue to the next step.
3. Open the /etc/selinux/config file (in some systems, the /etc/sysconfig/selinux file).
4. Change the line SELINUX=enforcing to SELINUX=permissive.
5. Save and close the file.

- Restart your system or run the following command to disable SELinux immediately:

```
setenforce 0
```

After you have installed and deployed Runtime or CDH, you can re-enable SELinux by changing SELINUX=permissive back to SELINUX=enforcing in `/etc/selinux/config` (or `/etc/sysconfig/selinux`), and then running the following command to immediately switch to enforcing mode:

```
setenforce 1
```

If you are having trouble getting Cloudera Software working with SELinux, contact your OS vendor for support. Cloudera is not responsible for developing or supporting SELinux policies.

Enable an NTP Service

Runtime requires that you configure a Network Time Protocol (NTP) service on each machine in your cluster. Most operating systems include the `ntpd` service for time synchronization.

About this task

RHEL 7 compatible operating systems use `chronyd` by default instead of `ntpd`. If `chronyd` is running (on any OS), Cloudera Manager uses it to determine whether the host clock is synchronized. Otherwise, Cloudera Manager uses `ntpd`.



Note: If you are using `ntpd` to synchronize your host clocks, but `chronyd` is also running, Cloudera Manager relies on `chronyd` to verify time synchronization, even if it is not synchronizing properly. This can result in Cloudera Manager reporting clock offset errors, even though the time is correct.

To fix this, either configure and use `chronyd` or disable it and remove it from the hosts.

To use `ntpd` for time synchronization:

Before you begin

Procedure

- Install the `ntp` package:

- RHEL compatible:

```
yum install ntp
```

- SLES:

```
zypper install ntp
```

- Ubuntu:

```
apt-get install ntp
```

- Edit the `/etc/ntp.conf` file to add NTP servers, as in the following example:

```
server 0.pool.ntp.org
server 1.pool.ntp.org
server 2.pool.ntp.org
```

3. Start the ntpd service:

- RHEL 7 Compatible:

```
sudo systemctl start ntpd
```

- SLES, Ubuntu:

```
sudo service ntpd start
```

4. Configure the ntpd service to run at boot:

- RHEL 7 Compatible:

```
sudo systemctl enable ntpd
```

- SLES, Ubuntu:

```
chkconfig ntpd on
```

5. Synchronize the system clock to the NTP server:

```
ntpdate -u <ntp_server>
```

6. Synchronize the hardware clock to the system clock:

```
hwclock --systohc
```

Impala Requirements

To perform as expected, Impala depends on the availability of the software, hardware, and configurations described in the following sections.

Product Compatibility Matrix

The ultimate source of truth about compatibility between various versions of Cloudera Runtime, Cloudera Manager, and various Runtime components is the Product Compatibility Matrix.

Supported Operating Systems

The relevant supported operating systems and versions for Impala are the same as for the corresponding Cloudera Runtime platforms. For details, see the *Operating System Requirements* topic.

Hive Metastore and Related Configuration

Impala can interoperate with data stored in Hive, and uses the same infrastructure as Hive for tracking metadata about schema objects such as tables and columns. The following components are prerequisites for Impala:

To install the metastore:

1. Install a MySQL or PostgreSQL database. Start the database if it is not started after installation.
2. Download the MySQL Connector or the PostgreSQL connector and place it in the `/usr/share/java/` directory.
3. Use the appropriate command line tool for your database to create the metastore database.
4. Use the appropriate command line tool for your database to grant privileges for the metastore database to the hive user.
5. Modify `hive-site.xml` to include information matching your particular database: its URL, username, and password. You will copy the `hive-site.xml` file to the Impala Configuration Directory later in the Impala installation process.

Java Dependencies

Although Impala is primarily written in C++, it does use Java to communicate with various Hadoop components:

- The officially supported JVMs for Impala are the OpenJDK JVM and Oracle JVM. Other JVMs might cause issues, typically resulting in a failure at `impalad` startup. In particular, the JamVM used by default on certain levels of Ubuntu systems can cause `impalad` to fail to start.
- Internally, the `impalad` daemon relies on the `JAVA_HOME` environment variable to locate the system Java libraries. Make sure the `impalad` service is not run from an environment with an incorrect setting for this variable.
- All Java dependencies are packaged in the `impala-dependencies.jar` file, which is located at `/usr/lib/impala/lib/`. These map to everything that is built under `fe/target/dependency`.

Networking Configuration Requirements

As part of ensuring best performance, Impala attempts to complete tasks on local data, as opposed to using network connections to work with remote data. To support this goal, Impala matches the hostname provided to each Impala daemon with the IP address of each DataNode by resolving the hostname flag to an IP address. For Impala to work with local data, use a single IP interface for the DataNode and the Impala daemon on each machine. Ensure that the Impala daemon's hostname flag resolves to the IP address of the DataNode. For single-homed machines, this is usually automatic, but for multi-homed machines, ensure that the Impala daemon's hostname resolves to the correct interface. Impala tries to detect the correct hostname at start-up, and prints the derived hostname at the start of the log in a message of the form:

```
Using hostname: impala-daemon-1.example.com
```

In the majority of cases, this automatic detection works correctly. If you need to explicitly set the hostname, do so by setting the `--hostname` flag.

Hardware Requirements

The memory allocation should be consistent across Impala executor nodes. A single Impala executor with a lower memory limit than the rest can easily become a bottleneck and lead to suboptimal performance.

This guideline does not apply to coordinator-only nodes.

Hardware Requirements for Optimal Join Performance

During join operations, portions of data from each joined table are loaded into memory. Data sets can be very large, so ensure your hardware has sufficient memory to accommodate the joins you anticipate completing.

While requirements vary according to data set size, the following is generally recommended:

- CPU

Impala version 2.2 and higher uses the SSE3 instruction set, which is included in newer processors.



Note: This required level of processor is the same as in Impala version 1.x. The Impala 2.0 and 2.1 releases had a stricter requirement for the SSE4.1 instruction set, which has now been relaxed.

- Memory

128 GB or more recommended, ideally 256 GB or more. If the intermediate results during query processing on a particular node exceed the amount of memory available to Impala on that node, the query writes temporary work data to disk, which can lead to long query times. Note that because the work is parallelized, and intermediate results for aggregate queries are typically smaller than the original data, Impala can query and join tables that are much larger than the memory available on an individual node.

- JVM Heap Size for Catalog Server

4 GB or more recommended, ideally 8 GB or more, to accommodate the maximum numbers of tables, partitions, and data files you are planning to use with Impala.

- Storage

DataNodes with 12 or more disks each. I/O speeds are often the limiting factor for disk performance with Impala. Ensure that you have sufficient disk space to store the data Impala will be querying.

User Account Requirements

For user account requirements, see the topic User Account Requirements in the Impala documentation.

Runtime Cluster Hosts and Role Assignments

Cluster hosts can be broadly described as master hosts, utility hosts, gateway hosts, or worker hosts.

- Master hosts run Hadoop master processes such as the HDFS NameNode and YARN Resource Manager.
- Utility hosts run other cluster processes that are not master processes such as Cloudera Manager and the Hive Metastore.
- Gateway hosts are client access points for launching jobs in the cluster. The number of gateway hosts required varies depending on the type and size of the workloads.
- Worker hosts primarily run DataNodes and other distributed processes such as Impalad.



Important: Cloudera recommends that you always enable high availability when Runtime is used in a production environment.

The following tables describe the recommended role allocations for different cluster sizes. Note that these configurations take into account services dependencies that might not be obvious. For example, running Atlas or Ranger requires also running HBase, Kafka, Solr, and ZooKeeper. For details see [Service Dependencies in Cloudera Manager](#).

3 - 10 Worker Hosts without High Availability

Master Hosts	Utility Hosts	Gateway Hosts	Worker Hosts
Master Host 1: <ul style="list-style-type: none"> • NameNode • YARN ResourceManager • JobHistory Server • ZooKeeper • Kudu master • Spark History Server • HBase master • Schema Registry 	One host for all Utility and Gateway roles: <ul style="list-style-type: none"> • Secondary NameNode • Cloudera Manager • Cloudera Manager Management Service • Cruise Control • Hive Metastore • HiveServer2 • Impala Catalog Server • Impala StateStore • Hue • Oozie • Gateway configuration • HBase backup master • Ranger Admin, Tagsync, Usersync servers • Atlas server • Solr server (CDP-INFRA-SOLR instance to support Atlas) • Streams Messaging Manager • Streams Replication Manager Service • ZooKeeper 		3 - 10 Worker Hosts: <ul style="list-style-type: none"> • DataNode • NodeManager • Impalad • Kudu tablet server • Kafka Broker • Kafka Connect • HBase RegionServer • Solr server (For Cloudera Search) • Streams Replication Manager Driver • ZooKeeper (Recommend 3 servers total)

3 - 20 Worker Hosts with High Availability

Master Hosts	Utility Hosts	Gateway Hosts	Worker Hosts
<p>Master Host 1:</p> <ul style="list-style-type: none"> NameNode JournalNode FailoverController YARN ResourceManager ZooKeeper JobHistory Server Kudu master HBase master Schema Registry <p>Master Host 2:</p> <ul style="list-style-type: none"> NameNode JournalNode FailoverController YARN ResourceManager ZooKeeper Kudu master HBase master Schema Registry <p>Master Host 3:</p> <ul style="list-style-type: none"> Kudu master (Kudu requires an odd number of masters for HA.) Spark History Server JournalNode (requires dedicated disk) ZooKeeper 	<p>Utility Host 1:</p> <ul style="list-style-type: none"> Cloudera Manager Cloudera Manager Management Service Cruise Control Hive Metastore Impala Catalog Server Impala StateStore Oozie Ranger Admin, Tagsync, Usersync servers Atlas server Solr server (CDP-INFRA-SOLR instance to support Atlas) Streams Messaging Manager Streams Replication Manager Service <p>Utility Host 2:</p> <ul style="list-style-type: none"> Ranger Admin server Atlas server Solr server (CDP-INFRA-SOLR instance to support Atlas) 	<p>One or more Gateway Hosts:</p> <ul style="list-style-type: none"> Hue HiveServer2 Gateway configuration 	<p>3 - 20 Worker Hosts:</p> <ul style="list-style-type: none"> DataNode NodeManager Impalad Kudu tablet server Kafka Broker (Recommend 3 brokers minimum) Kafka Connect HBase RegionServer Solr server (For Cloudera Search, recommend 3 servers minimum) Streams Replication Manager Driver

20 - 80 Worker Hosts with High Availability

Master Hosts	Utility Hosts	Gateway Hosts	Worker Hosts
<p>Master Host 1:</p> <ul style="list-style-type: none"> NameNode JournalNode FailoverController YARN ResourceManager ZooKeeper Kudu master HBase master Schema Registry <p>Master Host 2:</p> <ul style="list-style-type: none"> NameNode JournalNode FailoverController YARN ResourceManager ZooKeeper Kudu master HBase master Schema Registry <p>Master Host 3:</p> <ul style="list-style-type: none"> ZooKeeper JournalNode JobHistory Server Spark History Server Kudu master HBase master 	<p>Utility Host 1:</p> <ul style="list-style-type: none"> Cloudera Manager Cruise Control Ranger Admin server Atlas server Solr server (CDP-INFRA-SOLR instance to support Atlas) Streams Messaging Manager Streams Replication Manager Service <p>Utility Host 2:</p> <ul style="list-style-type: none"> Cloudera Manager Management Service Hive Metastore Impala Catalog Server Oozie Ranger Admin, Tagsync, Usersync servers Atlas server Solr server (CDP-INFRA-SOLR instance to support Atlas) 	<p>One or more Gateway Hosts:</p> <ul style="list-style-type: none"> Hue HiveServer2 Gateway configuration 	<p>20 - 80 Worker Hosts:</p> <ul style="list-style-type: none"> DataNode NodeManager Impalad Kudu tablet server Kafka Broker (Recommend 3 brokers minimum) Kafka Connect HBase RegionServer Solr server (For Cloudera Search, recommend 3 servers minimum) Streams Replication Manager Driver

80 - 200 Worker Hosts with High Availability

Master Hosts	Utility Hosts	Gateway Hosts	Worker Hosts
<p>Master Host 1:</p> <ul style="list-style-type: none"> NameNode JournalNode FailoverController YARN ResourceManager ZooKeeper Kudu master HBase master Schema Registry <p>Master Host 2:</p> <ul style="list-style-type: none"> NameNode JournalNode FailoverController YARN ResourceManager ZooKeeper Kudu master HBase master Schema Registry <p>Master Host 3:</p> <ul style="list-style-type: none"> ZooKeeper JournalNode JobHistory Server Spark History Server Kudu master HBase master 	<p>Utility Host 1:</p> <ul style="list-style-type: none"> Cloudera Manager Cruise Control Streams Messaging Manager Streams Replication Manager Service <p>Utility Host 2:</p> <ul style="list-style-type: none"> Hive Metastore Impala Catalog Server Impala StateStore Oozie <p>Utility Host 3:</p> <ul style="list-style-type: none"> Activity Monitor <p>Utility Host 4:</p> <ul style="list-style-type: none"> Host Monitor <p>Utility Host 5:</p> <ul style="list-style-type: none"> Ranger Admin, Tagsync, Usersync servers Atlas server Solr server <p>Utility Host 6:</p> <ul style="list-style-type: none"> Ranger Admin server Atlas server Solr server <p>Utility Host 7:</p> <ul style="list-style-type: none"> Reports Manager <p>Utility Host 8:</p> <ul style="list-style-type: none"> Service Monitor 	<p>One or more Gateway Hosts:</p> <ul style="list-style-type: none"> Hue HiveServer2 Gateway configuration 	<p>80 - 200 Worker Hosts:</p> <ul style="list-style-type: none"> DataNode NodeManager Impalad Kudu tablet server (Recommend 100 tablet servers maximum) Kafka Broker (Recommend 3 brokers minimum) Kafka Connect HBase RegionServer Solr server (For Cloudera Search, recommend 3 servers minimum) Streams Replication Manager Driver

200 - 500 Worker Hosts with High Availability

Master Hosts	Utility Hosts	Gateway Hosts	Worker Hosts
<p>Master Host 1:</p> <ul style="list-style-type: none"> NameNode JournalNode FailoverController ZooKeeper Kudu master HBase master <p>Master Host 2:</p> <ul style="list-style-type: none"> NameNode JournalNode FailoverController ZooKeeper Kudu master HBase master <p>Master Host 3:</p> <ul style="list-style-type: none"> YARN ResourceManager ZooKeeper JournalNode Kudu master HBase master Schema Registry <p>Master Host 4:</p> <ul style="list-style-type: none"> YARN ResourceManager ZooKeeper JournalNode Schema Registry <p>Master Host 5:</p> <ul style="list-style-type: none"> JobHistory Server Spark History Server ZooKeeper JournalNode <p>We recommend no more than three masters for Kudu and HBase.</p>	<p>Utility Host 1:</p> <ul style="list-style-type: none"> Cloudera Manager Cruise Control Streams Messaging Manager Streams Replication Manager Service <p>Utility Host 2:</p> <ul style="list-style-type: none"> Hive Metastore Impala Catalog Server Impala StateStore Oozie <p>Utility Host 3:</p> <ul style="list-style-type: none"> Activity Monitor <p>Utility Host 4:</p> <ul style="list-style-type: none"> Host Monitor <p>Utility Host 5:</p> <ul style="list-style-type: none"> Ranger Admin, Tagsync, Usersync servers Atlas server Solr server (CDP-INFRA-SOLR instance to support Atlas) <p>Utility Host 6:</p> <ul style="list-style-type: none"> Ranger Admin server Atlas server Solr server (CDP-INFRA-SOLR instance to support Atlas) <p>Utility Host 7:</p> <ul style="list-style-type: none"> Reports Manager <p>Utility Host 8:</p> <ul style="list-style-type: none"> Service Monitor 	<p>One or more Gateway Hosts:</p> <ul style="list-style-type: none"> Hue HiveServer2 Gateway configuration 	<p>200 - 500 Worker Hosts:</p> <ul style="list-style-type: none"> DataNode NodeManager Impalad Kudu tablet server (Recommend 100 tablet servers maximum) Kafka Broker (Recommend 3 brokers minimum) Kafka Connect HBase RegionServer Solr server (For Cloudera Search, recommend 3 servers minimum) Streams Replication Manager Driver

500 -1000 Worker Hosts with High Availability

Master Hosts	Utility Hosts	Gateway Hosts	Worker Hosts
Master Host 1: <ul style="list-style-type: none"> NameNode JournalNode FailoverController ZooKeeper Kudu master HBase master Master Host 2: <ul style="list-style-type: none"> NameNode JournalNode FailoverController ZooKeeper Kudu master HBase master Master Host 3: <ul style="list-style-type: none"> YARN ResourceManager ZooKeeper JournalNode Kudu master HBase master Schema Registry Master Host 4: <ul style="list-style-type: none"> YARN ResourceManager ZooKeeper JournalNode Schema Registry Master Host 5: <ul style="list-style-type: none"> JobHistory Server Spark History Server ZooKeeper JournalNode <p>We recommend no more than three masters for Kudu and HBase.</p>	Utility Host 1: <ul style="list-style-type: none"> Cloudera Manager Cruise Control Streams Messaging Manager Streams Replication Manager Service Utility Host 2: <ul style="list-style-type: none"> Hive Metastore Impala Catalog Server Impala StateStore Oozie Utility Host 3: <ul style="list-style-type: none"> Activity Monitor Utility Host 4: <ul style="list-style-type: none"> Host Monitor Utility Host 5: <ul style="list-style-type: none"> Ranger Admin, Tagsync, Usersync servers Atlas server Solr server (CDP-INFRA-SOLR instance to support Atlas) Utility Host 6: <ul style="list-style-type: none"> Ranger Admin server Atlas server Solr server (CDP-INFRA-SOLR instance to support Atlas) Utility Host 7: <ul style="list-style-type: none"> Reports Manager Utility Host 8: <ul style="list-style-type: none"> Service Monitor 	One or more Gateway Hosts: <ul style="list-style-type: none"> Hue HiveServer2 Gateway configuration 	500 - 1000 Worker Hosts: <ul style="list-style-type: none"> DataNode NodeManager Impalad Kudu tablet server (Recommend 100 tablet servers maximum) Kafka Broker (Recommend 3 brokers minimum) Kafka Connect HBase RegionServer Solr server (For Cloudera Search, recommend 3 servers minimum) Streams Replication Manager Driver

Related Information

[Service Dependencies in Cloudera Manager](#)

Allocating Hosts for Key Trustee Server and Key Trustee KMS

If you are enabling data-at-rest encryption for a Cloudera Runtime cluster, Cloudera recommends that you isolate the Key Trustee Server from other enterprise data hub (EDH) services by deploying the Key Trustee Server on dedicated hosts in a separate cluster managed by Cloudera Manager.

Cloudera also recommends deploying Key Trustee KMS on dedicated hosts in the same cluster as the EDH services that require access to Key Trustee Server. This architecture helps users avoid having to restart the Key Trustee Server when restarting a cluster.

For production environments in general, or if you have enabled high availability for HDFS and are using data-at-rest encryption, Cloudera recommends that you enable high availability for Key Trustee Server and Key Trustee KMS.

Configuring Local Package and Parcel Repositories

Cloudera hosts two types of software repositories that you can use to install products such as Cloudera Manager or Cloudera Runtime—parcel repositories and package repositories. These repositories are effective solutions in most cases, but custom installation solutions are sometimes required.

For example, using the Cloudera-hosted software repositories requires client access over the Internet. Typical installations use the latest available software. In some scenarios, these behaviors might not be desirable, such as:

- You need to install older product versions. For example, in a Runtime cluster, all hosts must run the same Runtime version. After completing an initial installation, you may want to add hosts. This could be to increase the size of your cluster to handle larger tasks or to replace older hardware.
- The hosts on which you want to install Cloudera products are not connected to the Internet, so they cannot reach the Cloudera repository (for a parcel installation, only the Cloudera Manager Server needs Internet access, but for a package installation, all cluster hosts require access to the Cloudera repository). Most organizations partition parts of their network from outside access. Isolating network segments improves security, but can add complexity to the installation process.

In both of these cases, using an internal repository allows you to meet the needs of your organization, whether that means installing specific versions of Cloudera software or installing Cloudera software on hosts without Internet access.

Understanding Package Management

Before you configure a custom package management solution in your environment, understand the concepts of package management tools and package repositories.

Package Management Tools

Packages (rpm or deb files) help ensure that installations complete successfully by satisfying package dependencies. When you install a particular package, all other required packages are installed at the same time. For example, `hadoop-0.20-hive` depends on `hadoop-0.20`.

Package management tools, such as `yum` (RHEL), `zypper` (SLES), and `apt-get` (Ubuntu) are tools that can find and install required packages. For example, on a RHEL compatible system, you might run the command `yum install hadoop-0.20-hive`. The `yum` utility informs you that the Hive package requires `hadoop-0.20` and offers to install it for you. `zypper` and `apt-get` provide similar functionality.

Package Repositories

Package management tools rely on package repositories to install software and resolve any dependency requirements. For information on creating an internal repository, see *Configuring a Local Package Repository*.

Repository Configuration Files

Information about package repositories is stored in configuration files, the location of which varies according to the package management tool.

- RHEL compatible (`yum`): `/etc/yum.repos.d`
- SLES (`zypper`): `/etc/zypp/zypper.conf`
- Ubuntu (`apt-get`): `/etc/apt/apt.conf` (Additional repositories are specified using `.list` files in the `/etc/apt/sources.list.d/` directory.)

For example, on a typical CentOS system, you might find:

```
ls -l /etc/yum.repos.d/
total 36
-rw-r--r--. 1 root root 1664 Dec 9 2015 CentOS-Base.repo
-rw-r--r--. 1 root root 1309 Dec 9 2015 CentOS-CR.repo
-rw-r--r--. 1 root root 649 Dec 9 2015 CentOS-Debuginfo.repo
-rw-r--r--. 1 root root 290 Dec 9 2015 CentOS-fasttrack.repo
-rw-r--r--. 1 root root 630 Dec 9 2015 CentOS-Media.repo
-rw-r--r--. 1 root root 1331 Dec 9 2015 CentOS-Sources.repo
-rw-r--r--. 1 root root 1952 Dec 9 2015 CentOS-Vault.repo
```



```
-rw-r--r--. 1 root root 951 Jun 24 2017 epel.repo
-rw-r--r--. 1 root root 1050 Jun 24 2017 epel-testing.repo
```

The .repo files contain pointers to one or more repositories. In the following excerpt from CentOS-Base.repo, there are two repositories defined: one named Base and one named Updates. The mirrorlist parameter points to a website that has a list of places where this repository can be downloaded.

```
[base]
name=CentOS-$releasever - Base
mirrorlist=http://mirrorlist.centos.org/?release=$releasever&arch=$basearch&repo=os&infra=$infra
#baseurl=http://mirror.centos.org/centos/$releasever/os/$basearch/
gpgcheck=1
gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-CentOS-7

#released updates
[updates]
name=CentOS-$releasever - Updates
mirrorlist=http://mirrorlist.centos.org/?release=$releasever&arch=$basearch&repo=updates&infra=$infra
#baseurl=http://mirror.centos.org/centos/$releasever/updates/$basearch/
gpgcheck=1
gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-CentOS-7
```

Listing Repositories

You can list the enabled repositories by running one of the following commands:

- RHEL compatible: yum repolist
- SLES: zypper repos
- Ubuntu: apt-get does not include a command to display sources, but you can determine sources by reviewing the contents of /etc/apt/sources.list and any files contained in /etc/apt/sources.list.d/.

The following shows an example of the output of yum repolist on a CentOS 7 system:

repo id	repo name	st
atus		
base/7/x86_64	CentOS-7 - Base	
9,591		
epel/x86_64	Extra Packages for Enterprise Linux 7 - x86_64	
12,382		
extras/7/x86_64	CentOS-7 - Extras	
392		
updates/7/x86_64	CentOS-7 - Updates	1
,962		
repolist: 24,327		

Configuring a Local Package Repository

You can create a package repository for Cloudera Manager either by hosting an internal web repository or by manually copying the repository files to the Cloudera Manager Server host for distribution to Cloudera Manager Agent hosts.

Creating a Permanent Internal Repository

The following sections describe how to create a permanent internal repository using Apache HTTP Server.

Setting Up a Web Server

To host an internal repository, you must install or use an existing Web server on an internal host that is reachable by the Cloudera Manager host, and then download the repository files to the Web server host.

About this task

The examples in this section use Apache HTTP Server as the Web server. If you already have a Web server in your organization, you can skip to *Downloading and Publishing the Package Repository*.

Procedure

1. Install Apache HTTP Server:

RHEL / CentOS

```
sudo yum install httpd
```

2. Start Apache HTTP Server:

RHEL 7

```
sudo systemctl start httpd
```

Downloading and Publishing the Package Repository

Download the package repository for the product you want to install.

Procedure

1. Download the package repository for the product you want to install:

Cloudera Manager 7

To download the files for a Cloudera Manager release, download the repository tarball for your operating system. Then unpack the tarball, move the files to the web server directory, and modify file permissions. For example:

```
sudo mkdir -p /var/www/html/cloudera-repos/cm7
```

```
wget https://[username]:[password]@archive.cloudera.com/p/cm7/7.2.3/repo-as-tarball/cm7.2.3-redhat7.tar.gz
```

```
tar xvfz cm7.2.3-redhat7.tar.gz -C /var/www/html/cloudera-repos/cm7 --strip-components=1
```

```
sudo chmod -R ugo+rX /var/www/html/cloudera-repos/cm7
```

2. Visit the Repository URL `http://<web_server>/cloudera-repos/` in your browser and verify the files you downloaded are present. If you do not see anything, your Web server may have been configured to not show indexes.

Creating a Temporary Internal Repository

You can quickly create a temporary remote repository to deploy packages on a one-time basis. Cloudera recommends using the same host that runs Cloudera Manager, or a gateway host.

About this task

This example uses Python SimpleHTTPServer as the Web server to host the `/var/www/html` directory, but you can use a different directory.

Procedure

1. Download the repository you need following the instructions in *Downloading and Publishing the Package Repository*.
2. Determine a port that your system is not listening on. This example uses port 8900.
3. Start a Python SimpleHTTPServer in the `/var/www/html` directory:

```
cd /var/www/html
```

```
python -m SimpleHTTPServer 8900
```

```
Serving HTTP on 0.0.0.0 port 8900 ...
```

4. Visit the Repository URL `http://<web_server>:8900/cloudera-repos/` in your browser and verify the files you downloaded are present.

Configuring Hosts to Use the Internal Repository

After you establish the repository, modify the client configuration to use it.

OS	Procedure
RHEL compatible	<p>Create <code>/etc/yum.repos.d/cloudera-repo.repo</code> files on cluster hosts with the following content, where <code><web_server></code> is the hostname of the Web server:</p> <pre>[cloudera-repo] name=cloudera-repo baseurl=http://<web_server>/cm7 enabled=1 gpgcheck=0</pre>

Configuring a Local Parcel Repository

You can create a parcel repository for Cloudera Manager either by hosting an internal Web repository or by manually copying the repository files to the Cloudera Manager Server host for distribution to Cloudera Manager Agent hosts.

Related Information

[Overview of Parcels](#)

Using an Internally Hosted Remote Parcel Repository

The following sections describe how to use an internal Web server to host a parcel repository.

Related Information

[Overview of Parcels](#)

Setting Up a Web Server

To host an internal repository, you must install or use an existing Web server on an internal host that is reachable by the Cloudera Manager host, and then download the repository files to the Web server host.

About this task

The examples on this page use Apache HTTP Server as the Web server. If you already have a Web server in your organization, you can skip to *Downloading and Publishing the Parcel Repository*.

Procedure

1. Install Apache HTTP Server:

RHEL / CentOS

```
sudo yum install httpd
```

2. Edit the Apache HTTP Server configuration file (`/etc/httpd/conf/httpd.conf` by default) to add or edit the following line in the `<IfModule mime_module>` section:

```
AddType application/x-gzip .gz .tgz .parcel
```

If the `<IfModule mime_module>` section does not exist, you can add it in its entirety as follows:



Note: This example configuration was modified from the default configuration provided after installing Apache HTTP Server on RHEL 7.

```
<IfModule mime_module>
```

```

#
# TypesConfig points to the file containing the list of mappings from
# filename extension to MIME-type.
#
TypesConfig /etc/mime.types
#
# AddType allows you to add to or override the MIME configuration
# file specified in TypesConfig for specific file types.
#
#AddType application/x-gzip .tgz
#
# AddEncoding allows you to have certain browsers uncompress
# information on the fly. Note: Not all browsers support this.
#
#AddEncoding x-compress .Z
#AddEncoding x-gzip .gz .tgz
#
# If the AddEncoding directives above are commented-out, then you
# probably should define those extensions to indicate media types:
#
AddType application/x-compress .Z
AddType application/x-gzip .gz .tgz .parcel

#
# AddHandler allows you to map certain file extensions to "handlers":
# actions unrelated to filetype. These can be either built into the se
rver
# or added with the Action directive (see below)
#
# To use CGI scripts outside of ScriptAliased directories:
# (You will also need to add "ExecCGI" to the "Options" directive.)
#
#AddHandler cgi-script .cgi

# For type maps (negotiated resources):
#AddHandler type-map var

#
# Filters allow you to process content before it is sent to the client
.
#
# To parse .shtml files for server-side includes (SSI):
# (You will also need to add "Includes" to the "Options" directive.)
#
AddType text/html .shtml
AddOutputFilter INCLUDES .shtml
</IfModule>

```



Warning: Skipping this step could result in an error message Hash verification failed when trying to download the parcel from a local repository, especially in Cloudera Manager 6 and higher.

3. Start Apache HTTP Server:

RHEL 7

```
sudo systemctl start httpd
```

Downloading and Publishing the Parcel Repository

Download the parcels that you want to install and publish the parcel directory.

Procedure

1. Download manifest.json and the parcel files for the product you want to install:

Runtime 7

Apache Impala, Apache Kudu, Apache Spark 2, and Cloudera Search are included in the Runtime parcel. To download the files for the latest Runtime 7 release, run the following commands on the Web server host:

```
sudo mkdir -p /var/www/html/cloudera-repos
sudo wget --recursive --no-parent --no-host-directories https://[username]:[password]@archive.cloudera.com/p/cdh7/7.1.5.0/parcels/ -P /var/www/html/cloudera-repos

sudo chmod -R ugo+rX /var/www/html/cloudera-repos/cdh7
```

Sqoop Connectors

To download the parcels for a Sqoop Connector release, run the following commands on the Web server host. This example uses the latest available Sqoop Connectors:

```
sudo mkdir -p /var/www/html/cloudera-repos
sudo wget --recursive --no-parent --no-host-directories http://archive.cloudera.com/sqoop-connectors/parcels/latest/ -P /var/www/html/cloudera-repos
sudo chmod -R ugo+rX /var/www/html/cloudera-repos/sqoop-connectors
```

If you want to create a repository for a different Sqoop Connector release, replace latest with the Sqoop Connector version that you want. You can see a list of versions in the parcels parent directory.

2. Visit the Repository URL `http://<Web_server>/cloudera-repos/` in your browser and verify the files you downloaded are present. If you do not see anything, your Web server may have been configured to not show indexes.

Related Information

[Overview of Parcels](#)

Configuring Cloudera Manager to Use an Internal Remote Parcel Repository

In Cloudera Manager's parcel settings, add a path to the internal parcel repository.

Procedure

1. Use one of the following methods to open the parcel settings page:
 - Navigation bar:
 - a. Click the parcel icon in the left navigation bar or click Hosts and click the Parcels tab.
 - b. Click the Configuration button.
 - Menu:
 - a. Select AdministrationSettings.
 - b. Select CategoryParcels.
2. Enter the path to the parcel. For example: `http://<web_server>/cloudera-parcels/cdh7/7.0.3.1/`

Using a Local Parcel Repository

To use a local parcel repository, complete the following steps:

Procedure

1. Open the Cloudera Manager Admin Console and click Parcels in the left-side navigation menu.

2. Select Configuration and verify that you have a Local Parcel Repository path set. By default, the directory is `/opt/cloudera/parcel-repo`.
3. Remove any Remote Parcel Repository URLs that you are not using, including ones that point to Cloudera archives.
4. Add the parcel you want to use to the local parcel repository directory that you specified. For instructions on downloading parcels, see [Downloading and Publishing the Parcel Repository](#) above.
5. In the command line, navigate to the local parcel repository directory.
6. Create a SHA1 hash for the parcel you added and save it to a file named `parcel_name.parcel.sha`.
For example, the following command generates a SHA1 hash for the parcel `CDH-6.1.0-1.cdh6.1.0.p0.770702-el7.parcel`:

```
sha1sum CDH-6.1.0-1.cdh6.1.0.p0.770702-el7.parcel | awk '{ print $1 }'  
> CDH-6.1.0-1.cdh6.1.0.p0.770702-el7.parcel.sha
```

7. Change the ownership of the parcel and hash files to `cloudera-scm`:

```
sudo chown -R cloudera-scm:cloudera-scm /opt/cloudera/parcel-repo/*
```

8. In the Cloudera Manager Admin Console, click **Parcels** page in the left-side navigation menu.
9. Click **Check for New Parcels** and verify that the new parcel appears.
10. Download, distribute, and activate the parcel.

Production Installation: Installing Cloudera Manager, Cloudera Runtime, and Managed Services

This procedure is recommended for installing Cloudera Manager and Cloudera Runtime for production environments. For a non-production trial install see *Installing the CDP Private Cloud Base Trial*.

Before you begin the installation, make sure you have reviewed the requirements and other considerations described in *Before You Install*.

The general steps in the installation procedure are as follows:

- [Step 1: Configure a Repository for Cloudera Manager](#) on page 70
- [Step 2: Install Java Development Kit](#) on page 71
- [Step 3: Install Cloudera Manager Server](#) on page 76
- [Step 4: Install and Configure Databases](#) on page 76
- [Step 5: Set up the Cloudera Manager Database](#) on page 107
- [Step 6: Install Runtime and Other Software](#) on page 109
- [Step 7: Set Up a Cluster Using the Wizard](#) on page 114

Step 1: Configure a Repository for Cloudera Manager

How to configure a package repository to install Cloudera Manager.

Cloudera Manager is installed using package management tools such as `yum` for RHEL compatible systems. These tools depend on access to repositories to install software. Cloudera maintains Internet-accessible repositories for Runtime and Cloudera Manager installation files.

You can also create your own internal repository for hosts that do not have Internet access. For more information on creating an internal repository for Cloudera Manager, see [Configuring a Local Package Repository](#) on page 65.

To use the Cloudera repository:

1. Download the repository file for your operating system and version on the Cloudera Manager server host:

RHEL / CentOS

```
https://[username]:[password]@archive.cloudera.com/p/cm7-scale/7.2.3/redhat7/yum/cloudera-manager.repo
```

For example:

```
sudo wget https://[username]:[password]@archive.cloudera.com/p/cm7-scale/7.2.3/redhat7/yum/cloudera-manager.repo
```

Move the cloudera-manager.repo file to the /etc/yum.repos.d/ directory.

2. Edit the repository file and add your username and password:

RHEL / CentOS

Open the /etc/yum.repos.d/cloudera-manager.repo file in a text editor. The file will look like this:

```
[cloudera-manager]
name=Cloudera Manager 7.2.3
baseurl=https://archive.cloudera.com/p/cm7/7.2.3/redhat7/yum/
gpgkey=https://archive.cloudera.com/p/cm7/7.2.3/redhat7/yum/RPM-GPG-KEY-cloudera
username=changeme
password=changeme
gpgcheck=1
enabled=1
autorefresh=0
type=rpm-md
```

Replace the two *changeme* placeholders with your username and password.

3. Import the repository signing GPG key:

- RHEL 7 compatible:

```
sudo rpm --import https://[username]:[password]@archive.cloudera.com/p/cm7/7.2.3/redhat7/yum/RPM-GPG-KEY-cloudera
```

4. Continue to *Step 2: Install Java Development Kit*.

Step 2: Install Java Development Kit

CDP Private Cloud Base requires a JDK installed on all hosts., you can either install OpenJDK or a Oracle JDK directly from Oracle.

There are several options for installing a JDK on your CDP Private Cloud Base hosts:

- Install OpenJDK 8 on the Cloudera Manager server host and then allow Cloudera Manager to install OpenJDK 8 on its managed hosts.
- Manually install a [supported JDK](#) on all cluster hosts before installing Cloudera software.

Requirements:

- The JDK must be 64-bit. Do not use a 32-bit JDK.
- The installed JDK must be a supported version as documented in .
- The same version of the JDK must be installed on each cluster host.
- The JDK must be installed at /usr/java/jdk-version.

**Important:**

- The RHEL-compatible operating system supported by CDP Private Cloud Base 7 uses AES-256 encryption by default for tickets. To support AES-256 bit encryption in JDK versions lower than 1.8u161, you must install the Java Cryptography (JCE) Unlimited Strength Jurisdiction Policy File on all cluster and Hadoop user machines. Cloudera Manager can automatically install the policy files, or you can install them manually. For JCE Policy File installation instructions, see the README.txt file included in the jce_policy-x.zip file. JDK 1.8u161 and higher enable unlimited strength encryption by default, and do not require policy files.
- On SLES platforms, do not install or try to use the IBM Java version bundled with the SLES distribution.

Related Information[Java Requirements](#)[Java Requirements](#)**Installing OpenJDK**

After you configure a repository, you can install OpenJDK on the Cloudera Manager Server host using your package manager.



Important: If you are using OpenJDK versions 1.8 u242 or 11.0.6 and have enabled Kerberos, you may experience authentication errors when running cluster services. To work around this problem:

1. Log in to the Cloudera Manager Admin Console.
2. Go to AdministrationSettings.
3. Select the Advanced category.
4. Locate the JVM Arguments for Java-based services parameter and enter the following:

```
-Dsun.security.krb5.disableReferrals=true
```

5. Restart any stale services.

- RHEL Compatible

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

You can use Cloudera Manager to install Open JDK 8 on the remaining cluster hosts in an upcoming step. Continue to *Step 3. Installing Cloudera Manager Server*.

Manually Installing OpenJDK

Before installing Cloudera Manager and Runtime, perform the steps in this section to install OpenJDK on all hosts in your cluster(s).

About this task

Note that the path for the default truststore for OpenJDK 8 is jre/lib/security/cacerts.

- The package names used when installing the OpenJDK 11 and OpenJDK 8 are different and are noted in the steps below.
- The path for the default truststore has changed from (OpenJDK 8) jre/lib/security/cacerts to (OpenJDK 11) lib/security/cacerts
- See the following blog post for general information about migrating to Java 11: [All You Need to Know For Migrating To Java 11](#).



Important: When you install CDP Private Cloud Base, Cloudera Manager includes an option to install Oracle JDK. De-select this option before continuing with the installation.

You must install a supported version of OpenJDK. If your deployment uses a version of OpenJDK lower than 1.8.0_181, see *TLS Protocol Error with OpenJDK*.



Note: If you intend to enable Auto-TLS, note the following:

You can specify a PEM file containing trusted CA certificates to be imported into the Auto-TLS truststore. If you want to use the certificates in the cacerts truststore that comes with OpenJDK, you must convert the truststore to PEM format first. However, OpenJDK ships with some intermediate certificates that cannot be imported into the Auto-TLS truststore. You must remove these certificates from the PEM file before importing the PEM file into the Auto-TLS truststore. This is not required when upgrading to OpenJDK from a cluster where Auto-TLS has already been enabled.

Procedure

1. Log in to each host and run the command for the version of the JDK you want to install:

RHEL

OpenJDK 8

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

OpenJDK 11

```
su -c yum install java-11-openjdk-devel
```

2. Tune the JDK (OpenJDK 11 only.)

OpenJDK 11 uses new defaults for garbage collection and other Java options specified when launching Java processes. Due to these changes you may need to tune the garbage collection by adjusting the Java options used to run cluster services, which are configured separately for each service using the service's configuration parameters. To locate the correct parameter, log in to the Cloudera Manager Admin Console, go to the cluster and service you want to configure and search for "Java Configuration Options".

When using OpenJDK 11, Cloudera Manager and most Cloudera Runtime services use G1GC as the default method of garbage collection. Java 8 used "ConcurrentMarkSweep" (CMS) for garbage collection. When using G1GC, the pauses for garbage collection are shorter, so components will usually be more responsive, but they are more sensitive to JVMs with overcommitted memory usage. See [Tuning JVM Garbage Collection](#) on page 73.

Manually Installing Oracle JDK

The Oracle JDK installer is available both as an RPM-based installer for RPM-based systems, and as a .tar.gz file. These instructions are for the .tar.gz file.

Procedure

1. Download the .tar.gz file for one of the 64-bit supported versions of the Oracle JDK from Java SE 8 Downloads.



Note: If you want to download the JDK directly using a utility such as wget, you must accept the Oracle license by configuring headers, which are updated frequently. Blog posts and Q&A sites can be a good source of information on how to download a particular JDK version using wget.

2. Extract the JDK to /usr/java/jdk-version. For example:

```
tar xvfz /path/to/jdk-8u<update_version>-linux-x64.tar.gz -C /usr/java/
```

3. Repeat this procedure on all cluster hosts.

Results

After you have finished, continue to *Step 3: Install Cloudera Manager Server*.

Tuning JVM Garbage Collection

When using OpenJDK 11, Cloudera Manager and most Cloudera Runtime services use G1GC as the default method of garbage collection. (Java 8 used "ConcurrentMarkSweep" (CMS) for garbage collection.) When using G1GC,

the pauses for garbage collection are shorter, so components will usually be more responsive, but they are more sensitive to overcommitted memory usage. You should monitor memory usage to determine whether memory is overcommitted.

Cloudera Manager alerts you when memory is overcommitted on cluster hosts. To view these alerts and adjust the allocations:

1. Log in to the Cloudera Manager Admin Console
2. Go to HomeConfigurationConfiguration Issues.
3. Look for entries labeled Memory Overcommit Validation Threshold and note the hostname of the affected host.
4. Go to HostsAll Hosts and click on the affected host.
5. Click the Resources tab.
6. Scroll down to the Memory section.

A list of roles instances and their memory allocations are displayed. The Description column displays the configuration property name where the memory allocation can be set.

7. To adjust the memory allocation, search for the configuration property and adjust the value to reduce the overcommitment of memory. You may need to move some roles to other hosts if there is not sufficient memory for the roles running on the host.
8. After making any changes, Cloudera Manager will indicate that the service has a stale configuration and prompt you to [restart the service](#).

You may also need to adjust the Java options used to start Java processes. You can add Java startup options using Cloudera Manager configuration properties that are available for all service roles. Cloudera has provided default arguments for some of the services where they are needed. You can add to these, or completely override all of the provided Java options. For more information on configuring G1GC, see [The OpenJDK documentation](#).

If default options are provided, the role configuration specifies a single value, {JAVA_GC_ARGS}. This value is a placeholder for the default Java Garbage Collection options provided with Cloudera Manager and Cloudera Runtime.

To modify Java options:

1. Log in to the Cloudera Manager Admin Console.
2. Go to the service where you want to modify the options.
3. Select the Configuration tab.
4. Enter "Java" in the search box.
5. Locate the Java Configuration Options property named for the role you want to modify. For example, in the HDFS service, you will see parameters like Java Configuration Options for DataNode and Java Configuration Options for JournalNode.
6. To add to the Java options, enter additional options before or after the {JAVA_GC_ARGS} placeholder, separated by spaces. For example:

```
{JAVA_GC_ARGS} -XX:MaxPermSize=512M
```

7. To replace the default Java options, delete the {JAVA_GC_ARGS} placeholder and replace it with one or more Java options, separated by spaces.
8. The service will now have a stale configuration and must be restarted. See [Restarting a service](#).

Table 26: Default Java Options

Service and Role	Default Java 8 Options	Default Java 11 Options
<ul style="list-style-type: none"> HDFS DataNode HDFS NameNode HDFS Secondary NameNode 	<pre>-XX:+UseParNewGC -XX: +UseConcMarkSweepGC - XX:CMSInitiatingOccupancyFraction=70 -XX: +CMSParallelRemarkEnabled</pre>	<pre>-XX:+UseConcMarkSweepGC -XX:CMSInitiatingOccupancyFraction=70 -XX:+CMSParallelRemarkEnabled</pre>
<ul style="list-style-type: none"> Hive Metastore Server HiveServer 2 WebHCat Server 	<pre>-XX:+UseParNewGC -XX: +UseConcMarkSweepGC - XX:CMSInitiatingOccupancyFraction=70 -XX: +CMSParallelRemarkEnabled</pre>	None, G1GC is enabled by default.
<ul style="list-style-type: none"> HBase REST Server HBase Thrift Server HBase Master HBase RegionServer 	<pre>-XX:+UseParNewGC -XX: +UseConcMarkSweepGC - XX:CMSInitiatingOccupancyFraction=70 -XX: +CMSParallelRemarkEnabled</pre>	None, G1GC is enabled by default.
<ul style="list-style-type: none"> HBase Region Server 	<pre>-XX:+UseParNewGC -XX: +UseConcMarkSweepGC - XX:CMSInitiatingOccupancyFraction=70 -XX: +CMSParallelRemarkEnabled -verbose:gc -XX:+PrintGCDetails -XX:+PrintGCDateStamps</pre>	<pre>-verbose:gc -Xlog:gc</pre>
<ul style="list-style-type: none"> MapReduce JobTracker MapReduce TaskTracker 	<pre>-XX:+UseParNewGC -XX: +UseConcMarkSweepGC - XX:CMSInitiatingOccupancyFraction=70 -XX: +CMSParallelRemarkEnabled</pre>	None, G1GC is enabled by default.
<ul style="list-style-type: none"> Solr Server 	<pre>-XX:+UseParNewGC -XX: +UseConcMarkSweepGC - XX:CMSInitiatingOccupancyFraction=70 -XX: +CMSParallelRemarkEnabled</pre>	None, G1GC is enabled by default.
<ul style="list-style-type: none"> YARN JobHistory Server YARN NodeManager YARN Resource Manager 	<pre>-XX:+UseParNewGC -XX: +UseConcMarkSweepGC - XX:CMSInitiatingOccupancyFraction=70 -XX:</pre>	<pre>-Dlibrary.leveldbjni .path={ {CMF_CONF_DIR} }</pre>

Step 3: Install Cloudera Manager Server

In this step you install the Cloudera Manager packages on the Cloudera Manager Server host, and optionally enable auto-TLS.

Install Cloudera Manager Packages

Cloudera Manager is installed on the Cloudera Manager Server host using packages.

Procedure

1. On the Cloudera Manager Server host, type the following commands to install the Cloudera Manager packages:

OS	Command
RHEL	<pre>sudo yum install cloudera-manager-daemons cloudera-manager-agent cloudera-manager-server</pre>

2. If you are using an Oracle database for Cloudera Manager Server, edit the `/etc/default/cloudera-scm-server` file on the Cloudera Manager server host. Locate the line that begins with `export CMF_JAVA_OPTS` and change the `-Xmx2G` option to `-Xmx4G`.
3. If you are installing on Ubuntu, and are planning to add the Kudu service to the cluster and are planning to enable Apache Ranger, run the following command on all cluster hosts:

```
sudo apt-get install gettext-base
```



Note: If you know in advance which hosts will be running the Kudu service roles, you only need to run this command on those hosts.

Step 4. Install and Configure Databases

Cloudera Manager uses various databases and datastores to store information about the Cloudera Manager configuration, as well as information such as the health of the system, or task progress.

Although you can deploy different types of databases in a single environment, doing so can create unexpected complications. Cloudera recommends choosing one supported database provider for all of the Cloudera databases.

Cloudera recommends installing the databases on different hosts than the services. Separating databases from services can help isolate the potential impact from failure or resource contention in one or the other. It can also simplify management in organizations that have dedicated database administrators.

For information about supported databases, see [Database Requirements](#)

Required Databases

The following components all require databases: Cloudera Manager Server, Oozie Server, Sqoop Server, Reports Manager, Hive Metastore Server, Hue Server, DAS server, Ranger, Schema Registry, and Streams Messaging Manager.

The type of data contained in the databases and their relative sizes are as follows:

- Cloudera Manager Server - Contains all the information about services you have configured and their role assignments, all configuration history, commands, users, and running processes. This relatively small database (< 100 MB) is the most important to back up.



Important: When you restart processes, the configuration for each of the services is redeployed using information saved in the Cloudera Manager database. If this information is not available, your cluster cannot start or function correctly. You must schedule and maintain regular backups of the Cloudera Manager database to recover the cluster in the event of the loss of this database.

- Oozie Server - Contains Oozie workflow, coordinator, and bundle data. Can grow very large. (Only available when installing CDH 5 or CDH 6 clusters.)
- Sqoop Server - Contains entities such as the connector, driver, links and jobs. Relatively small. (Only available when installing CDH 5 or CDH 6 clusters.)
- Reports Manager - Tracks disk utilization and processing activities over time. Medium-sized.
- Hive Metastore Server - Contains Hive metadata. Relatively small.
- Hue Server - Contains user account information, job submissions, and Hive queries. Relatively small.
- Sentry Server - Contains authorization metadata. Relatively small.
- Cloudera Navigator Audit Server - Contains auditing information. In large clusters, this database can grow large. (Only available when installing CDH 5 or CDH 6 clusters.)
- Cloudera Navigator Metadata Server - Contains authorization, policies, and audit report metadata. Relatively small. (Only available when installing CDH 5 or CDH 6 clusters.)
- DAS server - Contains Hive and Tez event logs and DAG information. Can grow very large.
- Ranger Admin - Contains administrative information such as Ranger users, groups, and access policies. Medium-sized.
- Schema Registry - Contains the schemas and their metadata, all the versions and branches. Usually small, but can be large when a lot of schemas are in use.



Important: For the Schema Registry database, you must set collation to be case sensitive.

- Streams Messaging Manager Server - Contains Kafka metadata, stores metrics, and alert definitions. Relatively small.

The Host Monitor and Service Monitor services use local disk-based datastores.

The JDBC connector for your database must be installed on the hosts where you assign the Activity Monitor and Reports Manager roles.

For instructions on installing and configuring databases for Cloudera Manager, Runtime, and other managed services, see the instructions for the type of database you want to use.

Related Information

[Database Requirements](#)

Install and Configure PostgreSQL for CDP

To use a PostgreSQL database, follow these procedures. For information on compatible versions of the PostgreSQL database, see [Database Requirements](#) on page 21.



Note: The following instructions are for a dedicated PostgreSQL database for use in production environments, and are unrelated to the embedded PostgreSQL database provided by Cloudera for trial installations.

Installing PostgreSQL Server

Install the PostgreSQL packages on the PostgreSQL server.

**Note:**

- If you already have a PostgreSQL database set up, you can skip to the section *Configuring and Starting the PostgreSQL Server* to verify that your PostgreSQL configurations meet the requirements for Cloudera Manager.
- Make sure that the data directory, which by default is `/var/lib/postgresql/data/`, is on a partition that has sufficient free space.
- Cloudera Manager supports the use of a custom schema name for the Cloudera Manager Server database, but not the Runtime component databases (such as Hive and Hue). For more information, see *Schemas* in the PostgreSQL documentation.

Install the PostgreSQL packages as follows:

RHEL:

```
sudo yum install postgresql-server
```

Installing the psycopg2 Python Package

Hue in Runtime 7 requires version 2.7.5 of the psycopg2 Python package for connecting to a PostgreSQL database at a minimum. The psycopg2 package is automatically installed as a dependency of Cloudera Manager Agent, but the version installed is often lower than 2.7.5.

If you are installing Runtime 7 and using PostgreSQL for the Hue database, you must install one of the recommended psycopg2 package versions on all Hue hosts.

Recommended psycopg2 package versions: 2.7.5, 2.7.6.1, and 2.7.7.

The following sample commands install version 2.7.5:

RHEL 7 Compatible

1. Install the python-pip package:

```
sudo yum install python-pip
```

2. Install psycopg2 2.7.5 using pip:

```
sudo pip install psycopg2==2.7.5 --ignore-installed
```

Configuring and Starting the PostgreSQL Server

By default, PostgreSQL only accepts connections on the loopback interface. You must reconfigure PostgreSQL to accept connections from the fully qualified domain names (FQDN) of the hosts hosting the services for which you are configuring databases. If you do not make these changes, the services cannot connect to and use the database on which they depend.

Before you begin

If you are making changes to an existing database, make sure to stop any services that use the database before continuing.

Procedure

1. Make sure that LC_ALL is set to en_US.UTF-8 and initialize the database as follows:

- RHEL 7:

```
echo 'LC_ALL="en_US.UTF-8"' >> /etc/locale.conf  
sudo su -l postgres -c "postgresql-setup initdb"
```

- SLES 12:

```
sudo su -l postgres -c "initdb --pgdata=/var/lib/pgsql/data --encoding=UTF-8"
```

- Ubuntu:

```
sudo service postgresql start
```

2. Enable MD5 authentication. Edit `pg_hba.conf`, which is usually found in `/var/lib/pgsql/data` or `/etc/postgresql/<version>/main`. Add the following line:

```
host all all 127.0.0.1/32 md5
```

If the default `pg_hba.conf` file contains the following line:

```
host all all 127.0.0.1/32 ident
```

then the host line specifying md5 authentication shown above must be inserted before this ident line. Failure to do so may cause an authentication error when running the `scm_prepare_database.sh` script. You can modify the contents of the md5 line shown above to support different configurations. For example, if you want to access PostgreSQL from a different host, replace 127.0.0.1 with your IP address and update `postgresql.conf`, which is typically found in the same place as `pg_hba.conf`, to include:

```
listen_addresses = '*'
```

3. Configure settings to ensure your system performs as expected. Update these settings in the `/var/lib/pgsql/data/postgresql.conf` or `/var/lib/postgresql/data/postgresql.conf` file. Settings vary based on cluster size and resources as follows:

- Small to mid-sized clusters - Consider the following settings as starting points. If resources are limited, consider reducing the buffer sizes and checkpoint segments further. Ongoing tuning may be required based on each host's resource utilization. For example, if the Cloudera Manager Server is running on the same host as other roles, the following values may be acceptable:
 - `max_connection` - In general, allow each database on a host 100 maximum connections and then add 50 extra connections. You may have to increase the system resources available to PostgreSQL, as described at *Connection Settings*.
 - `shared_buffers` - 256MB
 - `wal_buffers` - 8MB
 - `checkpoint_segments` - 16



Note: The `checkpoint_segments` setting is removed in PostgreSQL 9.5 and higher, replaced by `min_wal_size` and `max_wal_size`. The PostgreSQL 9.5 release notes provides the following formula for determining the new settings:

```
max_wal_size = (3 * checkpoint_segments) * 16MB
```

- `checkpoint_completion_target` - 0.9

- Large clusters - Can contain up to 1000 hosts. Consider the following settings as starting points.
 - `max_connection` - For large clusters, each database is typically hosted on a different host. In general, allow each database on a host 100 maximum connections and then add 50 extra connections. You may have to increase the system resources available to PostgreSQL, as described at Connection Settings.
 - `shared_buffers` - 1024 MB. This requires that the operating system can allocate sufficient shared memory. See PostgreSQL information on Managing Kernel Resources for more information on setting kernel resources.
 - `wal_buffers` - 16 MB. This value is derived from the `shared_buffers` value. Setting `wal_buffers` to be approximately 3% of `shared_buffers` up to a maximum of approximately 16 MB is sufficient in most cases.
 - `checkpoint_segments` - 128. The PostgreSQL Tuning Guide recommends values between 32 and 256 for write-intensive systems, such as this one.



Note: The `checkpoint_segments` setting is removed in PostgreSQL 9.5 and higher, replaced by `min_wal_size` and `max_wal_size`. The PostgreSQL 9.5 Release Notes provides the following formula for determining the new settings:

```
max_wal_size = (3 * checkpoint_segments) * 16MB
```

- `checkpoint_completion_target` - 0.9.

4. Configure the PostgreSQL server to start at boot.

OS	Command
RHEL 7 compatible	<pre>sudo systemctl enable postgresql</pre>

5. Restart the PostgreSQL database:

- RHEL 7 Compatible:

```
sudo systemctl restart postgresql
```

Creating Databases for Cloudera Software

You must create databases and service accounts for components that require databases.

About this task

The following components require databases:

- Cloudera Manager Server
- Cloudera Management Service roles:
 - Reports Manager
- Data Analytics Studio (DAS) Supported with PostgreSQL only.
- Hue
- Each Hive metastore
- Oozie
- Data Analytics Studio
- Schema Registry
- Streams Messaging Manager

The databases must be configured to support the PostgreSQL UTF8 character set encoding.

Record the values you enter for database names, usernames, and passwords. The Cloudera Manager installation wizard requires this information to correctly connect to these databases.



Note: The instructions for Cloudera Manager Server, Cloudera Management Service roles, Activity Monitor, Reports Manager, Hue, Hive metastores, Oozie, and Data Analytics Studio (DAS) are documented in this topic.

Additional configuration for Ranger is documented in the following two topics. Refer to those topics for detailed instructions on the Ranger database.



Note:

- For DAS, install the PostgreSQL database version 9.6.
- If you are creating more than one Data Hub clusters with DAS, then make sure that you create and use a separate Postgres database for each DAS instance. Ensure this especially when you are creating Data Hub clusters using the Cloudera Manager cluster templates. You can configure a unique database instance by specifying different host, name, or port.

To create databases for Cloudera Manager Server, Cloudera Management Service roles, Activity Monitor, Reports Manager, Hue, Hive metastores, Oozie, and DAS, complete the following steps:

Procedure

1. Connect to PostgreSQL:

```
sudo -u postgres psql
```

2. Create databases for each service you are using from the below table:

```
CREATE ROLE <user> LOGIN PASSWORD '<password>';
```

```
CREATE DATABASE <database> OWNER <user> ENCODING 'UTF8';
```

You can use any value you want for *<database>*, *<user>*, and *<password>*. The following examples are the default names provided in the Cloudera Manager configuration settings, but you are not required to use them:

Table 27: Databases for Cloudera Software

Service	Database	User
Cloudera Manager Server	scm	scm
Reports Manager	rman	rman
Hue	hue	hue
Hive Metastore Server	metastore	hive
Oozie	oozie	oozie
Data Analytics Studio (DAS) Supported with PostgreSQL only.	das	das
Schema Registry	schemaregistry	schemaregistry
Streams Messaging Manager	smm	smm

Record the databases, usernames, and passwords chosen because you will need them later.

3. For PostgreSQL 8.4 and higher, set `standard_conforming_strings=off` for the Hive Metastore and Oozie databases:

```
ALTER DATABASE <database> SET standard_conforming_strings=off;
```

What to do next

- If you plan to use Apache Ranger, see the following topic for instructions on creating and configuring the Ranger database. See [Configuring a PostgreSQL Database for Ranger](#) on page 103.

- If you plan to use Schema Registry or Streams Messaging Manager, see the following topic for instructions on configuring the database: [Configuring the Database for Streaming Components](#) on page 105
- After you install and configure PostgreSQL databases for Cloudera software, continue to *Step 5: Set up the Cloudera Manager Database* to configure a database for Cloudera Manager.

Install and Configure MySQL for Cloudera Software

To use a MySQL database, follow these procedures. For information on compatible versions of the MySQL database, see [Database Requirements](#) on page 21.

Installing the MySQL Server



Note:

- If you already have a MySQL database set up, you can skip to the section [Configuring and Starting the MySQL Server](#) on page 82 to verify that your MySQL configurations meet the requirements for Cloudera Manager.
- For MySQL 5.6 and 5.7, you must install the MySQL-shared-compat or MySQL-shared package. This is required for the Cloudera Manager Agent package installation.
- It is important that the datadir directory, which, by default, is /var/lib/mysql, is on a partition that has sufficient free space.
- Cloudera Manager installation fails if GTID-based replication is enabled in MySQL.

1. Install the MySQL database.

OS	Command
RHEL	<p>MySQL is no longer included with RHEL. You must download the repository from the MySQL site and install it directly. You can use the following commands to install MySQL. For more information, visit the MySQL website.</p> <pre>wget http://repo.mysql.com/mysql-community-release-el7-5.noarch.rpm</pre> <pre>sudo rpm -ivh mysql-community-release-el7-5.noarch.rpm</pre> <pre>sudo yum update</pre> <pre>sudo yum install mysql-server</pre> <pre>sudo systemctl start mysqld</pre>

Configuring and Starting the MySQL Server



Note: If you are making changes to an existing database, make sure to stop any services that use the database before continuing.

1. Stop the MySQL server if it is running.

OS	Command
RHEL 7 Compatible	<pre>sudo systemctl stop mysqld</pre>

2. Move old InnoDB log files /var/lib/mysql/ib_logfile0 and /var/lib/mysql/ib_logfile1 out of /var/lib/mysql/ to a backup location.
3. Determine the location of the [option file](#), my.cnf (/etc/my.cnf by default).

4. Update my.cnf so that it conforms to the following requirements:

- To prevent deadlocks, set the isolation level to READ-COMMITTED.
- Configure the InnoDB engine. Cloudera Manager will not start if its tables are configured with the MyISAM engine. (Typically, tables revert to MyISAM if the InnoDB engine is misconfigured.) To check which engine your tables are using, run the following command from the MySQL shell:

```
mysql> show table status;
```

- The default settings in the MySQL installations in most distributions use conservative buffer sizes and memory usage. Cloudera Management Service roles need high write throughput because they might insert many records in the database. Cloudera recommends that you set the innodb_flush_method property to O_DIRECT.
- Set the max_connections property according to the size of your cluster:
 - Fewer than 50 hosts - You can store more than one database (for example, both the Activity Monitor and Service Monitor) on the same host. If you do this, you should:
 - Put each database on its own physical disk for best performance. You can do this by manually setting up symbolic links or running multiple database instances (each instance uses a different data directory path).
 - Allow 100 maximum connections for each database and then add 50 extra connections. For example, for two databases, set the maximum connections to 250. If you store five databases on one host (the databases for Cloudera Manager Server, Activity Monitor, Reports Manager, and Hive metastore), set the maximum connections to 550.
 - More than 50 hosts - Do not store more than one database on the same host. Use a separate host for each database/host pair. The hosts do not need to be reserved exclusively for databases, but each database should be on a separate host.
- If the cluster has more than 1000 hosts, set the max_allowed_packet property to 16M. Without this setting, the cluster may fail to start due to the following exception: com.mysql.jdbc.PacketTooBigException.
- Binary logging is not a requirement for Cloudera Manager installations. Binary logging provides benefits such as MySQL replication or point-in-time incremental recovery after database restore. Examples of this configuration follow. For more information, see [The Binary Log](#).

Here is an option file with Cloudera recommended settings:

```
[mysqld]
datadir=/var/lib/mysql
socket=/var/lib/mysql/mysql.sock
transaction-isolation = READ-COMMITTED
# Disabling symbolic-links is recommended to prevent assorted security risks;
# to do so, uncomment this line:
symbolic-links = 0

key_buffer_size = 32M
max_allowed_packet = 16M
thread_stack = 256K
thread_cache_size = 64
query_cache_limit = 8M
query_cache_size = 64M
query_cache_type = 1
max_connections = 550
#expire_logs_days = 10
#max_binlog_size = 100M

#log_bin should be on a disk with enough free space.
#Replace '/var/lib/mysql/mysql_binary_log' with an appropriate path for your
#system and chown the specified folder to the mysql user.
log_bin=/var/lib/mysql/mysql_binary_log
```

```
#In later versions of MySQL, if you enable the binary log and do not set
#a server_id, MySQL will not start. The server_id must be unique within
#the replicating group.
server_id=1

binlog_format = mixed
read_buffer_size = 2M
read_rnd_buffer_size = 16M
sort_buffer_size = 8M
join_buffer_size = 8M

# InnoDB settings
innodb_file_per_table = 1
innodb_flush_log_at_trx_commit = 2
innodb_log_buffer_size = 64M
innodb_buffer_pool_size = 4G
innodb_thread_concurrency = 8
innodb_flush_method = O_DIRECT
innodb_log_file_size = 512M

[mysqld_safe]
log-error=/var/log/mysqld.log
pid-file=/var/run/mysqld/mysqld.pid

sql_mode=STRICT_ALL_TABLES
```

5. If AppArmor is running on the host where MySQL is installed, you might need to configure AppArmor to allow MySQL to write to the binary.
6. Ensure the MySQL server starts at boot:

OS	Command
RHEL 7 compatible	<pre>sudo systemctl enable mysqld</pre>

7. Start the MySQL server:

OS	Command
RHEL 7 Compatible	<pre>sudo systemctl start mysqld</pre>

8. Run `/usr/bin/mysql_secure_installation` to set the MySQL root password and other security-related settings. In a new installation, the root password is blank. Press the Enter key when you're prompted for the root password. For the rest of the prompts, enter the responses listed below in bold:

```
sudo /usr/bin/mysql_secure_installation
```

```
[...]
Enter current password for root (enter for none):
OK, successfully used password, moving on...
[...]
Set root password? [Y/n] Y
New password:
Re-enter new password:
Remove anonymous users? [Y/n] Y
[...]
Disallow root login remotely? [Y/n] N
[...]
Remove test database and access to it [Y/n] Y
[...]
Reload privilege tables now? [Y/n] Y
```

All done!

Installing the MySQL JDBC Driver

Install the JDBC driver on the Cloudera Manager Server host, as well as any other hosts running services that require database access.




Note: If you already have the JDBC driver installed on the hosts that need it, you can skip this section. However, MySQL 5.6 requires a 5.1 driver version 5.1.26 or higher.

Cloudera recommends that you consolidate all roles that require databases on a limited number of hosts, and install the driver on those hosts. Locating all such roles on the same hosts is recommended but not required. Make sure to install the JDBC driver on each host running roles that access the database.



Note: Cloudera recommends using only version 5.1 of the JDBC driver.

OS	Command
RHEL	<p> Important: Using the yum install command to install the MySQL driver package before installing a JDK installs OpenJDK, and then uses the Linux alternatives command to set the system JDK to be OpenJDK. If you intend to use an Oracle JDK, make sure that it is installed before installing the MySQL driver using yum install. If you want to use OpenJDK, you can install the driver using yum.</p> <p>Alternatively, use the following procedure to manually install the driver.</p> <ol style="list-style-type: none"> 1. Download the MySQL JDBC driver from http://www.mysql.com/downloads/connector/j/5.1.html (in .tar.gz format). As of the time of writing, you can download version 5.1.46 using wget as follows: <pre>wget https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-java-5.1.46.tar.gz</pre> 2. Extract the JDBC driver JAR file from the downloaded file. For example: <pre>tar zxvf mysql-connector-java-5.1.46.tar.gz</pre> 3. Copy the JDBC driver, renamed, to /usr/share/java/. If the target directory does not yet exist, create it. For example: <pre>sudo mkdir -p /usr/share/java/ cd mysql-connector-java-5.1.46 sudo cp mysql-connector-java-5.1.46-bin.jar /usr/share/java/mysql-connector-java.jar</pre>

Creating Databases for Cloudera Software

Create databases and service accounts for components that require databases:

- Cloudera Manager Server
- Cloudera Management Service roles:
 - Reports Manager
- Data Analytics Studio (DAS) Supported with PostgreSQL only.
- Hue
- Each Hive metastore
- Oozie
- Data Analytics Studio
- Schema Registry
- Streams Messaging Manager

1. Log in as the root user, or another user with privileges to create database and grant privileges:

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

```
Enter password:
```

2. Create databases for each service deployed in the cluster using the following commands. You can use any value you want for the `<database>`, `<user>`, and `<password>` parameters. The Databases for Cloudera Software table, below lists the default names provided in the Cloudera Manager configuration settings, but you are not required to use them.

Configure all databases to use the utf8 character set.

Include the character set for each database when you run the CREATE DATABASE statements described below.

```
CREATE DATABASE <database> DEFAULT CHARACTER SET utf8 DEFAULT COLLATE utf8_general_ci;
```

```
Query OK, 1 row affected (0.00 sec)
```

```
GRANT ALL ON <database>.* TO '<user>'@'%' IDENTIFIED BY '<password>';
```

```
Query OK, 0 rows affected (0.00 sec)
```

Table 28: Databases for Cloudera Software

Service	Database	User
Cloudera Manager Server	scm	scm
Reports Manager	rman	rman
Hue	hue	hue
Hive Metastore Server	metastore	hive
Oozie	oozie	oozie
Data Analytics Studio (DAS) Supported with PostgreSQL only.	das	das
Schema Registry	schemaregistry	schemaregistry
Streams Messaging Manager	smm	smm

3. Confirm that you have created all of the databases:

```
SHOW DATABASES ;
```

You can also confirm the privilege grants for a given user by running:

```
SHOW GRANTS FOR '<user>'@'%' ;
```

4. Record the values you enter for database names, usernames, and passwords. The Cloudera Manager installation wizard requires this information to correctly connect to these databases.

Next Steps

- If you plan to use Apache Ranger, see the following topic for instructions on creating and configuring the Ranger database. See [Configuring a Ranger Database: MySQL/MariaDB](#) on page 101.
- If you plan to use Schema Registry or Streams Messaging Manager, see the following topic for instructions on configuring the database: [Configuring the Database for Streaming Components](#) on page 105

- After you install and configure PostgreSQL databases for Cloudera software, continue to *Step 5: Set up the Cloudera Manager Database* to configure a database for Cloudera Manager.

Install and Configure MariaDB for Cloudera Software

To use a MariaDB database, follow these procedures. For information on compatible versions of MariaDB, see [Database Requirements](#) on page 21.

Installing MariaDB Server



Note:

- If you already have a MariaDB database set up, you can skip to the section [Configuring and Starting the MariaDB Server](#) on page 87 to verify that your MariaDB configurations meet the requirements for Cloudera Manager.
- It is important that the datadir directory (/var/lib/mysql by default), is on a partition that has sufficient free space. For more information, see [Hardware Requirements](#) on page 8.

1. Install MariaDB server:

OS	Command
RHEL compatible	<pre>sudo yum install mariadb-server</pre>

If these commands do not work, you might need to add a repository or use a different yum install command, particularly on RHEL 6 compatible operating systems. For more assistance, see the following topics on the MariaDB website:

- RHEL compatible: [Installing MariaDB with yum](#)

Configuring and Starting the MariaDB Server



Note: If you are making changes to an existing database, make sure to stop any services that use the database before continuing.

1. Stop the MariaDB server if it is running:

OS	Command
RHEL 7 compatible, SLES, and Ubuntu	<pre>sudo systemctl stop mariadb</pre>

2. If they exist, move old InnoDB log files /var/lib/mysql/ib_logfile0 and /var/lib/mysql/ib_logfile1 out of /var/lib/mysql/ to a backup location.
3. Determine the location of the [option file](#), my.cnf (/etc/my.cnf by default).
4. Update my.cnf so that it conforms to the following requirements:
 - To prevent deadlocks, set the isolation level to READ-COMMITTED.
 - The default settings in the MariaDB installations in most distributions use conservative buffer sizes and memory usage. Cloudera Management Service roles need high write throughput because they might insert

many records in the database. Cloudera recommends that you set the `innodb_flush_method` property to `O_DIRECT`.

- Set the `max_connections` property according to the size of your cluster:
 - Fewer than 50 hosts - You can store more than one database (for example, both the Activity Monitor and Service Monitor) on the same host. If you do this, you should:
 - Put each database on its own physical disk for best performance. You can do this by manually setting up symbolic links or running multiple database instances (each instance uses a different data directory path).
 - Allow 100 maximum connections for each database and then add 50 extra connections. For example, for two databases, set the maximum connections to 250. If you store five databases on one host (the databases for Cloudera Manager Server, Reports Manager, and Hive metastore), set the maximum connections to 550.
 - More than 50 hosts - Do not store more than one database on the same host. Use a separate host for each database/host pair. The hosts do not need to be reserved exclusively for databases, but each database should be on a separate host.
- If the cluster has more than 1000 hosts, set the `max_allowed_packet` property to 16M. Without this setting, the cluster may fail to start due to the following exception: `com.mysql.jdbc.PacketTooBigException`.
- Although binary logging is not a requirement for Cloudera Manager installations, it provides benefits such as MariaDB replication or point-in-time incremental recovery after a database restore. The provided example configuration enables the binary log. For more information, see [The Binary Log](#).

Here is an option file with Cloudera recommended settings:

```
[mysqld]
datadir=/var/lib/mysql
socket=/var/lib/mysql/mysql.sock
transaction-isolation = READ-COMMITTED
# Disabling symbolic-links is recommended to prevent assorted security risks;
# to do so, uncomment this line:
symbolic-links = 0
# Settings user and group are ignored when systemd is used.
# If you need to run mysqld under a different user or group,
# customize your systemd unit file for mariadb according to the
# instructions in http://fedoraproject.org/wiki/Systemd

key_buffer = 16M
key_buffer_size = 32M
max_allowed_packet = 32M
thread_stack = 256K
thread_cache_size = 64
query_cache_limit = 8M
query_cache_size = 64M
query_cache_type = 1

max_connections = 550
#expire_logs_days = 10
#max_binlog_size = 100M
#log_bin should be on a disk with enough free space.
#Replace '/var/lib/mysql/mysql_binary_log' with an appropriate path for your
#system and chown the specified folder to the mysql user.
log_bin=/var/lib/mysql/mysql_binary_log

#In later versions of MariaDB, if you enable the binary log and do not set
#a server_id, MariaDB will not start. The server_id must be unique within
#the replicating group.
server_id=1
```



```

binlog_format = mixed

read_buffer_size = 2M
read_rnd_buffer_size = 16M
sort_buffer_size = 8M
join_buffer_size = 8M
# InnoDB settings
innodb_file_per_table = 1
innodb_flush_log_at_trx_commit = 2
innodb_log_buffer_size = 64M
innodb_buffer_pool_size = 4G
innodb_thread_concurrency = 8
innodb_flush_method = O_DIRECT
innodb_log_file_size = 512M
[mysqld_safe]
log-error=/var/log/mariadb/mariadb.log
pid-file=/var/run/mariadb/mariadb.pid
#
# include all files from the config directory
#
!includedir /etc/my.cnf.d

```

5. If AppArmor is running on the host where MariaDB is installed, you might need to configure AppArmor to allow MariaDB to write to the binary.
6. Ensure the MariaDB server starts at boot:

OS	Command
RHEL 7 compatible, SLES, and Ubuntu	<code>sudo systemctl enable mariadb</code>

7. Start the MariaDB server:

OS	Command
RHEL 7 compatible, SLES, and Ubuntu	<code>sudo systemctl start mysqld</code>

8. Run `/usr/bin/mysql_secure_installation` to set the MariaDB root password and other security-related settings. In a new installation, the root password is blank. Press the Enter key when you're prompted for the root password. For the rest of the prompts, enter the responses listed below in bold:

```
sudo /usr/bin/mysql_secure_installation
```

```

[...]  

Enter current password for root (enter for none):  

OK, successfully used password, moving on...  

[...]  

Set root password? [Y/n] Y  

New password:  

Re-enter new password:  

[...]  

Remove anonymous users? [Y/n] Y  

[...]  

Disallow root login remotely? [Y/n] N  

[...]  

Remove test database and access to it [Y/n] Y  

[...]  

Reload privilege tables now? [Y/n] Y  

[...]  

All done! If you've completed all of the above steps, your MariaDB  

installation should now be secure.

```

Thanks for using MariaDB!

Installing the MySQL JDBC Driver for MariaDB


The MariaDB JDBC driver is not supported. Follow the steps in this section to install and use the MySQL JDBC driver instead.

Install the JDBC driver on the Cloudera Manager Server host, as well as any other hosts running services that require database access.

Cloudera recommends that you consolidate all roles that require databases on a limited number of hosts, and install the driver on those hosts. Locating all such roles on the same hosts is recommended but not required. Make sure to install the JDBC driver on each host running roles that access the database.



Note: Cloudera recommends using only version 5.1 of the JDBC driver.

OS	Command
RHEL	<p> Important: Using the yum install command to install the MySQL driver package before installing a JDK installs OpenJDK, and then uses the Linux alternatives command to set the system JDK to be OpenJDK. If you intend to use an Oracle JDK, make sure that it is installed before installing the MySQL driver using yum install. If you want to use OpenJDK, you can install the driver using yum.</p> <p>Alternatively, use the following procedure to manually install the driver.</p> <ol style="list-style-type: none"> 1. Download the MySQL JDBC driver from http://www.mysql.com/downloads/connector/j/5.1.html (in .tar.gz format). As of the time of writing, you can download version 5.1.46 using wget as follows: <pre>wget https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-java-5.1.46.tar.gz</pre> 2. Extract the JDBC driver JAR file from the downloaded file. For example: <pre>tar zxvf mysql-connector-java-5.1.46.tar.gz</pre> 3. Copy the JDBC driver, renamed, to /usr/share/java/. If the target directory does not yet exist, create it. For example: <pre>sudo mkdir -p /usr/share/java/ cd mysql-connector-java-5.1.46 sudo cp mysql-connector-java-5.1.46-bin.jar /usr/share/ java/mysql-connector-java.jar</pre>

Creating Databases for Cloudera Software

Create databases and service accounts for components that require databases:

- Cloudera Manager Server
- Cloudera Management Service roles:
 - Reports Manager
- Data Analytics Studio (DAS) Supported with PostgreSQL only.
- Hue
- Each Hive metastore
- Oozie
- Data Analytics Studio
- Schema Registry
- Streams Messaging Manager

1. Log in as the root user, or another user with privileges to create database and grant privileges:

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

```
Enter password:
```

2. Create databases for each service deployed in the cluster using the following commands. You can use any value you want for the `<database>`, `<user>`, and `<password>` parameters. The Databases for Cloudera Software table, below lists the default names provided in the Cloudera Manager configuration settings, but you are not required to use them.

Configure all databases to use the utf8 character set.

Include the character set for each database when you run the CREATE DATABASE statements described below.

```
CREATE DATABASE <database> DEFAULT CHARACTER SET utf8 DEFAULT COLLATE utf8_general_ci;
```

```
Query OK, 1 row affected (0.00 sec)
```

```
GRANT ALL ON <database>.* TO '<user>'@'%' IDENTIFIED BY '<password>';
```

```
Query OK, 0 rows affected (0.00 sec)
```

Table 29: Databases for Cloudera Software

Service	Database	User
Cloudera Manager Server	scm	scm
Reports Manager	rman	rman
Hue	hue	hue
Hive Metastore Server	metastore	hive
Oozie	oozie	oozie
Data Analytics Studio (DAS) Supported with PostgreSQL only.	das	das
Schema Registry	schemaregistry	schemaregistry
Streams Messaging Manager	smm	smm

3. Confirm that you have created all of the databases:

```
SHOW DATABASES;
```

You can also confirm the privilege grants for a given user by running:

```
SHOW GRANTS FOR '<user>'@'%' ;
```

4. Record the values you enter for database names, usernames, and passwords. The Cloudera Manager installation wizard requires this information to correctly connect to these databases.

Next Steps

- If you plan to use Apache Ranger, see the following topic for instructions on creating and configuring the Ranger database. See [Configuring a Ranger Database: MySQL/MariaDB](#) on page 101.
- If you plan to use Schema Registry or Streams Messaging Manager, see the following topic for instructions on configuring the database: [Configuring the Database for Streaming Components](#) on page 105

- After you install and configure PostgreSQL databases for Cloudera software, continue to *Step 5: Set up the Cloudera Manager Database* to configure a database for Cloudera Manager.

Install and Configure Oracle Database for Cloudera Software

To use an Oracle database, follow these procedures. For information on compatible versions of the Oracle database, see [Database Requirements](#) on page 21.

Collecting Oracle Database Information

To configure Cloudera Manager to work with an Oracle database, get the following information from your Oracle DBA:

- Hostname - The DNS name or the IP address of the host where the Oracle database is installed.
- SID - The name of the schema that will store Cloudera Manager information.
- Username - A username for each schema that is storing information. You could have four unique usernames for the four schema.
- Password - A password corresponding to each username.

Configuring the Oracle Server



Note: If you are making changes to an existing database, make sure to stop any services that use the database before continuing.

Adjusting Oracle Settings to Accommodate Larger Clusters

Cloudera Management services require high write throughput. Depending on the size of your deployments, your DBA may need to modify Oracle settings for monitoring services. These guidelines are for larger clusters and do not apply to the Cloudera Manager configuration database and to smaller clusters. Many factors help determine whether you need to change your database settings, but in most cases, if your cluster has more than 100 hosts, you should consider making the following changes:

- Enable direct and asynchronous I/O by setting the `FILESYSTEMIO_OPTIONS` parameter to `SETALL`.
- Increase the RAM available to Oracle by changing the `MEMORY_TARGET` parameter. The amount of memory to assign depends on the size of the Hadoop cluster.
- Create more redo log groups and spread the redo log members across separate disks or logical unit numbers.
- Increase the size of redo log members to be at least 1 GB.

Reserving Ports for HiveServer 2

HiveServer2 uses port 10000 by default, but Oracle database changes the local port range. This can cause HiveServer2 to fail to start.

Manually reserve the default port for HiveServer2. For example, the following command reserves port 10000 and inserts a comment indicating the reason:

```
echo << EOF > /etc/sysctl.conf
# HS2 uses port 10000
net.ipv4.ip_local_reserved_ports = 10000
EOF
```

```
sysctl -q -w net.ipv4.ip_local_reserved_ports=10000
```

Modifying the Maximum Number of Oracle Connections

Work with your Oracle database administrator to ensure appropriate values are applied for your Oracle database settings. You must determine the number of connections, transactions, and sessions to be allowed.

Allow 100 maximum connections for each service that requires a database and then add 50 extra connections. For example, for two services, set the maximum connections to 250. If you have five services that require a database on one host (the databases for Cloudera Manager Server, Activity Monitor, Reports Manager, and Hive metastore), set the maximum connections to 550.

From the maximum number of connections, you can determine the number of anticipated sessions using the following formula:

```
sessions = (1.1 * maximum_connections) + 5
```

For example, if a host has a database for two services, anticipate 250 maximum connections. If you anticipate a maximum of 250 connections, plan for 280 sessions.

Once you know the number of sessions, you can determine the number of anticipated transactions using the following formula:

```
transactions = 1.1 * sessions
```

Continuing with the previous example, if you anticipate 280 sessions, you can plan for 308 transactions.

Work with your Oracle database administrator to apply these derived values to your system.

Using the sample values above, Oracle attributes would be set as follows:

```
alter system set processes=250;  
alter system set transactions=308;  
alter system set sessions=280;
```

Ensuring Your Oracle Database Supports UTF8

The database you use must support UTF8 character set encoding. You can implement UTF8 character set encoding in Oracle databases by using the dbca utility. In this case, you can use the characterSet AL32UTF8 option to specify proper encoding. Consult your DBA to ensure UTF8 encoding is properly configured.

Installing the Oracle JDBC Connector

You must install the JDBC connector on the Cloudera Manager Server host and any other hosts that use a database.

Cloudera recommends that you assign all roles that require a database on the same host and install the connector on that host. Locating all such roles on the same host is recommended but not required. If you install a role, such as Activity Monitor, on one host and other roles on a separate host, you would install the JDBC connector on each host running roles that access the database.

1. Download the Oracle JDBC Driver from the Oracle website. For example, the version 6 JAR file is named ojdbc6.jar.

For more information about supported Java versions, see [Java Requirements](#).

To download the JDBC driver, visit the [Oracle JDBC and UCP Downloads](#) page, and click on the link for your Oracle Database version. Download the ojdbc6.jar file (or ojdbc8.jar, for Oracle Database 12.2).

2. Copy the Oracle JDBC JAR file to /usr/share/java/oracle-connector-java.jar. The Cloudera Manager databases and the Hive Metastore database use this shared file. For example:

```
sudo mkdir -p /usr/share/java  
sudo cp /tmp/ojdbc8-12.2.0.1.jar /usr/share/java/oracle-connector-java.jar  
sudo chmod 644 /usr/share/java/oracle-connector-java.jar
```

Creating Databases for Cloudera Software

Create schema and user accounts for components that require databases:

- Cloudera Manager Server

- Cloudera Management Service roles:
 - Reports Manager
- Data Analytics Studio (DAS) Supported with PostgreSQL only.
- Hue
- Each Hive metastore
- Oozie
- Data Analytics Studio
- Schema Registry
- Streams Messaging Manager

You can create the Oracle database, schema and users on the host where the Cloudera Manager Server will run, or on any other hosts in the cluster. For performance reasons, you should install each database on the host on which the service runs, as determined by the roles you assign during installation or upgrade. In larger deployments or in cases where database administrators are managing the databases the services use, you can separate databases from services, but use caution.

The database must be configured to support UTF-8 character set encoding.

Record the values you enter for database names, usernames, and passwords. The Cloudera Manager installation wizard requires this information to correctly connect to these databases.

1. Log into the Oracle client:

```
sqlplus system@localhost
```

```
Enter password: *****
```

2. Create a user and schema for each service you are using from the below table:

```
create user <user> identified by <password> default tablesp
ace <tablespace>;
grant CREATE SESSION to <user>;
grant CREATE TABLE to <user>;
grant CREATE SEQUENCE to <user>;
grant EXECUTE on sys.dbms_lob to <user>;
```

You can use any value you want for *<schema>*, *<user>*, and *<password>*. The following examples are the default names provided in the Cloudera Manager configuration settings, but you are not required to use them:

Table 30: Databases for Cloudera Software

Service	Database	User
Cloudera Manager Server	scm	scm
Reports Manager	rman	rman
Hue	hue	hue
Hive Metastore Server	metastore	hive
Oozie	oozie	oozie
Data Analytics Studio (DAS) Supported with PostgreSQL only.	das	das
Schema Registry	schemaregistry	schemaregistry
Streams Messaging Manager	smm	smm

3. Grant a quota on the tablespace (the default tablespace is SYSTEM) where tables will be created:

```
ALTER USER <user> quota 100m on <tablespace>;
```

or for unlimited space:

```
ALTER USER username quota unlimited on <tablespace>;
```

4. Set the following additional privileges for Oozie:

```
grant alter index to oozie;
grant alter table to oozie;
grant create index to oozie;
grant create sequence to oozie;
grant create session to oozie;
grant create table to oozie;
grant drop sequence to oozie;
grant select dictionary to oozie;
grant drop table to oozie;
alter user oozie quota unlimited on <tablespace>;
```



Important:

For security reasons, do not grant select any table privileges to the Oozie user.

For further information about Oracle privileges, see [Authorization: Privileges, Roles, Profiles, and Resource Limitations](#).

Next Steps

- If you plan to use Apache Ranger, see the following topic for instructions on creating and configuring the Ranger database. See [Configuring a Ranger Database: Oracle](#) on page 102.
- If you plan to use Schema Registry or Streams Messaging Manager, see the following topic for instructions on configuring the database: [Configuring the Database for Streaming Components](#) on page 105
- After you install and configure PostgreSQL databases for Cloudera software, continue to *Step 5: Set up the Cloudera Manager Database* to configure a database for Cloudera Manager.
- If you plan to use Hue in the cluster, see [Configuring the Hue Server to Store Data in the Oracle database](#) on page 95.

Configuring the Hue Server to Store Data in the Oracle database

You can connect Hue to your Oracle database while installing Cloudera Runtime (and Hue).

Connect Hue Service to Oracle

If you want to connect Hue service to Oracle with an existing CDH installation, then connect and restart Hue without saving the data in your current database. Alternatively, you can migrate the old data into Oracle.





New Cloudera Runtime Installation

See [Step 3: Install Cloudera Manager Server](#) on page 76 to install Cloudera Manager (and its Installation Wizard), which you will use here to install Cloudera Runtime and the Oracle client.

Install Hue in CDP with Oracle database 12c and higher

1. Download the zip files for the [Instant Client Package](#), both Basic and SDK (with headers).

Version 12.2.0.1.0

Name	Download	Description
Instant Client Package (ZIP)	 instantclient-basic-linux.x64-12.2.0.1.0.zip	Basic: All files required to run OCI, OCCI, and JDBC-OCI applications (68,965,195 bytes) (cksum - 3923339140)
Instant Client Package (ZIP)	 oracle-instantclient12.2-basic-12.2.0.1.0-1.x86_64.rpm	Basic: All files required to run OCI, OCCI, and JDBC-OCI applications (52,826,628 bytes) (cksum - 888077889)
Name	Download	Description
Instant Client Package (ZIP)	 instantclient-sdk-linux.x64-12.2.0.1.0.zip	SDK: Additional header files and an example makefile for developing Oracle applications with Instant Client (674,743 bytes) (cksum - 2114815674)
Instant Client Package (RPM)	 oracle-instantclient12.2-devel-12.2.0.1.0-1.x86_64.rpm	SDK: Additional header files and an example makefile for developing Oracle applications with Instant Client (606,864 bytes) (cksum - 2680490862)



Note: If you are using Oracle database 11g, then download the corresponding 11g Instant Client Package from the Oracle website.

2. Switch to the host with the downloaded files and upload zip to the Hue server host:

```
scp instantclient-*.zip root@<hue server hostname>:.
```

3. Arrange the client libraries to mirror the tree structure in the image as shown in the following example:

```
# Create nested directories: /usr/share/oracle/instantclient/lib/
mkdir -pm 755 /usr/share/oracle/instantclient/lib
# Unzip. The files expand into /usr/share/oracle/instantclient/instantc
lient_<ver>/
unzip '*.zip' -d /usr/share/oracle/instantclient/

# Move lib files from instantclient_<ver> to /usr/share/oracle/instantcl
ient/lib/
mv /usr/share/oracle/instantclient/`ls -l /usr/share/oracle/instantclient/
| grep instantclient_ | awk '{print $9}'`/lib* /usr/share/oracle/instan
tclient/lib/

# Move rest of the files to /usr/share/oracle/instantclient/
mv /usr/share/oracle/instantclient/`ls -l /usr/share/oracle/instantclient
/ | grep instantclient_ | awk '{print $9}'`/* /usr/share/oracle/instantc
lient/

# Create symbolic links. Remember to edit version numbers as necessary
cd /usr/share/oracle/instantclient/lib
ln -s libclntsh.so.<ver>.1 libclntsh.so
ln -s libocci.so.<ver>.1 libocci.so
# For example:
ln -s libclntsh.so.12.1 libclntsh.so
ln -s libocci.so.12.1 libocci.so
ln -s libclntsh.so.12.1
ln -s libocci.so.12.1 libocci.so.11.1
```

where <ver> is the version of the Instant Client Package. Replace <ver> with the actual version of the Instant Client Package.

4. Set the path for \$ORACLE_HOME and \$LD_LIBRARY_PATH as shown in the following example:

```
export ORACLE_HOME=/usr/share/oracle/instantclient
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$ORACLE_HOME
```

Apply temporary workaround for Oracle 12c client

Update the cx_Oracle package in your native Python environment and copy it to Hue's Python environment. The default cx_Oracle version that is shipped with Cloudera Manager is 5.2.1.

1. Install gcc and Python development tools:

```
## CentOS/RHEL (yum), SLES (zypper), Ubuntu/Debian (apt-get)
yum install -y python-setuptools python-devel gcc
#zypper install -y python-setuptools python-devel gcc
#apt-get install -y python-setuptools python-dev gcc
```

2. Install pip:

```
easy_install pip
```

3. Install cx_Oracle. Ensure that ORACLE_HOME and \$LD_LIBRARY_PATH are properly set so that pip knows which version to install.

```
echo $ORACLE_HOME $LD_LIBRARY_PATH
```

```
pip install cx_Oracle==5.3
```



Tip: You can also wget the proper cx_Oracle file yourself: https://pypi.python.org/pypi/cx_Oracle/.

4. Get the version of the new cx_Oracle package:

- CentOS/RHEL and SLES:

```
ls /usr/lib64/python2.7/site-packages/cx_Oracle*
```

- Ubuntu/Debian:

```
ls /usr/local/lib/python2.7/dist-packages/cx_Oracle*
```

5. If this is a new CDP installation, stop here to run the first 5 steps of the Cloudera Manager Installation Wizard. Do not go past Cluster Installation.
6. Navigate to Hue's python environment, \$HUE_HOME/build/env/lib/<python version>/site-packages.

```
cd /usr/lib/hue/build/env/lib/python2.7/site-packages
```



Note: The parcel path is created during step 5 of the Cluster Installation, so you must have completed this to continue.

7. Move the existing cx_Oracle file:

```
mv cx_Oracle-5.2.1-py2.7-linux-x86_64.egg cxfoo
```

8. Copy the new cx_Oracle module to Hue's python environment. The version can change:

- CentOS/RHEL and SLES:

```
cp -a /usr/lib64/python2.7/site-packages/cx_Oracle-5.3-py2.7.egg-info .
```

- Ubuntu/Debian:

```
cp -a /usr/local/lib/python2.7/dist-packages/cx_Oracle-5.3.egg-info .
```

Connect Hue to Oracle

Continuing with Cloudera Manager Installation Wizard ...

1. Stop at Database Setup to set connection properties (Cluster Setup, step 3).

a. Select Use Custom Database.

b. Under Hue, set the connection properties to the Oracle database.



Note: Copy and store the password for the Hue embedded database (just in case).

```
Database Hostname (and port): <fqdn of host with Oracle server>:1521
Database Type (or engine): Oracle
Database SID (or name): orcl
Database Username: hue
Database Password: <hue database password>
```

c. Click Test Connection and click Continue when successful.

2. Continue with the installation and click Finish to complete.

3. Add support for a multi-threaded environment:

a. Go to Clusters Hue Configuration.

b. Filter by Category, Hue-service and Scope, Advanced.

c. Add support for a multi-threaded environment by setting Hue Service Advanced Configuration Snippet (Safety Valve) for hue_safety_valve.ini:

```
[desktop]
[[database]]
options={"threaded":true}
```

d. Click Save Changes.

4. Restart the Hue service: select Actions Restart and click Restart.

5. Log on to Hue by clicking Hue Web UI.

Existing CDH Installation

If you are using Oracle database with Hue and are upgrading to CDP 7.x from CDH 5 or CDH 6, then do the following:

Deactivate the Oracle Client Parcel

1. Log on to Cloudera Manager.


2.



Go to the Parcels page by clicking Hosts Parcels (or clicking the parcels icon ).

3. Click the ConfigurationCheck for New Parcels.





4. Find ORACLE_INSTANT_CLIENT and click Download, Distribute, and Deactivate.

Parcel Name	Version	Status	Actions
ORACLE_INSTANT_CLIENT	11.2-1.oracleinstantclient1.0.0.p0.130 	Distributed, Activated	Deactivate

Install Hue with Oracle database 12c and higher

1. Download the zip files for the [Instant Client Package](#), both Basic and SDK (with headers).

Version 12.2.0.1.0

Name	Download	Description
Instant Client Package (ZIP)	 instantclient-basic-linux.x64-12.2.0.1.0.zip	Basic: All files required to run OCI, OCCI, and JDBC-OCI applications (68,965,195 bytes) (cksum - 3923339140)
Instant Client Package (ZIP)	 oracle-instantclient12.2-basic-12.2.0.1.0-1.x86_64.rpm	Basic: All files required to run OCI, OCCI, and JDBC-OCI applications (52,826,628 bytes) (cksum - 888077889)
Name	Download	Description
Instant Client Package (ZIP)	 instantclient-sdk-linux.x64-12.2.0.1.0.zip	SDK: Additional header files and an example makefile for developing Oracle applications with Instant Client (674,743 bytes) (cksum - 2114815674)
Instant Client Package (RPM)	 oracle-instantclient12.2-devel-12.2.0.1.0-1.x86_64.rpm	SDK: Additional header files and an example makefile for developing Oracle applications with Instant Client (606,864 bytes) (cksum - 2680490862)



Note: If you are using Oracle database 11g, then download the corresponding 11g Instant Client Package from the Oracle website.

2. Switch to the host with the downloaded files and upload zip to the Hue server host:

```
scp instantclient-*.zip root@<hue server hostname>:.
```

3. Arrange the client libraries to mirror the tree structure in the image as shown in the following example:

```
# Create nested directories: /usr/share/oracle/instantclient/lib/
mkdir -pm 755 /usr/share/oracle/instantclient/lib
# Unzip. The files expand into /usr/share/oracle/instantclient/instantc
lient_<ver>/
unzip '*.zip' -d /usr/share/oracle/instantclient/

# Move lib files from instantclient_<ver> to /usr/share/oracle/instantcl
ient/lib/
mv /usr/share/oracle/instantclient/`ls -l /usr/share/oracle/instantclient/
| grep instantclient_ | awk '{print $9}'`/lib* /usr/share/oracle/instan
tclient/lib/

# Move rest of the files to /usr/share/oracle/instantclient/
mv /usr/share/oracle/instantclient/`ls -l /usr/share/oracle/instantclient
/ | grep instantclient_ | awk '{print $9}'`/* /usr/share/oracle/instantc
lient/

# Create symbolic links. Remember to edit version numbers as necessary
```

```
cd /usr/share/oracle/instantclient/lib
ln -s libclntsh.so.<ver>.1 libclntsh.so
ln -s libocci.so.<ver>.1 libocci.so
```

where <ver> is the version of the Instant Client Package. Replace <ver> with the actual version of the Instant Client Package.

4. Set the path for \$ORACLE_HOME and \$LD_LIBRARY_PATH as shown in the following example:

```
export ORACLE_HOME=/usr/share/oracle/instantclient
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$ORACLE_HOME
```

Connect Hue to Oracle

If you are not migrating the current (or old) database, simply connect to your new Oracle database and restart Hue.

1. [migration only] Stop Hue Service
 - a. In Cloudera Manager, navigate to ClusterHue.
 - b. Select Actions Stop.



Note: If necessary, refresh the page to ensure the Hue service is stopped:



2. [migration only] Dump Current Database
 - a. Select Actions Dump Database.
 - b. Click Dump Database. The file is written to /tmp/hue_database_dump.json on the host of the Hue server.
 - c. Log on to the host of the Hue server in a command-line terminal.
 - d. Edit /tmp/hue_database_dump.json by removing all objects with useradmin.userprofile in the model field. For example:

```
# Count number of objects
grep -c useradmin.userprofile /tmp/hue_database_dump.json
```

```
vi /tmp/hue_database_dump.json
```

```
{
  "pk": 1,
  "model": "useradmin.userprofile",
  "fields": {
    "last_activity": "2016-10-03T10:06:13",
    "creation_method": "HUE",
    "first_login": false,
    "user": 1,
    "home_directory": "/user/admin"
  }
},
{
  "pk": 2,
  "model": "useradmin.userprofile",
  "fields": {
    "last_activity": "2016-10-03T10:27:10",
    "creation_method": "HUE",
    "first_login": false,
    "user": 2,
    "home_directory": "/user/alice"
  }
},
}
```

3. Connect to New Database

a. Configure Database connections:

- Go to Hue Configuration and filter by category, Database.
- Set database properties and click Save Changes:

```
Hue Database Type (or engine): Oracle
Hue Database Hostname: <fqdn of host with Oracle server>
Hue Database Port: 1521
Hue Database Username: hue
Hue Database Password: <hue database password>
Hue Database Name (or SID): orcl
```

b. Add support for a multi-threaded environment:

- Filter by Category, Hue-service and Scope, Advanced.
- Set Hue Service Advanced Configuration Snippet (Safety Valve) for hue_safety_valve.ini and click Save Changes:

```
[desktop]
[[database]]
options={"threaded":true}
```

4. [migration only] Synchronize New Database

- Select Actions Synchronize Database
- Click Synchronize Database.

5. [migration only] Load Data from Old Database



Important: All user tables in the Hue database must be empty.

```
sqlplus hue/<your hue password> < delete_from_tables.ddl
```

6. Re/Start Hue service

- Navigate to ClusterHue.
- Select Actions Start, and click Start.
- Click Hue Web UI to log on to Hue with a custom Oracle database.

Configuring a database for Ranger

Additional steps to configure databases for Ranger.

After you have installed a database, use these steps to configure the database for Ranger . Ranger should use separate databases.

Configuring a Ranger Database: MySQL/MariaDB

Prior to upgrading your cluster to CDP Private Cloud Base you must configure the MySQL or MariaDB database instance for Ranger by creating a Ranger database and user. Before you begin the transition, review the support policies of database and admin policy support for transactions.

Before you begin

A supported version of MySQL or MariaDB must be running and available to be used by Ranger. See [Database Requirements](#).

When using MySQL or MariaDB, the storage engine used for the Ranger admin policy store tables **MUST** support transactions. InnoDB is an example of engine that supports transactions. A storage engine that does not support transactions is not suitable as a policy store.

Procedure

1. Log in to the host where you want to set up the MySQL database for Ranger.
2. Make sure you have the MYSQL connector version 5.7 or higher in the `/usr/share/java/` directory with name `mysql-connector-java.jar`.
3. Edit the following file: `/etc/my.cnf` and add the following line:

```
log_bin_trust_function_creators = 1
```

4. Restart the database:

```
systemctl restart mysqld
```

or:

```
systemctl restart mariadb
```

5. Log in to mysql:

```
mysql -u root
```

6. Run the following commands to create the Ranger database and user.

Substitute the following in the command:

- (optional) Replace `rangeradmin` with a username of your choice. Note this username, you will need to enter it later when running the Upgrade Cluster command.



Note: For Ranger KMS, use (for example) `rangerkms` rather than `rangeradmin`.

- (optional) Replace `cloudera` with a password of your choice. Note this password, you will need to enter it later when running the Upgrade Cluster command.
- `<Ranger Admin Role hostname>` – the name of the host where the Ranger Admin role will run. Note this host, you will need to enter it later when running the Upgrade Cluster command.

```
CREATE DATABASE ranger;
CREATE USER 'rangeradmin'@'%' IDENTIFIED BY 'cloudera';
CREATE USER 'rangeradmin'@'localhost' IDENTIFIED BY 'cloudera';
CREATE USER 'rangeradmin'@'<Ranger Admin Role hostname>' IDENTIFIED BY 'cloudera';
GRANT ALL PRIVILEGES ON ranger.* TO 'rangeradmin'@'%';
GRANT ALL PRIVILEGES ON ranger.* TO 'rangeradmin'@'localhost';
GRANT ALL PRIVILEGES ON ranger.* TO 'rangeradmin'@'<Ranger Admin Role hostname>';
FLUSH PRIVILEGES;
```

7. Use the `exit;` command to exit MySQL.
8. Test connecting to the database using the following command:

```
mysql -u rangeradmin -pcloudera
```

9. After testing the connection, use the `exit;` command to exit MySQL.
10. Continue with the cluster installation or upgrade to complete the transition.

Configuring a Ranger Database: Oracle

Prior to upgrading your cluster to CDP Private Cloud Base you must configure the Oracle database instance for Ranger by creating a Ranger database and user. Before you begin the transition, review the support policies of database and admin policy support for transactions.

Before you begin

A supported version of Oracle must be running and available to be used by Ranger. See [Database Requirements](#).

Procedure

1. On the Ranger host, install the appropriate JDBC .jar file.
 - a) Download the Oracle JDBC (OJDBC) driver from <https://www.oracle.com/technetwork/database/features/jdbc/index-091264.html>.
 - b) Copy the .jar file to the Java share directory.

```
sudo cp /tmp/ojdbc8-12.2.0.1.jar /usr/share/java/oracle-connector-java.jar
```

Make sure the .jar file has the appropriate permissions. For example:

```
sudo chmod 644 /usr/share/java/oracle-connector-java.jar
```

2. Log in to the host where the Oracle database is running and launch Oracle sqlplus:

```
sqlplus sys/root as sysdba
```

3. Create the Ranger database and user. Run the following commands:

```
# sqlplus sys/root as sysdba
CREATE USER rangeradmin IDENTIFIED BY rangeradmin;
GRANT SELECT_CATALOG_ROLE TO rangeradmin;
GRANT CONNECT, RESOURCE TO rangeradmin;
QUIT;
GRANT CREATE SESSION,CREATE PROCEDURE,CREATE TABLE,CREATE VIEW,CREATE SEQUENCE,CREATE PUBLIC SYNONYM,CREATE ANY SYNONYM,CREATE TRIGGER,UNLIMITED TABLESPACE TO rangeradmin;
ALTER USER rangeradmin DEFAULT TABLESPACE <tablespace>;
ALTER USER rangeradmin quota unlimited on <tablespace>;
```



Note: For Ranger KMS, use rangerkms rather than rangeradmin.

What to do next

Continue installing or upgrading your cluster.

Configuring a PostgreSQL Database for Ranger

Complete the following steps to configure a PostgreSQL database instance for Ranger or Ranger KMS.

Configuring a PostgreSQL Database for Ranger on RHEL7/Centos7

Before you begin



Important: Ranger and Ranger KMS should use separate databases.

Procedure

1. Run the following command to install PostgreSQL server:

```
sudo yum install postgresql-server
```

2. Initialize the Postgres database and start PostgreSQL:

```
sudo postgresql-setup initdb
sudo systemctl start postgresql
```

3. Optional: Configure PostgreSQL to start on boot:

```
sudo systemctl enable postgresql
```

4. Update the postgresql.conf file, which is usually found in /var/lib/pgsql/data or /var/lib/postgresql/data:

- Uncomment and change #listen_addresses = 'localhost' to listen_addresses = '*'
- Uncomment the #port = line and specify the port number (the default is 5432)
- Optional: Uncomment and change #standard_conforming_strings= to standard_conforming_strings = off

5. Update the pg_hba.conf file, which is usually found in /var/lib/pgsql/data or /etc/postgresql/<version>/main:

- Add the following line to allow connection to the Ranger database from any host:

```
host      ranger          rangeradmin    0.0.0.0/0          md5
```



Note: For Ranger KMS, use rangerkms rather than rangeradmin.

6. Restart PostgreSQL:

```
sudo systemctl restart postgresql
```

7. The PostgreSQL database administrator should be used to create the Ranger databases. The following series of commands could be used to create the rangeradmin user and grant it adequate privileges. Be sure to replace 'password' with a strong password.

```
echo "CREATE DATABASE ranger;" | sudo -u postgres psql -U postgres
echo "CREATE USER rangeradmin WITH PASSWORD 'password';" | sudo -u postgres psql -U postgres
echo "GRANT ALL PRIVILEGES ON DATABASE ranger TO rangeradmin;" | sudo -u postgres psql -U postgres
```



Note: For Ranger KMS, use rangerkms rather than rangeradmin.

8. Install the PostgreSQL JDBC driver. If you would like to use the PostgreSQL JDBC driver version shipped with the OS repositories, run the following command:

```
yum install postgresql-jdbc*
```

You can also download the JDBC driver from the official PostgreSQL JDBC Driver website – <https://jdbc.postgresql.org/>.

9. Rename the Postgres JDBC driver .jar file to postgresql-connector-java.jar and copy it to the /usr/share/java directory. The following copy command can be used if the Postgres JDBC driver .jar file is installed from the OS repositories:

```
cp /usr/share/java/postgresql-jdbc.jar /usr/share/java/postgresql-connector-java.jar
```

10. Confirm that the .jar file is in the Java share directory:

```
ls /usr/share/java/postgresql-connector-java.jar
```


11. Change the access mode of the .jar file to 644:

```
chmod 644 /usr/share/java/postgresql-connector-java.jar
```

What to do next

Ensure that the Ranger Solr and Ranger HDFS plugins are enabled. See [Additional Steps for Apache Ranger](#) on page 117 for details.

Configuring the Database for Streaming Components

Additional steps to configure the databases for Schema Registry and Streams Messaging Manager (SMM).

Configure PostgreSQL for Streaming Components

If you are installing Schema Registry or Streams Messaging Manager (SMM), you must configure the database to store metadata.

About this task

After you install PostgreSQL, configure the database to store:

- Schema Registry data such as the schemas and their metadata, all the versions and branches.
- SMM data such as Kafka metadata, stores metrics, and alert definitions.



Important: For the Schema Registry database, you must set collation to be case sensitive.

Procedure

1. Log in to Postgres:

```
sudo su postgres  
psql
```

2. For the Schema Registry metadata store, create a database called registry with the password registry:

```
create database registry;  
CREATE USER registry WITH PASSWORD 'registry';  
GRANT ALL PRIVILEGES ON DATABASE "registry" to registry;
```

3. For the SMM metadata store, create a database called streamsmgmr with the password streamsmgmr:

```
create database streamsmgmr;  
CREATE USER streamsmgmr WITH PASSWORD 'streamsmgmr';  
GRANT ALL PRIVILEGES ON DATABASE "streamsmgmr" to streamsmgmr;
```

If you cannot grant all privileges, grant the following privileges that SMM and Schema Registry require at a minimum:

- CREATE/ALTER/DROP TABLE
- CREATE/ALTER/DROP INDEX
- CREATE/ALTER/DROP SEQUENCE
- CREATE/ALTER/DROP PROCEDURE

For example:

```
grant create session to streamsmgmr;
```

```
grant create table to streamsmgmr;
grant create sequence to streamsmgmr;
```

Configuring MySQL for Streaming Components

If you intend to use MySQL to store the metadata for Streams Messaging Manager or Schema Registry, you must configure the MySQL database.

About this task

Configure the database to store:

- In Schema Registry, the schemas and their metadata, all the versions and branches.
- In SMM, the Kafka metadata, stores metrics, and alert definitions.



Important: For the Schema Registry database, you must set collation to be case sensitive.

Procedure

1. Log in to the host.

- a) Run the following command for Schema Registry:

```
ssh [MY_SCHEMA_REGISTRY_HOST]
```

- b) Run the following command for Streams Messaging Manager:

```
ssh [MY_STREAMS_MESSAGING_MANAGER_HOST]
```

2. Launch the MySQL monitor:

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

3. Create the database for the Schema Registry and the SMM metastore:

```
create database registry;
create database streamsmgmr;
```

4. Create Schema Registry and SMM user accounts, replacing the final IDENTIFIED BY string with your password:

```
CREATE USER 'registry'@'%' IDENTIFIED BY 'R12$%34qw';
CREATE USER 'streamsmgmr'@'%' IDENTIFIED BY 'R12$%34qw';
```

5. Assign privileges to the user account:

```
GRANT ALL PRIVILEGES ON registry.* TO 'registry'@'%' WITH GRANT OPTION ;
GRANT ALL PRIVILEGES ON streamsmgmr.* TO 'streamsmgmr'@'%' WITH GRANT OPTION ;
```

If you cannot grant all privileges, grant the following privileges that SMM and Schema Registry require at a minimum:

- CREATE/ALTER/DROP TABLE
- CREATE/ALTER/DROP INDEX
- CREATE/ALTER/DROP SEQUENCE

- CREATE/ALTER/DROP PROCEDURE

For example:

```
grant create session to streamsmgmr;
grant create table to streamsmgmr;
grant create sequence to streamsmgmr;
```

6. Commit the operation:

```
commit;
```

Step 5: Set up the Cloudera Manager Database

Cloudera Manager Server includes a script that can create and configure a database for itself.

The script can:

- Create the Cloudera Manager Server database configuration file.
- (PostgreSQL) Create and configure a database for Cloudera Manager Server to use.
- (PostgreSQL) Create and configure a user account for Cloudera Manager Server.

Although the script can create a database, the following procedures assume that you have already created the database as described in *Step 4: Install and Configure Databases*.

The following sections describe the syntax for the script and demonstrate how to use it:

Syntax for scm_prepare_database.sh

Review the syntax of the scm_prepare_database.sh script before you run it to configure the Cloudera Manager database.

The syntax for the scm_prepare_database.sh script is as follows:

```
sudo /opt/cloudera/cm/schema/scm_prepare_database.sh [option
s] <databaseType> <databaseName> <databaseUser> <password>
```



Note: You can also run scm_prepare_database.sh without options to see the syntax.

To create a new database, you must specify the -u and -p parameters for a user with privileges to create databases. If you have already created the database as instructed in *Step 4: Install and Configure Databases*, do not specify these options.

The following tables describe the parameters and options for the scm_prepare_database.sh script:

Table 31: Parameters

Parameter (Required in bold)	Description
<databaseType>	One of the supported database types: <ul style="list-style-type: none"> • PostgreSQL: postgresql • MariaDB: mysql • MySQL: mysql • Oracle: oracle
<databaseName>	The name of the Cloudera Manager Server database to use. For PostgreSQL databases, the script can create the specified database if you specify the -u and -p options with the credentials of a user that has privileges to create databases and grant privileges. The default database name provided in the Cloudera Manager configuration settings is scm, but you are not required to use it.

Parameter (Required in bold)	Description
<databaseUser>	The username for the Cloudera Manager Server database to create or use. The default username provided in the Cloudera Manager configuration settings is scm, but you are not required to use it.
<password>	<p>The password for the <databaseUser> to create or use. If you do not want the password visible on the screen or stored in the command history, do not specify the password, and you are prompted to enter it as follows:</p> <pre>Enter SCM password:</pre>

Table 32: Options

Option	Description
-? --help	Display help.
--config-path	The path to the Cloudera Manager Server configuration files. The default is /etc/cloudera-scm-server.
-f --force	If specified, the script does not stop if an error occurs.
-h --host	The IP address or hostname of the host where the database is installed. The default is to use localhost.
-p --password	<p>The admin password for the database application. Use with the -u option. The default is no password. Do not put a space between -p and the password (for example, -phunter2). If you do not want the password visible on the screen or stored in the command history, use the -p option without specifying a password, and you are prompted to enter it as follows:</p> <pre>Enter database password:</pre> <p>If you have already created the database, do not use this option.</p>
-P --port	<p>The port number to use to connect to the database. The default port is:</p> <ul style="list-style-type: none"> PostgreSQL: 5432 MariaDB: 3306 MySQL: 3306 Oracle: 1521 <p>This option is used for a remote connection only.</p>
--scm-host	The hostname where the Cloudera Manager Server is installed. If the Cloudera Manager Server and the database are installed on the same host, do not use this option or the -h option.
--scm-password-script	A script to execute whose stdout provides the password for user SCM (for the database).
-u --user	The admin username for the database application. Use with the -p option. Do not put a space between -u and the username (for example, -uroot). If this option is supplied, the script creates a user and database for the Cloudera Manager Server. If you have already created the database, do not use this option.

The following examples demonstrate the syntax and output of the scm_prepare_database.sh script for different scenarios:

Example 1: Running the script when MySQL or MariaDB is co-located with the Cloudera Manager Server

This example assumes that you have already created the Cloudera Management Server database and database user, naming both scm:

```
sudo /opt/cloudera/cm/schema/scm_prepare_database.sh mysql scm scm
```

```
Enter SCM password:
JAVA_HOME=/usr/java/jdk1.8.0_141-cloudera
Verifying that we can write to /etc/cloudera-scm-server
Creating SCM configuration file in /etc/cloudera-scm-server
Executing: /usr/java/jdk1.8.0_141-cloudera/bin/java -cp /usr/share/java/m
ysql-connector-java.jar:/usr/share/java/oracle-connector-java.jar:/usr/share
```

```
/java/postgresql-connector-java.jar:/opt/cloudera/cm/schema/./lib/* com.cloudera.enterprise.dbutil.DbCommandExecutor /etc/cloudera-scm-server/db.properties com.cloudera.cmf.db.
[ main] DbCommandExecutor INFO Successfully connected to database.
All done, your SCM database is configured correctly!
```

Example 2: Running the script when MySQL or MariaDB is installed on another host

This example demonstrates how to run the script on the Cloudera Manager Server host (cm01.example.com) and connect to a remote MySQL or MariaDB host (db01.example.com):

```
sudo /opt/cloudera/cm/schema/scm_prepare_database.sh mysql -h db01.example.com --scm-host cm01.example.com scm scm
```

```
Enter database password:
JAVA_HOME=/usr/java/jdk1.8.0_141-cloudera
Verifying that we can write to /etc/cloudera-scm-server
Creating SCM configuration file in /etc/cloudera-scm-server
Executing: /usr/java/jdk1.8.0_141-cloudera/bin/java -cp /usr/share/java/mysql-connector-java.jar:/usr/share/java/oracle-connector-java.jar:/usr/share/java/postgresql-connector-java.jar:/opt/cloudera/cm/schema/./lib/* com.cloudera.enterprise.dbutil.DbCommandExecutor /etc/cloudera-scm-server/db.properties com.cloudera.cmf.db.
[ main] DbCommandExecutor INFO Successfully connected to database.
All done, your SCM database is configured correctly!
```

Example 3: Running the script to configure Oracle

```
sudo /opt/cloudera/cm/schema/scm_prepare_database.sh -h cm-oracle.example.com oracle orcl sample_user sample_pass
```

```
JAVA_HOME=/usr/java/jdk1.8.0_141-cloudera
Verifying that we can write to /etc/cloudera-scm-server
Creating SCM configuration file in /etc/cloudera-scm-server
Executing: /usr/java/jdk1.8.0_141-cloudera/bin/java -cp /usr/share/java/mysql-connector-java.jar:/usr/share/java/oracle-connector-java.jar:/usr/share/java/postgresql-connector-java.jar:/opt/cloudera/cm/schema/./lib/*cloudera.enterprise.dbutil.DbCommandExecutor /etc/cloudera-scm-server/db.properties com.cloudera.cmf.db.
[ main] DbCommandExecutor INFO Successfully connected to database.
All done, your SCM database is configured correctly!
```

Step 6: Install Runtime and Other Software

After you set up the Cloudera Manager database, start Cloudera Manager Server and log in to the Cloudera Manager Admin Console. Then proceed through the installation wizard.

Procedure

1. Start Cloudera Manager Server:

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

2. If you want to configure the Cloudera Manager server to start automatically when the host reboots, run the following command:

```
sudo systemctl enable cloudera-scm-server
```

3. Wait several minutes for the Cloudera Manager Server to start. To observe the startup process, run the following on the Cloudera Manager Server host:

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

When you see this log entry, the Cloudera Manager Admin Console is ready:

```
INFO WebServerImpl:com.cloudera.server.cmf.WebServerImpl: Started Jetty server.
```

If the Cloudera Manager Server does not start, see *Troubleshooting Installation Problems*.

4. 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 you enabled auto-TLS, you are redirected to `https://<server_host>:7183`, and a security warning is displayed. You might need to indicate that you trust the certificate, or click to proceed to the Cloudera Manager Server host.

5. Log into Cloudera Manager Admin Console. The default credentials are:

Username: admin

Password: admin



Note: Cloudera Manager does not support changing the admin username for the installed account. You can change the password using Cloudera Manager after you run the installation wizard. Although you cannot change the admin username, you can add a new user, assign administrative privileges to the new user, and then delete the default admin account.

Results

After logging in, the installation wizard launches. The following sections guide you through each step of the installation wizard.

Installation Wizard

Proceed through the installation wizard to accept licenses, install and configure Cloudera Runtime, and more.

Upload License File

On the Upload License File page, you can select either the trial version of CDP Data Center or upload a license file:

1. Choose one of the following options:
 - Upload Cloudera Data Platform License
 - Try Cloudera Data Platform for 60 days. The CDP Data Center trial does not require a license file, but the trial expires after 60 days.
2. If you choose the CDP Data Center Edition Trial, you can upload a license file at a later time. Read the license agreement and click the checkbox labeled Yes, I accept the Cloudera Standard License Terms and Conditions if you accept the terms and conditions of the license agreement. Then click Continue.
3. If you have a license file for CDP Data Center, upload the license file:
 - a. Select Upload Cloudera Data Platform License.
 - b. Click Upload License File.
 - c. Browse to the location of the license file, select the file, and click Open.
 - d. Click Upload.
 - e. Click Continue.

- Click Continue to proceed with the installation.

The Welcome page displays.

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.

Cluster Basics

The Cluster Basics page allows you to specify the Cluster Name

For new installations, a Regular Cluster (also called a base cluster) is the only option. You can add a compute cluster after you finish installing the base cluster.

For more information on regular and compute clusters, and data contexts, see [Virtual Private Clusters and Cloudera SDX](#).

Enter a cluster name and click Continue.

Setup Auto-TLS

The Setup Auto-TLS page provides instructions for initializing the certificate manager for auto-TLS if you have not done so already. If you already initialized the certificate manager in *Step 3: Install Cloudera Manager Server*, the wizard displays a message indicating that auto-TLS has been initialized. Click Continue to proceed with the installation.

If you have not already initialized the certificate manager, and you want to enable auto-TLS, follow the instructions provided on the page before continuing. When you reload the page as instructed, you are redirected to `http s://<server_host>:7183`, and a security warning is displayed. You might need to indicate that you trust the certificate, or click to proceed to the Cloudera Manager Server host. You might also be required to log in again and re-complete the previous steps in the wizard.

If you do not want to enable auto-TLS at this time, click Continue to proceed.

Specify Hosts

Choose which hosts will run Runtime and other managed services.



Note: If you have enabled Auto-TLS, you must include the Cloudera Manager server host when you specify hosts.

- To enable Cloudera Manager to automatically discover hosts on which to install Runtime and managed services, enter the cluster hostnames or IP addresses in the Hostnames field. You can specify hostname and IP address ranges as follows:

Expansion Range	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].example.com	host1.example.com, host2.example.com, host3.example.com
host[07-10].example.com	host07.example.com, host08.example.com, host09.example.com, host10.example.com



Important: Unqualified hostnames (short names) must be unique in a Cloudera Manager instance. For example, you cannot have both `host01.example.com` and `host01.standby.example.com` managed by the same Cloudera Manager Server.

You can specify multiple addresses and address ranges by separating them with commas, semicolons, tabs, or blank spaces, or by placing them on separate lines. Use this technique to make more specific searches instead of searching overly wide ranges. Only scans that reach hosts running SSH will be selected for inclusion in your

cluster by default. You can enter an address range that spans over unused addresses and then clear the nonexistent hosts later in the procedure, but wider ranges require more time to scan.

2. Click Search. If there are a large number of hosts on your cluster, wait a few moments to allow them to be discovered and shown in the wizard. If the search is taking too long, you can stop the scan by clicking Abort Scan. You can modify the search pattern and repeat the search as many times as you need until you see all of the expected hosts.



Note: Cloudera Manager scans hosts by checking for network connectivity. If there are some hosts where you want to install services that are not shown in the list, make sure you have network connectivity between the Cloudera Manager Server host and those hosts, and that firewalls and SELinux are not blocking access.

3. Verify that the number of hosts shown matches the number of hosts where you want to install services. Clear host entries that do not exist or where you do not want to install services.
4. Click Continue.

The Select Repository screen displays.

Select Repository



Important: You cannot install software using both parcels and packages in the same cluster.

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*.

In the CDH and other software section:

1. Select the repository type to use for the installation. In the Install Method section select one of the following:
 - Use Parcels (Recommended)

A parcel is a binary distribution format containing the program files, along with additional metadata used by Cloudera Manager. Parcels are required for rolling upgrades. For more information, see *Parcels*.
2. Select the version of Cloudera Runtime or CDH to install. If you do not see the version you want to install:
 - Parcels – Click the Parcel Repository & Network Settings link to add the repository URL for your version. If you are using a local Parcel repository, enter its URL as the repository URL.

Repository URLs for CDH 6 parcels are documented in [CDH 6 Download Information](#)

Repository URLs for the Cloudera Runtime 7 parcels are documented in [Cloudera Runtime Download Information](#)



Important: If you are installing Cloudera Runtime 7.1.5.0 and you have selected to use a 60-day trial license, use the following Parcel Repository URL:

```
https://archive.cloudera.com/cdh7/7.1.5.0/parcels/
```

After adding the repository, click Save Changes and wait a few seconds for the version to appear. If your Cloudera Manager host uses an HTTP proxy, click the Proxy Settings button to configure your proxy.

Note that if you have a Cloudera Enterprise license and are using Cloudera Manager 6.3.3 or higher to install a CDH version 6.3.3 or higher, or a Cloudera Runtime version 7.0 or higher using parcels, you do not need to add a username and password or "@" to the parcel repository URL. Cloudera Manager will authenticate to the

Cloudera archive using the information in your license key file. Use a link to the repository in the following format:

```
https://archive.cloudera.com/p/cdh6/6.x.x/parcels/
```

If you are using a version of CM older than 6.3.3 to install CDH 6.3.3 or higher parcels, you must include the username/password and "@" in the repository URL during installation or when you configure a CDH 6.3.3 or higher parcel repository. After you add the repository, click Save Changes and wait a few seconds for the version to appear. If your Cloudera Manager host uses an HTTP proxy, click the Proxy Settings button to configure your proxy.



Note: Cloudera Manager only displays CDH versions it can support. If an available CDH version is too new for your Cloudera Manager version, it is not displayed. If the parcels do not appear on the Parcels page, ensure that the Parcel URL you entered is correct.



Note: Cloudera Manager only displays Cloudera Runtime versions it can support. If an available CDH version is too new for your Cloudera Manager version, it is not displayed.

3. If you selected Use Parcels, specify any Additional Parcels you want to install.
4. Click Continue.

Select JDK



Note: CDP Data Center is no longer bundled with Oracle JDK software. Cloudera provides a supported version of OpenJDK.

If you installed your own JDK version, such as Oracle JDK 8, in *Step 2: Install Java Development Kit*, select Manually manage JDK.

To allow Cloudera Manager to automatically install the OpenJDK on cluster hosts, select Install a Cloudera-provided version of OpenJDK.

To install the default OpenJDK that is provided by your operating system, select Install a system-provided version of OpenJDK.

After checking the applicable boxes, click Continue.

Enter Login Credentials

1. Select root for the root account, or select Another user and enter the username for an account that has password-less sudo privileges.
2. Select an authentication method:
 - If you choose password authentication, enter and confirm the password.
 - If you choose public-key authentication, provide a passphrase and path to the required key files.

You can modify the default SSH port if necessary.

3. Specify the maximum number of host installations to run at once. The default and recommended value is 10. You can adjust this based on your network capacity.
4. Click Continue.

The Install Agents page displays.

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.

If you selected the option to manually install agents, see *Manually Install Cloudera Manager Agent Packages* for the procedure and then continue with the next steps on this page.

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

If you are using parcels, the Install Parcels page displays. If you chose to install using packages, the Inspect Cluster page displays.

Install Parcels

If you selected parcels for the installation method, the Install Parcels page reports the installation progress of the parcels you selected earlier. 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.

After the installation is complete, click Continue.

The Inspect Cluster page displays.

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. You can click Advanced Options to customize some ping parameters.
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, and you want to abandon the cluster creation wizard to address them, select the radio button labeled Quit the wizard and Cloudera Manager will delete the temporarily created cluster and then click Continue.

Otherwise, after addressing any identified problems, 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.

Continue to *Step 7: Set Up a Cluster Using the Wizard*.

Step 7: Set Up a Cluster Using the Wizard

After you complete the Add Cluster - Installation wizard, the Add Cluster - Configuration wizard automatically starts. The following sections guide you through each page of the wizard.

Select Services

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

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

Choose one of the following:

Regular (Base) Clusters

Data Engineering

Process develop, and serve predictive models.

Services included: HDFS, YARN, YARN Queue Manager, Ranger, Atlas, Hive, Hive on Tez, Spark, Oozie, Hue, and Data Analytics Studio

Data Mart

Browse, query, and explore your data in an interactive way.

Services included: HDFS, Ranger, Atlas, Hive, and Hue

Operational Database

Real-time insights for modern data-driven business.

Services included: HDFS, Ranger, Atlas, and HBase

Custom Services

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

Compute Clusters

Data Engineering

Process develop, and serve predictive models.

Services included: Spark, Oozie, Hive on Tez, Data Analytics Studio, HDFS, YARN, and YARN Queue Manager

Spark

Spark for Compute

Services included: Core Configuration, Spark, Oozie, YARN, and YARN Queue Manager

Data Mart

Impala for Compute

Services included: Core Configuration, Impala, and Hue

Streams Messaging (Simple)

Simple Kafka cluster for streams messaging

Services included: Kafka, Schema Registry, and Zookeeper

Streams Messaging (Full)

Advanced Kafka cluster with monitoring and replication services for streams messaging

Services included: Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager, Cruise Control, and Zookeeper

Custom Services

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

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.

After assigning all of the roles for your services, click Continue. The Setup Database page displays.

Setup Database

On the Setup Database page, you can enter the database hosts, names, usernames, and passwords you created in *Step 4: Install and Configure Databases*.

For services that support it, you can add finer-grained customizations using a JDBC URL override.



Important: The Hive service is currently the only service that supports the JDBC URL override.

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

For MariaDB, select MySQL.

For services that support it, to specify a JDBC URL override, select Yes in the Use JDBC URL Override dropdown menu. You must also specify the database type, username, and password.

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. The Review Changes page displays.

Enter Required Parameters

The **Enter Required Parameters** page lists required parameters for DAS, the Cloudera Manager API client, Hive, and Ranger.

Atlas

The Atlas Admin user, Ranger Admin user, Usersync user, Tagsync user, and KMS Keyadmin user are created during cluster deployment. In this page you must give a password for each of these users.



Note: Passwords for the Atlas Admin, Ranger Admin, Usersync, Tagsync, and KMS Keyadmin users must be a minimum of 8 characters long, with at least one alphabetic and one numeric character. The following characters are not valid: " ' \ ` ' .

Cloudera Manager API Client

If you do not have an existing user for the Cloudera Manager API client, use the default username and password "admin" for both the The Existing Cloudera Manager API Client Username and The Existing Cloudera Manager API Client Password.

DAS

The DAS database hostname, database name, database username, and database password were configured when you created the required DAS database. The default database name is "das" and the default database user is "das".

Hive

If your database supports TLS connections, then configure the following parameters:

- Enable TLS/SSL to the Hive Metastore Database parameter,
- Set the Hive Metastore Client SSL/TLS Trust Store File parameter to a JKS truststore file that contains a CA certificate trusting the database's certificate.
- Set the Hive Metastore Client SSL/TLS Trust Store Password parameter to that truststore's password.

Ranger

The Ranger database host, name, user, and user password were configured when you created the required Ranger database. If you ran the `gen_embedded_ranger_db.sh` script to create the Ranger database, the output of the script contained the host and database user password. Enter those here. The default database name is "ranger" and the default database user is "rangeradmin."

Review Changes

The Review Changes page lists default and suggested settings for several configuration parameters, including data directories.



Warning: Do not place DataNode data directories on NAS devices. When resizing an NAS, block replicas can be deleted, which results in missing blocks.

Review and make any necessary changes, and then click Continue. The Command Details page displays.

Command Details

The Command Details page lists the details of the First Run command.

You can expand the running commands to view the details of any step, including log files and command output. You can filter the view by selecting Show All Steps, Show Only Failed Steps, or Show Only Running Steps.

After the First Run command completes, click Continue to go to the Summary page.

If cluster deployment fails, be sure to click Resume in the wizard after you fix any issues. If you do not click Resume, the Ranger service will not enable all of the necessary plugins.

Summary

The Summary page reports the success or failure of the setup wizard.

Click Finish to complete the wizard. The installation is complete.

Cloudera recommends that you change the default password as soon as possible by clicking the logged-in username at the top right of the home screen and clicking Change Password.

Additional Steps for Apache Ranger

After installing Cloudera Manager and adding a cluster, there are additional steps required to complete the installation of Apache Ranger.

Related Information

[Configure a resource-based policy: Solr](#)

[Enabling Solr clients to authenticate with a secure Solr](#)

Enable Plugins

About this task

The Ranger plugins for HDFS and Solr may not be enabled by default. Ranger plugins enable Cloudera Manager stack components – such as HDFS and Solr – to connect to Ranger and access its authorization and audit services. Verify that the HDFS and Solr plugins are enabled after you install and start the Ranger service.

Procedure

1. To enable the HDFS plugin:
 - a) Login to Cloudera Manager.
 - b) Go to the HDFS Service status page.
 - c) Click the Configuration tab.
 - d) Search for the Enable Ranger Authorization configuration property.
 - e) If the Enable Ranger Authorization property is not selected, select it and save the changes.
 - f) Go to the Ranger Service status page and click ActionsSetup Ranger Plugin Service.
 - g) Restart the HDFS service.
2. To enable the Ranger Solr plugin:
 - a) Login to Cloudera Manager.
 - b) Go to the Solr Service status page.
 - c) Click the Configuration tab.
 - d) Search for the Enable Ranger Authorization configuration property.
 - e) If the Enable Ranger Authorization property is not selected, select it and save the changes.



Note: Don't select the Ranger Service dependency parameter. This is used for enabling a Solr service instance that is not used by the Ranger service.

- f) Restart the Solr service.

Add Solr WebUI Users

Procedure

Add the username of any users to the Ranger Solr policy who should have access to the Solr Web UI in the Ranger Policy for Solr. The user should have full access privileges.

Update the Time-to-live configuration for Ranger Audits

Procedure

1. Download the Ranger audits configurations to your SolrServer or Solr gateway host, by running the following command on the host:

```
solrctl instancedir --get ranger_audits /tmp/ranger_audits
```

2. Open the following file in a text editor:

```
tmp/ranger_audits/conf/solrconfig.xml
```

3. Edit the TTL section in this file to change the value of the following parameter to the appropriate value (the default value is 90 days):

```
<str name="fieldName">ttl</str>
<str name="value">+90DAYS</str>
```

4. Upload the new configuration by running the following command on the host:

```
solrctl --jaas [***solr-jaas.conf***] instancedir --update ranger_audits /
tmp/ranger_audits
```

For information on creating a jaas.conf file, see *Enabling Solr clients to authenticate with a secure Solr*.

5. Reload the Ranger_audits collection with the Solr credentials so that the collection can pick up the modified configuration by running the following command:

```
solrctl collection --reload ranger_audits
```

What to do next

1. Verify Ranger Configurations

- Verify that the username of any users who should have access to the Solr Web UI to the Ranger policy for Solr has been added to the Ranger Policy for Sol. The user should have full access privileges.
- Verify that the Time-to-live value is set appropriately by examining this file on the SolrServer or Solr gateway host:

- a. Download the configuration:

```
solrctl instancedir --get ranger_audits /tmp/ranger_audits
```

- b. Open the tmp/ranger_audits/conf/solrconfig.xml file and examine the ttl parameter (identified by: `<str name="fieldName">ttl</str>`).
- c. If you need to change the value, edit the file and then reload the configuration by running the following command:

```
solrctl collection --reload ranger_audits
```

Installing Apache Knox

This document provides instructions on how to install Apache Knox using the CDP Private Cloud Base installation process.

About this task

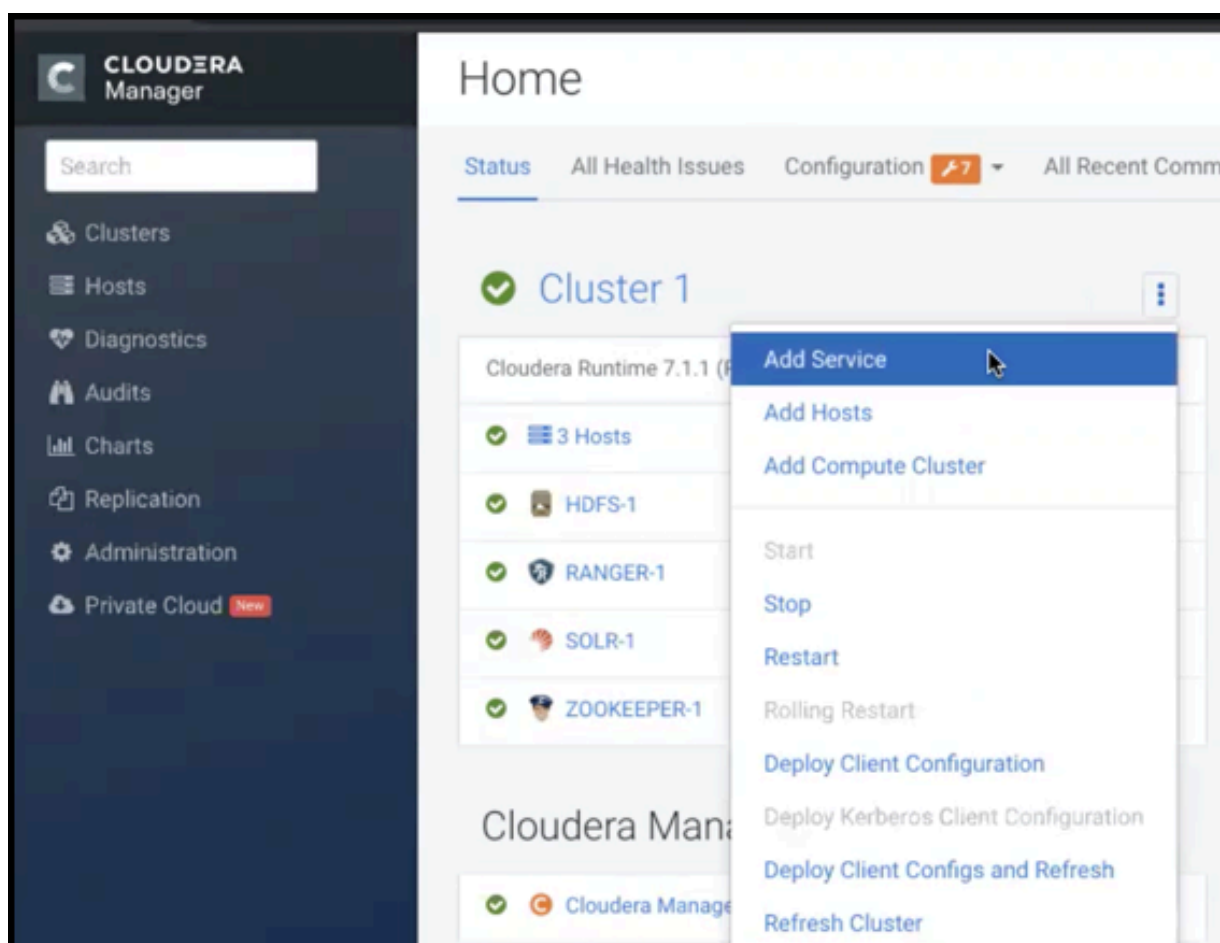
Apache Knox is an application gateway for interacting with the REST APIs and UIs. The Knox Gateway provides a single access point for all REST and HTTP interactions in your Cloudera Data Platform cluster.

Before you begin

When installing Knox, you must have Kerberos enabled on your cluster.

Procedure

1. From your Cloudera Manager homepage, go to Status tab \$Cluster Name ... Add Service



2. From the list of services, select Knox and click Continue.
3. On the **Select Dependencies** page, choose the dependencies you want Knox to set up:

HDFS, Ranger, Solr, Zookeeper

For users that require Apache Ranger for authorization. HDFS with Ranger. HDFS depends on Zookeeper, and Ranger depends on Solr.

HDFS, Zookeeper

HDFS depends on Zookeeper.

No optional dependencies

For users that do not wish to have Knox integrate with HDFS or Ranger.

4. On the **Assign Roles** page, select role assignments for your dependencies and click Continue:

Knox service roles	Description	Required?
Knox Gateway	If Knox is installed, at least one instance of this role should be installed. This role represents the Knox Gateway which provides a single access point for all REST and HTTP interactions with Apache Hadoop clusters.	Required
KnoxIDBroker*	It is strongly recommended that this role is installed on its own dedicated host. As its name suggests this role will allow you to take advantage of Knox's Identity Broker capabilities, an identity federation solution that exchanges cluster authentication for temporary cloud credentials.*	Optional*
Gateway	This role comes with the CSD framework. The gateway structure is used to describe the client configuration of the service on each host where the gateway role is installed.	Optional

* Note: KnoxIDBroker appears in the Assign Roles page, but it is not currently supported in CDP Private Cloud.

5. On the **Review Changes** page, most of the default values are acceptable, but you must Enable Kerberos Authentication and supply the Knox Master Secret. There are additional parameters you can specify or change, listed in “Knox Install Role Parameters”.
- Click Enable Kerberos Authentication
Kerberos is required where Knox is enabled.
 - Supply the Knox Master Secret, e.g. knoxsecret.
 - Click Continue.
6. The **Command Details** page shows the status of your operation. After completion, your system admin can view logs for your installation under stdout.

Related Information

[Apache Knox Install Role Parameters](#)

Apache Knox Install Role Parameters

Reference information on all the parameters available for Knox service roles.

Service-level parameters

Table 33: Required service-level parameters

Name	In Wizard	Type	Default Value
kerberos.auth.enabled*	Yes	Boolean	false
ranger_knox_plugin_hdfs_audit_directory	No	Text	\${ranger_base_audit_url}/knox
autorestart_on_stop	No	Boolean	false
knox_pam_realm_service	No	Text	login
save_alias_command_input_password	No	Text	-

Knox Gateway role parameters

Table 34: Required parameters for Knox Gateway role

Name	In Wizard	Type	Default Value
gateway_master_secret	Yes	Password	-
gateway_conf_dir	Yes	Path	/var/lib/knox/gateway/conf
gateway_data_dir	Yes	Path	/var/lib/knox/gateway/data
gateway_port	No	Port	8443
gateway_path	No	Text	gateway
gateway_heap_size	No	Memory	1 GB (min = 256 MB; soft min = 512 MB)
gateway_ranger_knox_plugin_conf_path	No	Path	/var/lib/knox/ranger-knox-plugin
gateway_ranger_knox_plugin_policy_cache_directory	No	Path	/var/lib/ranger/knox/gateway/policy-cache
gateway_ranger_knox_plugin_hdfs_audit_spool_directory	No	Path	/var/log/knox/gateway/audit/hdfs/spool
gateway_ranger_knox_plugin_solr_audit_spool_directory	No	Path	/var/log/knox/gateway/audit/solr/spool

Table 35: Optional parameters for Knox Gateway role

Name	Type	Default Value
gateway_default_topology_name	Text	cdp-proxy
gateway_auto_discovery_enabled	Boolean	true
gateway_cluster_configuration_monitor_interval	Time	60 seconds (minimum = 30 seconds)
gateway_auto_discovery_advanced_configuration_monitor_interval	Time	10 seconds (minimum = 5 seconds)
gateway_cloudera_manager_descriptors_monitor_interval	Time	10 seconds (minimum = 5 seconds)
gateway_auto_discovery_cdp_proxy_enabled_*	Boolean	true
gateway_auto_discovery_cdp_proxy_api_enabled_*	Boolean	true
gateway_descriptor_cdp_proxy	Text Array	Contains the required properties of cdp-proxy topology
gateway_descriptor_cdp_proxy_api	Text Array	Contains the required properties of cdp-proxy-api topology
gateway_sso_authentication_provider	Text Array	Contains the required properties of the authentication provider used by the UIs using the Knox SSO capabilities (Admin UI and Home Page). Defaults to PAM authentication.
gateway_api_authentication_provider	Text Array	Contains the required properties of the authentication provider used by pre-defined topologies such as admin, metadata or cdp-proxy-api. Defaults to PAM authentication.

Knox IDBroker role parameters



Note: Knox IDBroker is not currently supported in CDP Private Cloud.

Table 36: Required parameters for Knox IDBroker role

Name	In Wizard	Type	Default Value
idbroker_master_secret	Yes	Password	-
idbroker_conf_dir	Yes	Path	/var/lib/knox/idbroker/conf
idbroker_data_dir	Yes	Path	/var/lib/knox/idbroker/data
idbroker_gateway_port	No	Port	8444
idbroker_gateway_path	No	Text	gateway
idbroker_heap_size	No	Memory	1 GB (min = 256 MB; soft min = 512 MB)

Table 37: Optional parameters for Knox IDBroker role

Name	Type	Default Value
idbroker_aws_user_mapping	Text	-
idbroker_aws_group_mapping	Text	-
idbroker_aws_user_default_group_mapping	Text	-
idbroker_aws_credentials_key	Password	-
idbroker_aws_credentials_secret	Password	-
idbroker_gcp_user_mapping	Text	-
idbroker_gcp_group_mapping	Text	-
idbroker_gcp_user_default_group_mapping	Text	-
idbroker_gcp_credential_key	Password	-
idbroker_gcp_credential_secret	Password	-
idbroker_azure_user_mapping	Text	-
idbroker_azure_group_mapping	Text	-
idbroker_azure_user_default_group_mapping	Text	-
idbroker_azure_adls2_tenant_name	Text	-
idbroker_azure_vm_assumer_identity	Text	-
idbroker_reloadable_refresh_interval_ms	Time	10 seconds (minimum = 1 second)
idbroker_kerberos_dt_proxyuser_block	Text Array	A comma-separated list of proxy user configuration used in Knox's dt topology in case Kerberos is enabled
idbroker_knox_token_ttl_ms	Time	1 hour (minimum = 1 second)

Related Information[Installing Apache Knox](#)

Custom Installation Solutions

Some installations may require custom solutions such as creating virtual images of cluster hosts, configuring a custom Java home location, or creating a Runtime cluster using a template.

Related Information[CDP Private Cloud Base Installation Guide](#)

Creating Virtual Images of Cluster Hosts

You can create virtual machine images, such as PXE-boot images, Amazon AMIs, and Azure VM images of cluster hosts with pre-deployed Cloudera software that you can use to quickly spin up virtual machines.

You can create virtual machine images, such as PXE-boot images, Amazon AMIs, and Azure VM images of cluster hosts with pre-deployed Cloudera software that you can use to quickly spin up virtual machines. These images use parcels to install Runtime software. This topic describes the procedures to create images of the Cloudera Manager host and worker host and how to instantiate hosts from those images.

Creating a Pre-Deployed Cloudera Manager Host

Complete the steps below to create a Cloudera Manager virtual machine image.

Procedure

1. Instantiate a virtual machine image (an AMI, if you are using Amazon Web Services) based on a supported operating system and start the virtual machine. See the documentation for your virtualization environment for details.
2. Install Cloudera Manager and configure a database. You can configure either a local or remote database.
3. Wait for the Cloudera Manager Admin console to become active.
4. Log in to the Cloudera Manager Admin console.
5. Download any parcels for Runtime or other services managed by Cloudera Manager. Do not distribute or activate the parcels.
6. Log in to the Cloudera Manager server host:
 - a) Run the following command to stop the Cloudera Manager service: `service cloudera-scm-server stop`
 - b) Run the following command to disable autostarting of the `cloudera-scm-server` service:
 - RHEL 7.x /CentOS 7.x.x:

```
systemctl disable cloudera-scm-server.service
```
 - Ubuntu:

```
update-rc.d -f cloudera-scm-server remove
```
7. Create an image of the Cloudera Manager host.
8. If you installed the Cloudera Manager database on a remote host, also create an image of the database host.



Note: Ensure that there are no clients using the remote database while creating the image.

Instantiating a Cloudera Manager Image

Complete the following steps to create a new Cloudera Manager instance from a virtual machine image.

Procedure

1. Instantiate the Cloudera Manager image.
2. If the Cloudera Manager database will be hosted on a remote host, also instantiate the database host image.

3. Ensure that the cloudera-scm-server service is not running by running the following command on the Cloudera Manager host:

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

If it is running, stop it using the following command:

```
service cloudera-scm-server stop
```

4. On the Cloudera Manager host, create a file named uuid in the /etc/cloudera-scm-server directory. Add a globally unique identifier to this file using the following command:

```
cat /proc/sys/kernel/random/uuid > /etc/cloudera-scm-server/uuid
```

The existence of this file informs Cloudera Manager to reinitialize its own unique identifier when it starts.

5. Run the following command to start the Cloudera Manager service:

```
service cloudera-scm-server start
```

6. Run the following command to enable automatic restart for the cloudera-scm-server:

- SLES:

```
chkconfig cloudera-scm-server on
```

- RHEL 7.x /CentOS 7.x.x:

```
systemctl enable cloudera-scm-server.service
```

- Ubuntu:

```
update-rc.d -f cloudera-scm-server defaults
```

Creating a Pre-Deployed Worker Host

Complete the steps below to create a pre-deployed worker host.

Procedure

1. Instantiate a virtual machine image (an AMI, if you are using Amazon Web Services) based on a supported operating system and start the virtual machine. See the documentation for your virtualization environment for details.
2. Download the parcels required for the worker host from the public parcel repository, or from a repository that you have created and save them to a temporary directory. See *Cloudera Manager 7 Download Information*.
3. From the same location where you downloaded the parcels, download the *parcel_name.parcel.sha1* file for each parcel.
4. Calculate and compare the sha1 of the downloaded parcel to ensure that the parcel was downloaded correctly. For example:

```
shasum KAFKA-2.0.2-1.2.0.2.p0.5-el6.parcel | awk '{print $1}' > KAFKA-2.0.2-1.2.0.2.p0.5-el6.parcel.sha  
  
diff KAFKA-2.0.2-1.2.0.2.p0.5-el6.parcel.sha1 KAFKA-2.0.2-1.2.0.2.p0.5-el6.parcel.sha
```

5. Unpack the parcel:

a) Create the following directories:

- /opt/cloudera/parcels
- /opt/cloudera/parcel-cache

b) Set the ownership for the two directories you just created so that they are owned by the username that the Cloudera Manager agent runs as.

c) Set the permissions for each directory using the following command:

```
chmod 755 directory
```

Note that the contents of these directories will be publicly available and can be safely marked as world-readable.

d) Running as the same user that runs the Cloudera Manager agent, extract the contents of the parcel from the temporary directory using the following command:

```
tar -zxvf parcelfile -C /opt/cloudera/parcels/
```

e) Add a symbolic link from the product name of each parcel to the /opt/cloudera/parcels directory.

For example, to link /opt/cloudera/parcels/CDH-6.0.0-1.cdh6.0.0.p0.309038 to /opt/cloudera/parcels/CDH, use the following command:

```
ln -s /opt/cloudera/parcels/CDH-6.0.0-1.cdh6.0.0.p0.309038 /opt/cloudera/parcels/CDH
```

f) Mark the parcels to not be deleted by the Cloudera Manager agent on start up by adding a .dont_delete marker file (this file has no contents) to each subdirectory in the /opt/cloudera/parcels directory. For example:

```
touch /opt/cloudera/parcels/CDH/.dont_delete
```

6. Verify the file exists:

```
ls -l /opt/cloudera/parcels/parcelname
```

You should see output similar to the following:

```
ls -al /opt/cloudera/parcels/CDH
total 100
drwxr-xr-x  9 root root  4096 Sep 14 14:53 .
drwxr-xr-x  9 root root  4096 Sep 14 06:34 ..
drwxr-xr-x  2 root root  4096 Sep 12 06:39 bin
-rw-r--r--  1 root root    0 Sep 14 14:53 .dont_delete
drwxr-xr-x 26 root root  4096 Sep 12 05:10 etc
drwxr-xr-x  4 root root  4096 Sep 12 05:04 include
drwxr-xr-x  2 root root 69632 Sep 12 06:44 jars
drwxr-xr-x 37 root root  4096 Sep 12 06:39 lib
drwxr-xr-x  2 root root  4096 Sep 12 06:39 meta
drwxr-xr-x  5 root root  4096 Sep 12 06:39 share
```

7. Install the Cloudera Manager agent. If you have not already done so, *Step 1: Configure a Repository for Cloudera Manager*.**8. Create an image of the worker host.** See the documentation for your virtualization environment for details.

Instantiating a Worker Host

Complete the steps below to instantiate a worker host.

Procedure

1. Instantiate the Cloudera worker host image.

2. Edit the following file and set the `server_host` and `server_port` properties to reference the Cloudera Manager server host.
3. If necessary perform additional steps to configure TLS/SSL.
4. Start the agent service:

```
service cloudera-scm-agent start
```

Configuring a Custom Java Home Location

Although not recommended, the Java Development Kit (JDK),, may be installed at a custom location if necessary. These steps assume you have already installed the JDK as documented in *Step 2: Install Java Development Kit*.

About this task

Cloudera strongly recommends installing the JDK at `/usr/java/jdk-version`, which allows Cloudera Manager to auto-detect and use the correct JDK version. If you install the JDK anywhere else, you must follow these instructions to configure Cloudera Manager with your chosen location. The following procedure changes the JDK location for Cloudera Management Services and Runtime cluster processes only. It does not affect the JDK used by other non-Cloudera processes, or gateway roles. To modify the Cloudera Manager configuration to ensure the JDK can be found:

Procedure

1. Open the Cloudera Manager Admin Console.
2. In the left-side navigation bar, click **Hosts** Hosts Configuration. If you are configuring the JDK location on a specific host only, click **Hosts** All Hosts, select the specific host that you want to configure, and click the **Configuration** tab.
3. Select **Category** Advanced.
4. Set the **Java Home Directory** property to the custom location.
5. Click **Save Changes**.
6. Restart all services.

Manually Install Cloudera Software Packages

This topic shows how to manually install Cloudera Manager packages. Package installations of Cloudera Runtime are not supported in CDP Private Cloud Base .

Before manual installation, you must configure a repository. See [Step 1: Configure a Repository for Cloudera Manager](#) on page 70.

Install Cloudera Manager Packages

Cloudera Manager is installed on the Cloudera Manager Server host using packages.

Procedure

1. On the Cloudera Manager Server host, type the following commands to install the Cloudera Manager packages:

OS	Command
RHEL	<pre>sudo yum install cloudera-manager-daemons cloudera-manager-agent cloudera-manager-server</pre>

- If you are using an Oracle database for Cloudera Manager Server, edit the `/etc/default/cloudera-scm-server` file on the Cloudera Manager server host. Locate the line that begins with `export CMF_JAVA_OPTS` and change the `-Xmx2G` option to `-Xmx4G`.

Manually Install Cloudera Manager Agent Packages

The Cloudera Manager Agent is responsible for starting and stopping processes, unpacking configurations, triggering installations, and monitoring all hosts in a cluster. You can install the Cloudera Manager agent manually on all hosts, or Cloudera Manager can install the Agents in a later step. To use Cloudera Manager to install the agents, skip this section.

About this task

To install the Cloudera Manager Agent packages manually, do the following on every cluster host (including those that will run one or more of the Cloudera Management Service roles: Service Monitor, Activity Monitor, Event Server, Alert Publisher, or Reports Manager):

Procedure

- Use one of the following commands to install the Cloudera Manager Agent packages:

OS	Command
RHEL, if you have a yum repo configured:	<pre>\$ sudo yum install cloudera-manager-agent cloudera-manager-daemons</pre>
Ubuntu	<pre>\$ sudo apt-get install cloudera-manager-agent cloudera-manager-daemons</pre>

- On every cluster host, configure the Cloudera Manager Agent to point to the Cloudera Manager Server by setting the following properties in the `/etc/cloudera-scm-agent/config.ini` configuration file:

Property	Description
<code>server_host</code>	Name of the host where Cloudera Manager Server is running.
<code>server_port</code>	Port on the host where Cloudera Manager Server is running.

- Start the Agents by running the following command on all hosts:

```
sudo systemctl start cloudera-scm-agent
```

If the agent starts without errors, no response displays.

When the Agent starts, it contacts the Cloudera Manager Server. If communication fails between a Cloudera Manager Agent and Cloudera Manager Server, see *Troubleshooting Installation Problems*. When the Agent hosts reboot, `cloudera-scm-agent` starts automatically.

Installation Reference

Reference information related to CDP Private Cloud Base installation.

Related Information

[CDP Private Cloud Base Installation Guide](#)

Ports

Cloudera Manager, Cloudera Runtime components, managed services, and third-party components use the ports listed in the tables that follow.

Before you deploy Cloudera Manager, Cloudera Runtime, managed services, and third-party components, make sure these ports are open on each system. If you are using a firewall, such as iptables or firewalld, and cannot open all the listed ports, you must disable the firewall completely to ensure full functionality.

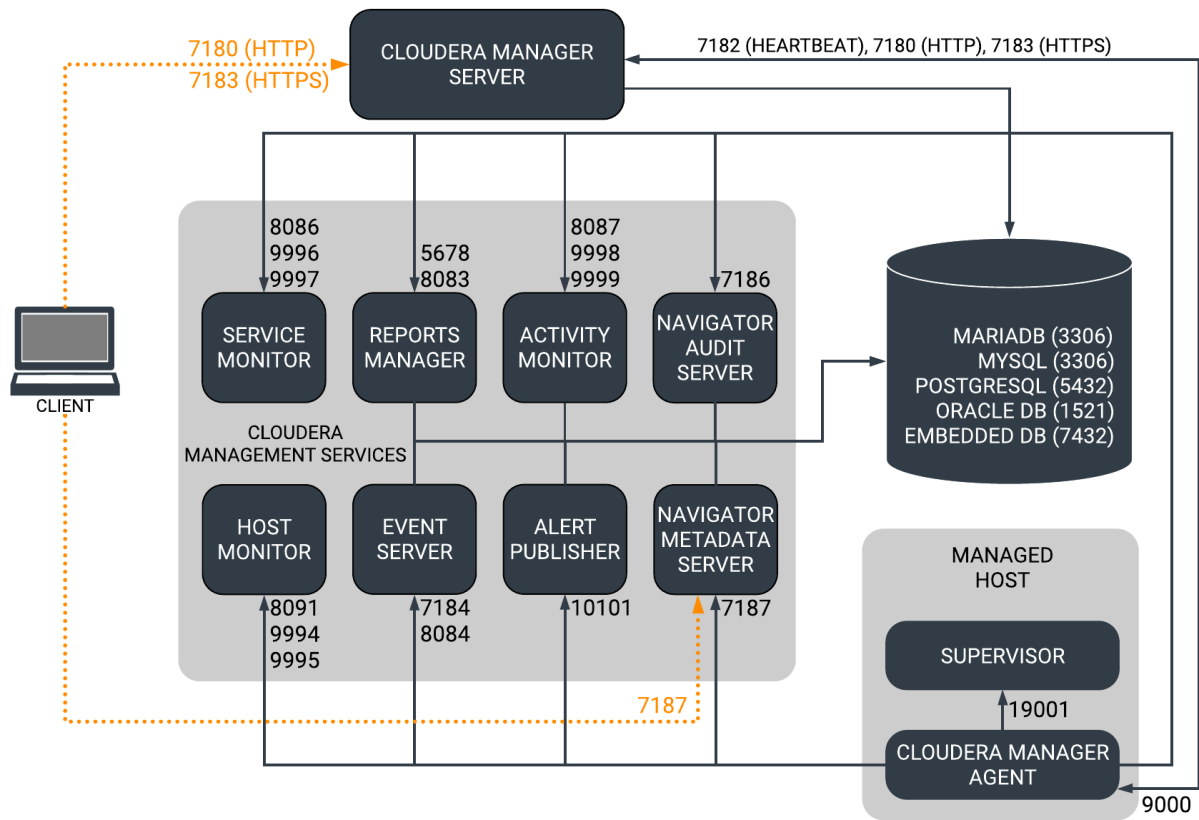
In the tables in the subsections that follow, the Access Requirement column for each port is usually either "Internal" or "External." In this context, "Internal" means that the port is used only for communication among the components (for example the JournalNode ports in an HA configuration); "External" means that the port can be used for either internal or external communication (for example, ports used by NodeManager and the JobHistory Server Web UIs).

Unless otherwise specified, the ports access requirement is unidirectional, meaning that inbound connections to the specified ports must be allowed. In most modern stateful firewalls, it is not necessary to create a separate rule for return traffic on a permitted session.

Ports Used by Cloudera Manager

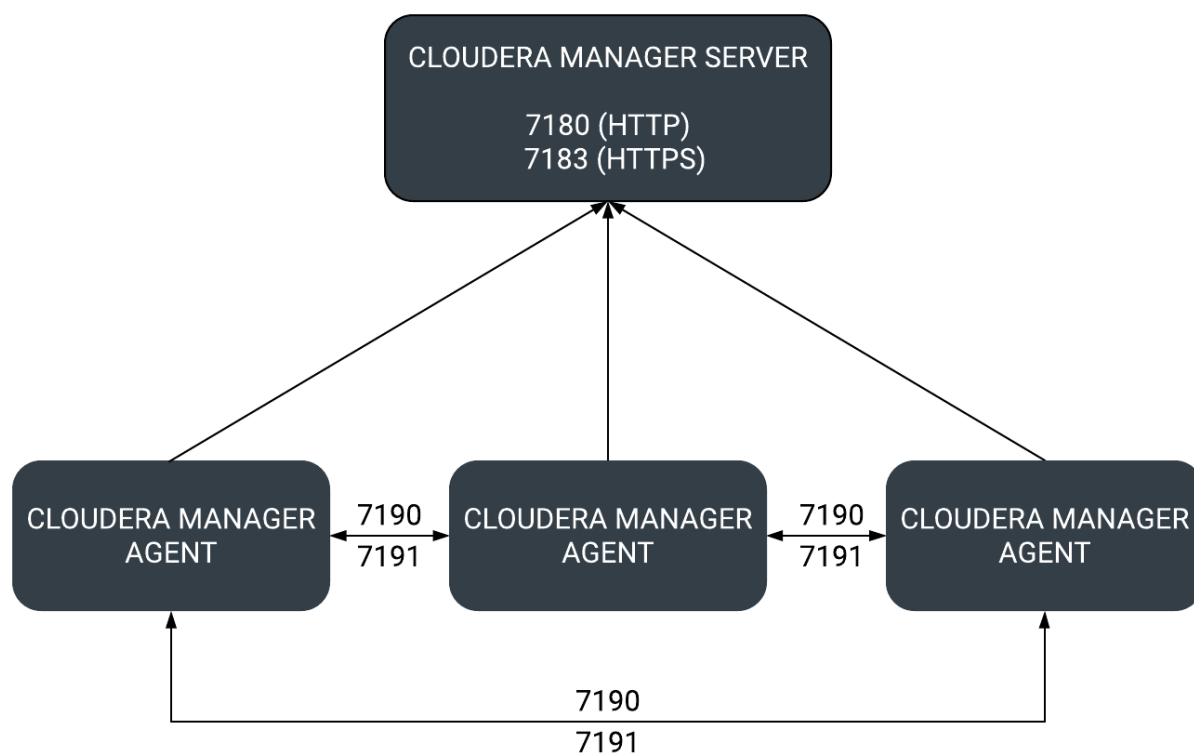
The diagrams and tables below provide an overview of some of the ports used by Cloudera Manager and Cloudera Management Service roles.

Figure 1: Ports Used by Cloudera Manager



When peer-to-peer distribution is enabled for parcels, the Cloudera Manager Agent can obtain the parcel from the Cloudera Manager Server or from other agents, as follows:

Figure 2: Ports Used in Peer-to-Peer Parcel Distribution



For further details, see the following tables. All ports listed are TCP.

In the following tables, Internal means that the port is used only for communication among the components; External means that the port can be used for either internal or external communication.

Table 38: External Ports

Component	Service	Port	Configuration	Description
Cloudera Manager Server	HTTP (Web UI)	7180	AdministrationSettingsCategory and AddressesHTTP Port for Admin Console	HTTP Port used by the web console.
	HTTPS (Web UI)	7183	AdministrationSettingsCategory and AddressesHTTPS Port for Admin Console	HTTPS Port used by the web console if HTTPS is enabled. If enabled, port 7180 remains open, but redirects all requests to HTTPS on port 7183.
Backup and Disaster Recovery	HTTP (Web UI)	7180	AdministrationSettingsCategory and AddressesHTTP Port for Admin Console	HTTP Port for communication to peer (source) Cloudera Manager.
	HTTPS (Web UI)	7183	AdministrationSettingsCategory and AddressesHTTPS Port for Admin Console	HTTPS Port for communication to peer (source) Cloudera Manager when HTTPS is enabled.

Component	Service	Port	Configuration	Description
	HDFS NameNode	8020	HDFS serviceConfigurationCategoryPorts and AddressesNameNode Port	HDFS and Hive/ Impala replication: communication from destination HDFS and MapReduce hosts to source HDFS NameNode(s). Hive/ Impala Replication: communication from source Hive hosts to destination HDFS NameNode(s).
	HDFS DataNode	50010	HDFS serviceConfigurationCategoryPorts and AddressesDataNode Transceiver Port	HDFS and Hive/ Impala replication: communication from destination HDFS and MapReduce hosts to source HDFS DataNode(s). Hive/ Impala Replication: communication from source Hive hosts to destination HDFS DataNode(s).
Telemetry Publisher	HTTP	10110	ClustersCloudera Management ServiceCategoryPorts and AddressesTelemetry Publisher Server Port	The port where the Telemetry Publisher Server listens for requests
Telemetry Publisher	HTTP (Debug)	10111	ClustersCloudera Management ServiceCategoryPorts and AddressesTelemetry Publisher Web UI Port	The port where Telemetry Publisher starts a debug web server. Set to -1 to disable debug server.

Table 39: Internal Ports

Component	Service	Port	Configuration	Description
Cloudera Manager Server	Avro (RPC)	7182	AdministrationSettingsCategoryPorts and AddressesAgent Port to connect to Server	Used for Agent to Server heartbeats
	Embedded PostgreSQL database	7432		The optional embedded PostgreSQL database used for storing configuration information for Cloudera Manager Server.
	Peer-to-peer parcel distribution	7190, 7191	HostsAll HostsConfigurationP2P Parcel Distribution Port	Used to distribute parcels to cluster hosts during installation and upgrade operations.
Cloudera Manager Agent	HTTP (Debug)	9000	/etc/cloudera-scm-agent/config.ini	
Event Server	Custom protocol	7184	Cloudera Management ServiceConfigurationCategoryPorts and AddressesEvent Publish Port	Port on which the Event Server listens for the publication of events.
	Custom protocol	7185	Cloudera Management ServiceConfigurationCategoryPorts and AddressesEvent Query Port	Port on which the Event Server listens for queries for events.
	HTTP (Debug)	8084	Cloudera Management ServiceConfigurationCategoryPorts and AddressesEvent Server Web UI Port	Port for the Event Server's Debug page. Set to -1 to disable debug server.

Component	Service	Port	Configuration	Description
Alert Publisher	Custom protocol	10101	Cloudera Management ServiceConfigurationCategoryPorts and AddressesAlerts: Listen Port	Port where the Alert Publisher listens for internal API requests.
Service Monitor	HTTP (Debug)	8086	Cloudera Management ServiceConfigurationCategoryPorts and AddressesService Monitor Web UI Port	Port for Service Monitor's Debug page. Set to -1 to disable the debug server.
	HTTPS (Debug)		Cloudera Management ServiceConfigurationCategoryPorts and AddressesService Monitor Web UI HTTPS Port	Port for Service Monitor's HTTPS Debug page.
	Custom protocol	9997	Cloudera Management ServiceConfigurationCategoryPorts and AddressesService Monitor Listen Port	Port where Service Monitor is listening for agent messages.
	Internal query API (Avro)	9996	Cloudera Management ServiceConfigurationCategoryPorts and AddressesService Monitor Nozzle Port	Port where Service Monitor's query API is exposed.
Activity Monitor	HTTP (Debug)	8087	Cloudera Management ServiceConfigurationCategoryPorts and AddressesActivity Monitor Web UI Port	Port for Activity Monitor's Debug page. Set to -1 to disable the debug server.
	HTTPS (Debug)		Cloudera Management ServiceConfigurationCategoryPorts and AddressesActivity Monitor Web UI HTTPS Port	Port for Activity Monitor's HTTPS Debug page.
	Custom protocol	9999	Cloudera Management ServiceConfigurationCategoryPorts and AddressesActivity Monitor Listen Port	Port where Activity Monitor is listening for agent messages.
	Internal query API (Avro)	9998	Cloudera Management ServiceConfigurationCategoryPorts and AddressesActivity Monitor Nozzle Port	Port where Activity Monitor's query API is exposed.
Host Monitor	HTTP (Debug)	8091	Cloudera Management ServiceConfigurationCategoryPorts and AddressesHost Monitor Web UI Port	Port for Host Monitor's Debug page. Set to -1 to disable the debug server.
	HTTPS (Debug)	9091	Cloudera Management ServiceConfigurationCategoryPorts and AddressesHost Monitor Web UI HTTPS Port	Port for Host Monitor's HTTPS Debug page.
	Custom protocol	9995	Cloudera Management ServiceConfigurationCategoryPorts and AddressesHost Monitor Listen Port	Port where Host Monitor is listening for agent messages.
	Internal query API (Avro)	9994	Cloudera Management ServiceConfigurationCategoryPorts and AddressesHost Monitor Nozzle Port	Port where Host Monitor's query API is exposed.
Reports Manager	Queries (Thrift)	5678	Cloudera Management ServiceConfigurationCategoryPorts and AddressesReports Manager Server Port	The port where Reports Manager listens for requests.

Component	Service	Port	Configuration	Description
	HTTP (Debug)	8083	Cloudera Management ServiceConfigurationCategoryPorts and AddressesReports Manager Web UI Port	The port where Reports Manager starts a debug web server. Set to -1 to disable debug server.

Ports Used by Cloudera Navigator Key Trustee Server

The Cloudera Navigator Key Trustee Server uses certain ports to store and retrieve encryption information and information required for high availability.

All ports listed are TCP.

In the following table, the Access Requirement column for each port is usually either "Internal" or "External." In this context, "Internal" means that the port is used only for communication among the components; "External" means that the port can be used for either internal or external communication.

Component	Service	Port	Access Requirement	Configuration	Comment
Cloudera Navigator Key Trustee Server	HTTPS (key management)	11371	External	Key Trustee Server serviceConfigurationCategoryPorts and AddressesKey Trustee Server Port	Navigator Key Trustee Server clients (including Key Trustee KMS and Navigator Encrypt) access this port to store and retrieve encryption keys.
	PostgreSQL database	11381	External	Key Trustee Server serviceConfigurationCategoryPorts and AddressesKey Trustee Server Database Port	The Navigator Key Trustee Server database listens on this port. The Passive Key Trustee Server connects to this port on the Active Key Trustee Server for replication in Cloudera Navigator Key Trustee Server High Availability.

Ports Used by Cloudera Runtime Components

Cloudera Runtime components use a number of ports for associated services.

All ports listed are TCP.

In the following tables, Internal means that the port is used only for communication among the components; External means that the port can be used for either internal or external communication.

Table 40: External Ports

Component	Service	Port	Configuration	Comment
Apache Atlas	Non-SSL	31000	atlas.server.http.port	
	SSL	31443	atlas.server.https.port	This port is used only when Atlas is in SSL mode.
Apache Hadoop HDFS	DataNode	9866	dfs.datanode.address	DataNode server address and port for data transfer.
		9864	dfs.datanode.http.address	DataNode HTTP server port.
		9865	dfs.datanode.https.address	DataNode HTTPS server port.
		9867	dfs.datanode.ipc.address	DataNode IPC server port.
	NameNode	8020	fs.default.name or fs.defaultFS	fs.default.name is deprecated (but still works)
		8022	dfs.namenode.servicerpc-address	Optional port used by HDFS daemons to avoid sharing the RPC port used by clients (8020). Cloudera recommends using port 8022.

Component	Service	Port	Configuration	Comment
		9870	dfs.http.address or dfs.namenode.http-address	dfs.http.address is deprecated (but still works)
		9871	dfs.https.address or dfs.namenode.https-address	dfs.https.address is deprecated (but still works)
	NFS gateway	2049		nfs port (nfs3.server.port)
		4242		mountd port (nfs3.mountd.port)
		111		portmapper or rpcbind port.
		50079	nfs.http.port	The NFS gateway daemon uses this port to serve metrics. The port is configurable on versions 5.10 and higher.
		50579	nfs.https.port	The NFS gateway daemon uses this port to serve metrics. The port is configurable on versions 5.10 and higher.
	HttpFS	14000		HttpFS server port
		14001		HttpFS admin port
Apache Hadoop YARN (MRv2)	ResourceManager	8032	yarn.resourcemanager.address	
		8033	yarn.resourcemanager.admin.address	
		8088	yarn.resourcemanager.webapp.address	
		8090	yarn.resourcemanager.webapp.https.address	
	NodeManager	8042	yarn.nodemanager.webapp.address	
		8044	yarn.nodemanager.webapp.https.address	
	JobHistory Server	19888	mapreduce.jobhistory.webapp.address	
		19890	mapreduce.jobhistory.webapp.https.address	
	ApplicationMaster			The ApplicationMaster serves an HTTP service using an ephemeral port that cannot be restricted. This port is never accessed directly from outside the cluster by clients. All requests to the ApplicationMaster web server is routed using the YARN ResourceManager (proxy service). Locking down access to ephemeral port ranges within the cluster's network might restrict your access to the ApplicationMaster UI and its logs, along with the ability to look at running applications.
Apache Flume	Flume Agent	41414		
Apache Hadoop KMS	Key Management Server	16000	kms_http_port	Applies to both Java KeyStore KMS and Key Trustee KMS.
Apache HBase	Master	16000	hbase.master.port	IPC
		16010	hbase.master.info.port	HTTP
	RegionServer	16020	hbase.regionserver.port	IPC
		16030	hbase.regionserver.info.port	HTTP

Component	Service	Port	Configuration	Comment
	REST	20550	hbase.rest.port	The default REST port in HBase is 8080. Because this is a commonly used port, Cloudera Manager sets the default to 20550 instead.
	REST UI	8085		
	Thrift Server	9090	Pass -p <port> on CLI	
	Thrift Server	9095		
		9090	Pass --port <port> on CLI	
	Lily HBase Indexer	11060		
Apache Hive	Metastore	9083		
	HiveServer2	10000	hive.server2.thrift.port	The Beeline command interpreter requires that you specify this port on the command line. If you use Oracle database, you must manually reserve this port.
	HiveServer2 Web User Interface (UI)	10002	hive.server2.webui.port in hive-site.xml	
Hue	Server	8888		
	Load Balancer	8889		
Apache Impala	Impala Daemon	21000		Used to transmit commands and receive results by impala-shell and version 1.2 of the Cloudera ODBC driver.
		21050		Used to transmit commands and receive results by applications, such as Business Intelligence tools, using JDBC, the Beeswax query editor in Hue, and version 2.0 or higher of the Cloudera ODBC driver.
		25000		Impala web interface for administrators to monitor and troubleshoot.
	StateStore Daemon	25010		StateStore web interface for administrators to monitor and troubleshoot.
	Catalog Daemon	25020		Catalog service web interface for administrators to monitor and troubleshoot.
Apache Kafka	Kafka Broker	9092	port	The primary communication port used by producers and consumers; also used for inter-broker communication.
		9093	ssl_port	A secured communication port used by producers and consumers; also used for inter-broker communication.
	Kafka Connect	38083	rest.port	Kafka Connect Rest Port.
		38085	secure.rest.port	Kafka Connect Secure Rest Port.
Apache Kudu	Master	7051		Kudu Master RPC port.
		8051		Kudu Master HTTP server port.
	TabletServer	7050		Kudu TabletServer RPC port.
		8050		Kudu TabletServer HTTP server port.

Component	Service	Port	Configuration	Comment
Apache Oozie	Oozie Server	11000	OOZIE_HTTP_PORT in oozie-env.sh	HTTP
		11443		HTTPS
Apache Ranger	Non-SSL	6080	ranger.service.http.port	
	SSL	6182	ranger.service.https.port	This port is used only when Ranger is in SSL mode.
	Admin Unix Auth Service Port	5151	ranger.unixauth.service.port	
Apache Solr	Solr Server	8983		HTTP port for all Solr-specific actions, update/query.
	Solr Server	8985		HTTPS port for all Solr-specific actions, update/query.
Apache Spark	Default Master RPC port	7077		
	Default Worker RPC port	7078		
	Default Master web UI port	18080		
	Default Worker web UI port	18081		
	History Server	18088	history.port	
Apache Sqoop	Metastore	16000	sqoop.metastore.server.port	
Apache ZooKeeper	Server (with Cloudera Runtime or Cloudera Manager)	2181	clientPort	Client port.
Cruise Control	Cruise Control Server	8899	webserver.http.port	This is the main port that enables access to the Cruise Control Server
Schema Registry	Schema Registry Server	7788	schema.registry.port	REST endpoint for Schema Registry.
		7789	schema.registry.adminPort	Page for monitoring the Schema Registry service to determine for example the health state and CPU usage.
		7790	schema.registry.ssl.port	When SSL is enabled, REST endpoint for Schema Registry.
		7791	schema.registry.ssl.adminPort	When SSL is enabled, the page for monitoring the Schema Registry service to determine for example the health state and CPU usage.
Streams Messaging Manager	Streams Messaging Manager Rest Admin Server	8585	streams.messaging.manager.port	Streams Messaging Manager Port
		8587	streams.messaging.manager.ssl.port	Streams Messaging Manager Port (SSL)
		8586	streams.messaging.manager.adminPort	Streams Messaging Manager Admin Port
		8588	streams.messaging.manager.ssl.adminPort	Streams Messaging Manager Admin Port (SSL)
	Streams Messaging Manager UI Server	9991	streams.messaging.manager.ui.port	The port on which server accepts connections. This port is used for both secured and unsecured connections.

Component	Service	Port	Configuration	Comment
Streams Replication Manager	SRM Service	6670	streams.replication.manager.service.port	SRM Service port.
		6671	streams.replication.manager.service.ssl.port	SRM Service port when SSL is enabled.

Table 41: Internal Ports

Component	Service	Port	Configuration	Comment
Apache Hadoop HDFS	Secondary NameNode	9868	dfs.secondary.http.address or dfs.namenode.secondary.http-address	dfs.secondary.http.address is deprecated (but still works)
		9869	dfs.secondary.https.address	
	JournalNode	8485	dfs.namenode.shared.edits.dir	
		8480	dfs.journalnode.http-address	
		8481	dfs.journalnode.https-address	
	Failover Controller	8019		Used for NameNode HA
Apache Hadoop YARN (MRv2)	ResourceManager	8030	yarn.resourcemanager.scheduler.address	
		8031	yarn.resourcemanager.resource-tracker.address	
	NodeManager	8040	yarn.nodemanager.localizer.address	
		8041	yarn.nodemanager.address	
	JobHistory Server	10020	mapreduce.jobhistory.address	
		10033	mapreduce.jobhistory.admin.address	
	Shuffle HTTP	13562	mapreduce.shuffle.port	
Apache Hadoop KMS	Key Management Server	16001	kms_admin_port	Applies to both Java KeyStore KMS and Key Trustee KMS.
Apache HBase	HQuorumPeer	2181	hbase.zookeeper.property.clientPort	HBase-managed ZooKeeper mode
		2888	hbase.zookeeper.peerport	HBase-managed ZooKeeper mode
		3888	hbase.zookeeper.leaderport	HBase-managed ZooKeeper mode
Apache Impala	Impala Daemon	22000		Internal use only. Impala daemons use this port to communicate with each other.
		23000		Internal use only. Impala daemons listen on this port for updates from the statestore daemon.
	StateStore Daemon	24000		Internal use only. The statestore daemon listens on this port for registration/unregistration requests.
	Catalog Daemon	23020		Internal use only. The catalog daemon listens on this port for updates from the statestore daemon.
		26000		Internal use only. The catalog service uses this port to communicate with the Impala daemons.

Component	Service	Port	Configuration	Comment
Apache Kafka	Kafka Broker	9092	port	The primary communication port used by producers and consumers; also used for inter-broker communication.
		9093	ssl_port	A secured communication port used by producers and consumers; also used for inter-broker communication.
		9393	jmx_port	Internal use only. Used for administration via JMX.
		9394	kafka.http.metrics.port	Internal use only. This is the port via which the HTTP metric reporter listens. It is used to retrieve metrics through HTTP instead of JMX.
	Kafka Connect	38084	metrics.jetty.server.port	Metrics Jetty Server Port
	Kafka MirrorMaker	24042	jmx_port	Internal use only. Used to administer the producer and consumer of the MirrorMaker.
Apache Solr	Solr Server	8993		Infra-Solr HTTP port
	Solr Server	8995		Infra-Solr HTTPS port
Apache Spark	Shuffle service	7337		
Apache ZooKeeper	Server (with Cloudera Runtime only)	2888	X in server.N =host:X:Y	Peer
	Server (with Cloudera Runtime only)	3888	X in server.N =host:X:Y	Peer
	Server (with Cloudera Runtime and Cloudera Manager)	3181	X in server.N =host:X:Y	Peer
	Server (with Cloudera Runtime and Cloudera Manager)	4181	X in server.N =host:X:Y	Peer
	ZooKeeper JMX port	9010		<p>ZooKeeper will also use another randomly selected port for RMI. To allow Cloudera Manager to monitor ZooKeeper, you must do one of the following:</p> <ul style="list-style-type: none"> • Open up all ports when the connection originates from the Cloudera Manager Server • Do the following: <ol style="list-style-type: none"> 1. Open a non-ephemeral port (such as 9011) in the firewall. 2. Install Oracle Java 7u4 JDK or higher. 3. Add the port configuration to the advanced configuration snippet, for example: <code>-Dcom.sun.management.jmxremote.rmi.port=9011</code> 4. Restart ZooKeeper.
Streams Messaging Manager	Streams Messaging Manager Rest Admin Server	6670	streams.replication.manager.port	Streams Replication Manager rest port
		6671	streams.replication.manager.port	Streams Replication Manager rest port on SSL
		7180	cm.metrics.port	Cloudera Manager's HTTP port.

Component	Service	Port	Configuration	Comment
		7183	cm.metrics.port	Cloudera Manager's HTTPS port
		9997	cm.metrics.service.monitor.port	Cloudera Manager Service Monitor port
		38083	kafka.connect.port	Kafka Connect port
		3306	streams.messaging.manager.storage.connector.port	Streams Messaging Manager database port

Ports Used by DistCp

DistCp uses various ports for HDFS and HttpFS services.

All ports listed are TCP.

In the following table, the Access Requirement column for each port is usually either "Internal" or "External." In this context, "Internal" means that the port is used only for communication among the components; "External" means that the port can be used for either internal or external communication.

Component	Service	Qualifier	Port	Access Requirement	Configuration	Comment
Hadoop HDFS	NameNode		8020	External	fs.default.name	fs.default.name
					or fs.defaultFS	is deprecated (but still works)
	DataNode	Secure	1004	External	dfs.datanode.address	
	DataNode		50010	External	dfs.datanode.address	
WebHDFS	NameNode		50070	External	dfs.http.address	dfs.http.address
					or dfs.namenode.http-address	is deprecated (but still works)
	DataNode	Secure	1006	External	dfs.datanode.http.address	
HttpFS	web		14000			

Ports Used by Third-Party Components

Third-party components such as PostgreSQL and LDAP use a number of ports for associated services.

In the following table, the Access Requirement column for each port is usually either "Internal" or "External." In this context, "Internal" means that the port is used only for communication among the components; "External" means that the port can be used for either internal or external communication.

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
Ganglia	ganglia-gmond		8649	UDP/TCP	Internal		
	ganglia-web		80	TCP	External	Via Apache <code>httpd</code>	
Kerberos	KRB5 KDC Server	Secure	88	UDP/TCP	External	<code>kdc_ports</code> and <code>kdc_tcp_ports</code> in either the <code>[kdcdefaults]</code> or <code>[realms]</code> sections of <code>kdc.conf</code>	By default only UDP
	KRB5 Admin Server	Secure	749	TCP	External	<code>kadmind_port</code> in the <code>[realms]</code> section of <code>kdc.conf</code>	
	kpasswd		464	UDP/TCP	External		
SSH	ssh		22	TCP	External		
PostgreSQL			5432	TCP	Internal		
MariaDB			3306	TCP	Internal		
MySQL			3306	TCP	Internal		
LDAP	LDAP Server		389	TCP	External		
	LDAP Server over TLS/SSL	TLS/SSL	636	TCP	External		
	Global Catalog		3268	TCP	External		
	Global Catalog over TLS/SSL	TLS/SSL	3269	TCP	External		

Service Dependencies in Cloudera Manager

The following tables list service dependencies that exist between various services in a Cloudera Manager deployment.

When configuring CDP Runtime for production environments, be sure that Kerberos is enabled for user authentication. Cloudera supports security services such as Ranger and Atlas when they run on clusters where Kerberos is enabled to authenticate users.

Service dependencies for Spark 2 on YARN and Cloudera Data Science Workbench are listed separately.

Table 42: Service Dependencies

Service	Dependencies	Optional Dependencies
ADLS Connector		
Atlas	<ul style="list-style-type: none"> HDFS HBase Kafka (Kafka broker role only) Solr 	Ranger
Cruise Control	<ul style="list-style-type: none"> Kafka Zookeeper 	
Data Context Connector		
HBase	<ul style="list-style-type: none"> HDFS ZooKeeper 	<ul style="list-style-type: none"> Atlas Ranger
HDFS		<ul style="list-style-type: none"> ADLS Connector or S3 Connector KMS, Thales KMS, Key Trustee, or Luna KMS Ranger ZooKeeper
Hive	HDFS	<ul style="list-style-type: none"> Atlas HBase Kudu Ranger Spark on YARN YARN ZooKeeper
Hive-on-Tez	<ul style="list-style-type: none"> HDFS Hive Tez 	<ul style="list-style-type: none"> Atlas HBase Ranger YARN ZooKeeper
Hue	<ul style="list-style-type: none"> HDFS Hive 	<ul style="list-style-type: none"> Atlas HBase Hive-on-Tez Impala Oozie Solr ZooKeeper
Impala	<ul style="list-style-type: none"> HDFS Hive 	<ul style="list-style-type: none"> Atlas HBase Kudu Ranger YARN ZooKeeper
Kafka	ZooKeeper	<ul style="list-style-type: none"> HDFS Ranger

Service	Dependencies	Optional Dependencies
Key-Value Store Indexer	<ul style="list-style-type: none"> HBase Solr 	Ranger
Kudu		Ranger
Livy	<ul style="list-style-type: none"> Spark-on-YARN YARN 	Hive
Oozie	YARN	<ul style="list-style-type: none"> Hive Spark on YARN ZooKeeper
Ozone		<ul style="list-style-type: none"> HDFS Ranger
Ranger	<ul style="list-style-type: none"> HBase HDFS Kafka Solr 	
S3 Connector		
Schema Registry		<ul style="list-style-type: none"> HDFS Ranger
Solr	<ul style="list-style-type: none"> HDFS ZooKeeper 	Ranger
Spark on YARN	YARN	<ul style="list-style-type: none"> Atlas HBase
Streams Messaging Manager	Kafka	<ul style="list-style-type: none"> Ranger Schema Registry Zookeeper
Streams Replication Manager		Kafka
Tez	YARN	
YARN	<ul style="list-style-type: none"> HDFS ZooKeeper 	Ranger
Zeppelin	<ul style="list-style-type: none"> HDFS Spark-on-YARN YARN 	<ul style="list-style-type: none"> Livy
ZooKeeper		

Related Information

[Runtime Cluster Hosts and Role Assignments](#)

Cloudera Manager sudo command options

To install, configure, start and stop the Cloudera Manager (CM), manage files, and so on, you can use the CM sudo commands.

Following is the list of sudo commands run by Cloudera Manager.



Note: In the list, RH6 = RHEL 6 / CentOS 6 / Oracle 6, RH7+ = RHEL 7 / CentOS 7 / Oracle 7, and later, and SLES 11 and later, Ubuntu = All Ubuntu versions, and SLES = All SLES versions. For those command supported in all the Operating System (OS) versions, an OS flavor is not specified.

- `sudo yum` (RH6, RH7+) - Install or remove software.
- `sudo apt-get` (Ubuntu) - Install or remove software.
- `sudo apt-key` (Ubuntu) - Update Repository key.
- `sudo sed` - Edit one or more text files (stream editor).
- `sudo systemctl` (RH7+, Ubuntu) - Start, stop, or configure software.
- `sudo service` (RH6) - Start or stop software.
- `sudo /sbin/chkconfig sudo chkconfig` (RH6) - Configure software.
- `sudo /usr/sbin/update-rc.d` (Ubuntu) - Configure software.
- `sudo id` - Used for user identification.
- `sudo rm` - Remove files.
- `sudo mv` - Move or rename files.
- `sudo chown` - Modify file ownership.
- `sudo install` - Install software.
- `sudo service` (RH6) - Start, stop, or restart the Cloudera Manager Server and Cloudera Manager Agents on the cluster hosts.
- `sudo systemctl` (RH7+, Ubuntu) - Start, stop, or restart the Cloudera Manager Server and Cloudera Manager Agents on the cluster hosts.
- `sudo cp` - Used for file copy.
- `sudo /opt/cloudera/cm-agent/bin/cm` - Used for certificate management and troubleshooting.
- `sudo mkdir` - Used for directory creation.
- `sudo /opt/cloudera/parcels/keycloak/cloudera_keycloak.sh` - Configure and startup Keycloak.
- `sudo keytrustee` - Used for Keytrustee backup.
- `sudo ln` - Manage file links.
- `sudo chmod` - Manage file permissions.
- `sudo wget` - Used to host local repositories for CM and CDH.
- `sudo -u postgres psql postgres` - Connect to PSQL as postgres user.
- `sudo -E tar` - Archive CM agent data directories prior to updates or changes.
- `sudo zypper clean --all` (SLES) - Clean up the repository cache for SLES package manager (zypper).
- `sudo ktadmin enable-synchronous-replication` - Enable synchronous replication on the active Key Trustee Server.
- `sudo ktadmin enable-synchronous-replication` - Enable synchronous replication on the active Key Trustee Server.
- `sudo rpm` (RH6, RH7+) - Install or remove the CM RPM packages.

Introduction to Parcels

Parcels are a packaging format that facilitate upgrading software from within Cloudera Manager.

You can download, distribute, and activate a new software version all from within Cloudera Manager. Cloudera Manager downloads a parcel to a local directory. Once the parcel is downloaded to the Cloudera Manager Server host, an Internet connection is no longer needed to deploy the parcel. For detailed information about parcels, see [Overview of Parcels](#).

If your Cloudera Manager Server does not have Internet access, you can obtain the required parcel files and put them into a parcel repository. For more information, see [Configuring a Local Parcel Repository](#) on page 67.

After You Install

The following topics describe post-installation actions, such as deploying client configuration and some simple tests to validate the installation and confirm that everything is working as expected.

Related Information

[CDP Private Cloud Base Installation Guide](#)

Deploying Clients

Client configuration files are generated automatically by Cloudera Manager based on the services you install.

Cloudera Manager deploys these configurations automatically at the end of the installation workflow. You can also download the client configuration files to deploy them manually.

If you modify the configuration of your cluster, you might need to redeploy the client configuration files. If a service's status is "Client configuration redeployment required," you need to redeploy those files.

Testing the Installation

Begin testing the installation from the **Home** page, where you can start by checking the health of the services.

To begin testing, start the Cloudera Manager Admin Console. Once you've logged in, the **Home** page should look something like this:

On the left side of the screen is a list of services currently running with their status information. All the services

should be running with Good Health . You can click each service to view more detailed information about each service. You can also test your installation by either checking each Host's heartbeats, running a MapReduce job, or interacting with the cluster with an existing Hue application.

Checking Host Heartbeats

One way to check whether all the Agents are running is to look at the time since their last heartbeat. You can do this by clicking the Hosts tab where you can see a list of all the hosts along with the value of their Last Heartbeat.

By default, every Agent must heartbeat successfully every 15 seconds. A recent value for the Last Heartbeat means that the Server and Agents are communicating successfully.

Running a MapReduce Job

Run a PiEstimator job to manually verify that the CDP Private Cloud Base installation was successful.

About this task



Note: If you have a secure cluster, use the kinit command line tool to authenticate to Kerberos.

Procedure

1. Log into a host in the cluster.
2. Run the Hadoop PiEstimator example using the following command:

```
yarn jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 10 100
```

3. In Cloudera Manager, navigate to **Cluster** *ClusterName* **yarn Applications**.
4. Check the results of the job.

You will see an entry like the following:

05/22/2014	10:45 AM	-	Name: QuasiMonteCarlo	Pool: root.hdfs	
05/22/2014	10:46 AM		Mapper: QuasiMonteCarlo\$QmcMapper	Reducer: QuasiMonteCarlo\$QmcReducer	<div> <div>Actions ▾</div> <div>Details</div> </div>
Type: MapReduce ID: job_1400700704311_0001 Duration: 54.27s User: hdfs CPU Time: 34.15s File Bytes Read: 98 B File Bytes Written: 992.7 KiB HDFS Bytes Read: 2.7 KiB HDFS Bytes Written: 215 B Memory Allocation: 184.7M Pool: root.hdfs					

Testing with Hue

You can test the cluster by running Hue.

About this task

Hue is a graphical user interface that allows you to interact with your clusters by running applications that let you browse HDFS and cloud object storage such as S3 and ABFS, manage a Hive metastore, and run Hive, Impala, and Search queries, and Oozie workflows.

Procedure

1. From Cloudera Manager, go to **Clusters Hue service**.
2. Click **Web UI** link and select the Hue web URL, which opens Hue in a new window.

By default, Authentication Backend is set to **AllowFirstUserDjangoBackend**. This makes the first user who logs into Hue the Superuser and allows you to set the username and password, and create other users.

You can change the Authentication Backend as per your requirements from Hue configurations in Cloudera Manager.

3. You can run a query or browse the database that you have set up for Hue.

For more information, see the Hue documentation.

Secure Your Cluster

After completing your Cloudera Enterprise installation and making sure that everything is working properly, secure your cluster by enabling authentication, authorization, auditing, and encryption.

For comprehensive instructions on securing your cluster, see the Security documentation.

Related Information

[Security Overview](#)

Troubleshooting Installation Problems

This topic describes common installation issues and suggested solutions.

Failed to start server reported by cloudera-manager-installer.bin

"Failed to start server" reported by cloudera-manager-installer.bin. /var/log/cloudera-scm-server/cloudera-scm-server.log contains a message beginning Caused by: java.lang.ClassNotFoundException: com.mysql.jdbc.Driver...

Possible reason:

You might have SELinux enabled.

Possible solution:

Disable SELinux by running `sudo setenforce 0` on the Cloudera Manager Server host. To disable it permanently, edit /etc/selinux/config.

Installation interrupted and installer does not restart

Possible reason:

You need to do some manual cleanup.

Possible solution:

See *Uninstalling Cloudera Manager and Managed Software*.

Cloudera Manager Server fails to start with MySQL

Cloudera Manager Server fails to start and the Server is configured to use a MySQL database to store information about service configuration.

Possible reason:

Tables might be configured with the ISAM engine. The Server does not start if its tables are configured with the MyISAM engine, and an error such as the following appears in the log file:

```
Tables ... have unsupported engine type ... . InnoDB is required.
```

Possible solution:

Make sure that the InnoDB engine is configured, not the MyISAM engine. To check what engine your tables are using, run the following command from the MySQL shell: `mysql> show table status;`

For more information, see [Install and Configure MySQL for Cloudera Software](#) on page 82.

Agents fail to connect to Server

Agents fail to connect to Server. You get an Error 113 ('No route to host') in /var/log/cloudera-scm-agent/cloudera-scm-agent.log.

Possible reason:

You might have SELinux or iptables enabled.

Possible solution:

Check /var/log/cloudera-scm-server/cloudera-scm-server.log on the Server host and /var/log/cloudera-scm-agent/cloudera-scm-agent.log on the Agent hosts. Disable SELinux and iptables.

Cluster hosts do not appear

Some cluster hosts do not appear when you click Find Hosts in install or update wizard.

Possible reason:

You might have network connectivity problems.

Possible solution:

- Make sure all cluster hosts have SSH port 22 open.
- Check other common causes of loss of connectivity such as firewalls and interference from SELinux.

"Access denied" in install or update wizard

"Access denied" in install or update wizard during database configuration for Activity Monitor or Reports Manager.

Possible reason:

Hostname mapping or permissions are not set up correctly.

Possible solution:

- For hostname configuration, see *Configure Network Names*.
- For permissions, make sure the values you enter into the wizard match those you used when you configured the databases. The value you enter into the wizard as the database hostname must match the value you entered for the hostname (if any) when you configured the database.

For example, if you had entered the following when you created the database

```
grant all on activity_monitor.* TO 'amon_user'@'myhost1.myco.com' IDENTIFIED BY 'amon_password';
```

the value you enter here for the database hostname must be myhost1.myco.com. If you did not specify a host, or used a wildcard to allow access from any host, you can enter either the fully qualified domain name (FQDN), or localhost. For example, if you entered

```
grant all on activity_monitor.* TO 'amon_user'@'%' IDENTIFIED BY 'amon_password';
```

the value you enter for the database hostname can be either the FQDN or localhost.

Databases fail to start.

Activity Monitor, Reports Manager, or Service Monitor databases fail to start.

Possible reason:

MySQL binlog format problem.

Possible solution:

Set `binlog_format=mixed` in `/etc/my.cnf`. For more information, see [this MySQL bug report](#). See also [Step 4. Install and Configure Databases](#) on page 76.

Cloudera services fail to start

Possible reason:

Java might not be installed or might be installed at a custom location.

Possible solution:

See *Configuring a Custom Java Home Location* for more information on resolving this issue.

Activity Monitor displays a status of BAD

The Activity Monitor displays a status of BAD in the Cloudera Manager Admin Console. The log file contains the following message:

```
ERROR 1436 (HY000): Thread stack overrun: 7808 bytes used of a 131072 byte stack, and 128000 bytes needed.
Use 'mysqld -O thread_stack=#' to specify a bigger stack.
```

Possible reason:

The MySQL thread stack is too small.

Possible solution:

1. Update the `thread_stack` value in `my.cnf` to 256KB. The `my.cnf` file is normally located in `/etc` or `/etc/mysql`.
2. Restart the `mysql` service: `$ sudo service mysql restart`
3. Restart Activity Monitor.

Activity Monitor fails to start

The Activity Monitor fails to start. Logs contain the error read-committed isolation not safe for the statement binlog format.

Possible reason:

The `binlog_format` is not set to mixed.

Possible solution:

Modify the `mysql.cnf` file to include the entry for binlog format as specified in *Install and Configure MySQL for Cloudera Software*.

Create Hive Metastore Database Tables command fails

The Create Hive Metastore Database Tables command fails due to a problem with an escape string.

Possible reason:

PostgreSQL versions 9 and higher require special configuration for Hive because of a backward-incompatible change in the default value of the `standard_conforming_strings` property. Versions up to PostgreSQL 9.0 defaulted to off, but starting with version 9.0 the default is on.

Possible solution:

As the administrator user, use the following command to turn `standard_conforming_strings` off:

```
ALTER DATABASE <hive_db_name> SET standard_conforming_strings = off;
```

Oracle invalid identifier

If you are using an Oracle database and the Cloudera Navigator AnalyticsAuditActivity tab displays "No data available" and there is an Oracle error about "invalid identifier" with the query containing the reference to `dbms_crypto` in the log.

Possible reason:

You have not granted execute permission to `sys.dbms_crypto`.

Possible solution:

Run `GRANT EXECUTE ON sys.dbms_crypto TO nav;`, where `nav` is the user of the Navigator Audit Server database.

Related Information

[CDP Private Cloud Base Installation Guide](#)

Uninstalling Cloudera Manager and Managed Software

Complete the following tasks to uninstall the Cloudera Manager Server, Agents, managed software, and databases.

Related Information

[CDP Private Cloud Base Installation Guide](#)

Record User Data Paths

Record the location of the user data paths by checking the configuration in each service.

The user data paths listed in the topic *Remove User Data*, `/var/lib/flume-ng` `/var/lib/hadoop*` `/var/lib/hue` `/var/lib/navigator` `/var/lib/oozie` `/var/lib/solr` `/var/lib/sqoop*` `/var/lib/zookeeper` `data_drive_path/dfs` `data_drive_path/mapred` `data_drive_path/yarn`, are the default settings. However, at some point they might have been reconfigured in Cloudera Manager. If you want to remove all user data from the cluster and have changed the paths, either when you installed Runtime and managed services or at some later time, note the location of the paths by checking the configuration in each service.

Stop all Services

Stop all services for each cluster managed by Cloudera Manager.

Procedure

1. On the HomeStatus tab, click three dots to the right of the cluster name and select Stop.
2. Click Stop in the confirmation screen. The Command Details window shows the progress of stopping services. When All services successfully stopped appears, the task is complete and you can close the Command Details window.
3. On the HomeStatus tab, click the three dots to the right of the Cloudera Management Service entry and select Stop. The Command Details window shows the progress of stopping services.

Results

When All services successfully stopped appears, the task is complete and you can close the Command Details window.

Deactivate and Remove Parcels

If you installed using packages, skip this step and go to *Uninstall the Cloudera Manager Server*; you will remove packages in *Uninstall Cloudera Manager Agent and Managed Software*. If you installed using parcels remove them as follows:

Procedure

1.



Click the parcel indicator in the left-hand navigation bar.

2. In the Location selector on the left, select All Clusters.
3. For each activated parcel, select ActionsDeactivate. When this action has completed, the parcel button changes to Activate.
4. For each activated parcel, select ActionsRemove from Hosts. When this action has completed, the parcel button changes to Distribute.
5. For each activated parcel, select ActionsDelete. This removes the parcel from the local parcel repository.

What to do next

There might be multiple parcels that have been downloaded and distributed, but that are not active. If this is the case, you should also remove those parcels from any hosts onto which they have been distributed, and delete the parcels from the local repository.

Delete the Cluster

On the Home page, Click the drop-down list next to the cluster you want to delete and select Delete.

Uninstall the Cloudera Manager Server

The commands for uninstalling the Cloudera Manager Server depend on the method you used to install it. Refer to steps below that correspond to the method you used to install the Cloudera Manager Server.

Procedure

1. If you used the cloudera-manager-installer.bin file (the trial installer): Run the following command on the Cloudera Manager Server host:
2. If you did not use the cloudera-manager-installer.bin file: If you installed the Cloudera Manager Server using a different installation method such as Puppet, run the following commands on the Cloudera Manager Server host:
 - a) Stop the Cloudera Manager Server and its database:

```
sudo /usr/share/cmf/uninstall-cloudera-manager.sh
```

```
sudo service cloudera-scm-server stop  
sudo service cloudera-scm-server-db stop
```

- b) Uninstall the Cloudera Manager Server and its database. This process described also removes the embedded PostgreSQL database software, if you installed that option. If you did not use the embedded PostgreSQL database, omit the cloudera-manager-server-db steps.

RHEL

```
sudo yum remove cloudera-manager-server  
sudo yum remove cloudera-manager-server-db-2
```

Uninstall Cloudera Manager Agent and Managed Software

To uninstall Cloudera Manager Agent and managed software, stop the Cloudera Manager Agent on all hosts, remove the parcel installation, and run the clean command.

About this task

Do the following on all Agent hosts:

Procedure

1. Stop the Cloudera Manager Agent.

```
sudo systemctl stop supervisord
```

- To uninstall managed software, run the following commands:

RHEL: \$ sudo yum remove 'cloudera-manager-*' Too difficult/impossible to hide entire rows and columns in this table, so I adding it to a draft comment for future use when packages and other OS's are supported. For DC 7.0 we only have RHEL compatible and Runtime parcel installs.

RHEL

```
sudo yum remove 'cloudera-manager-*
```

SLES

```
sudo zypper remove 'cloudera-manager-*
```

Ubuntu

```
sudo apt-get purge 'cloudera-manager-*
```

- Run the clean command:

RHEL

```
sudo yum clean all
```

SLES

```
sudo zypper clean
```

Ubuntu

```
sudo apt-get clean
```

Remove Cloudera Manager, User Data, and Databases

Permanently remove Cloudera Manager data, the Cloudera Manager lock file, and user data. Then stop and remove the databases.

Procedure

- On all Agent hosts, kill any running Cloudera Manager and managed processes:

```
for u in cloudera-scm flume hadoop hdfs hbase hive httpfs hue impala llama
mapred oozie solr spark sqoop sqoop2 yarn zookeeper; do sudo kill $(ps -u
$u -o pid=); done
```



Note: This step should not be necessary if you stopped all the services and the Cloudera Manager Agent correctly.

- If you are uninstalling on RHEL, run the following commands on all Agent hosts to permanently remove Cloudera Manager data. If you want to be able to access any of this data in the future, you must back it up before removing it. If you used an embedded PostgreSQL database, that data is stored in /var/lib/cloudera-scm-server-db.

```
sudo umount cm_processes
sudo rm -Rf /usr/share/cmf /var/lib/cloudera* /var/cache/yum/cloudera* /
var/log/cloudera* /var/run/cloudera*
```

- On all Agent hosts, run this command to remove the Cloudera Manager lock file:

```
sudo rm /tmp/.scm_prepare_node.lock
```

4. This step permanently removes all user data. To preserve the data, copy it to another cluster using the `distcp` command before starting the uninstall process.

a) On all Agent hosts, run the following commands:

```
sudo rm -Rf /var/lib/flume-ng /var/lib/hadoop* /var/lib/hue /var/lib/navigator /var/lib/oozie /var/lib/solr /var/lib/sqoop* /var/lib/zookeeper
```

b) Run the following command on each data drive on all Agent hosts (adjust the paths for the data drives on each host):

```
sudo rm -Rf data_drive_path/dfs data_drive_path/mapred data_drive_path/yarn
```

5. Stop and remove the databases. If you chose to store Cloudera Manager or user data in an external database, see the database vendor documentation for details on how to remove the databases.

Uninstalling a Runtime Component From a Single Host

The following procedure removes Runtime software components from a single host that is managed by Cloudera Manager.

Procedure

1. In the Cloudera Manager Administration Console, select **Hosts**All Hosts. A list of hosts in the cluster displays.
2. Select the host where you want to uninstall Runtime software.
3. Click the **Actions for Selected** button and select **Remove From Cluster**. Cloudera Manager removes the roles and host from the cluster.
4. Optionally, manually delete the `krb5.conf` file used by Cloudera Manager.