cloudera
Ask Bigger Questions

Cloudera Manager Installation Guide

Important Notice

Cloudera, Inc.
1001 Page Mill Road Bldg 2
Palo Alto, CA 94304
info@cloudera.com
US: 1-888-789-1488
Intl: 1-650-362-0488
www.cloudera.com

Release Information

Version: 5.1.x
Date: December 4, 2014

Table of Contents

About this Guide

This guide explains how to install Cloudera Manager and CDH. Cloudera Manager 5 supports managing CDH 4 and CDH 5.

Introduction to Cloudera Manager Installation

Cloudera Manager automates the installation and configuration of CDH and managed services on a cluster, requiring only that you have root SSH access to your cluster's hosts, and access to the internet or a local repository with installation files for all these hosts. Cloudera Manager installation software consists of:

- A small self-executing Cloudera Manager installation program to install the Cloudera Manager Server and other packages in preparation for host installation.
- Cloudera Manager wizard for automating CDH and managed service installation and configuration on the cluster hosts. Cloudera Manager provides two methods for installing CDH and managed services: traditional packages (RPMs or Debian packages) or parcels. Parcels simplify the installation process, and more importantly allows you to download, distribute, and activate new minor versions of CDH and managed services from within Cloudera Manager.

About the Cloudera Manager Installation Program

The Cloudera Manager installation program, which you install on the host where you want the Cloudera Manager Server to run:

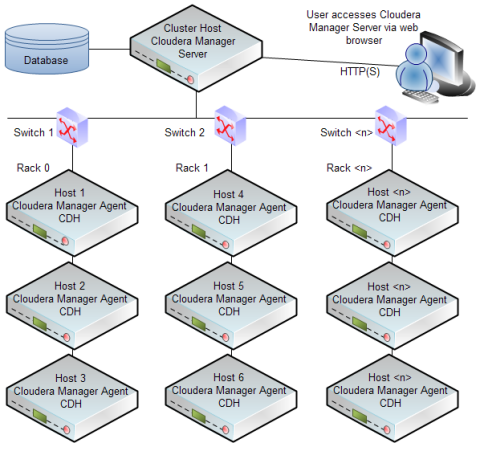
- Installs the package repositories for Cloudera Manager and the Oracle Java Development Kit (JDK) 1.7
- Installs the Cloudera Manager Server
- Installs and configures an embedded PostgreSQL database for use by the Cloudera Manager Server and some Cloudera Management Service roles, Hive Metastore, and Cloudera Navigator Audit Server

Installing Cloudera Manager for the First Time

To install Cloudera Manager, you:

- Optionally install a database application on the Cloudera Manager Server host or on a host that the Cloudera Manager Server can access, and (depending on the configuration you decide on) on other hosts as well.
- Run the Cloudera Manager installer on one host.
- Install CDH, managed services, and Cloudera Manager Agents on the other hosts.

The following illustrates a sample installation:



About the Cloudera Manager "First Run" Wizard

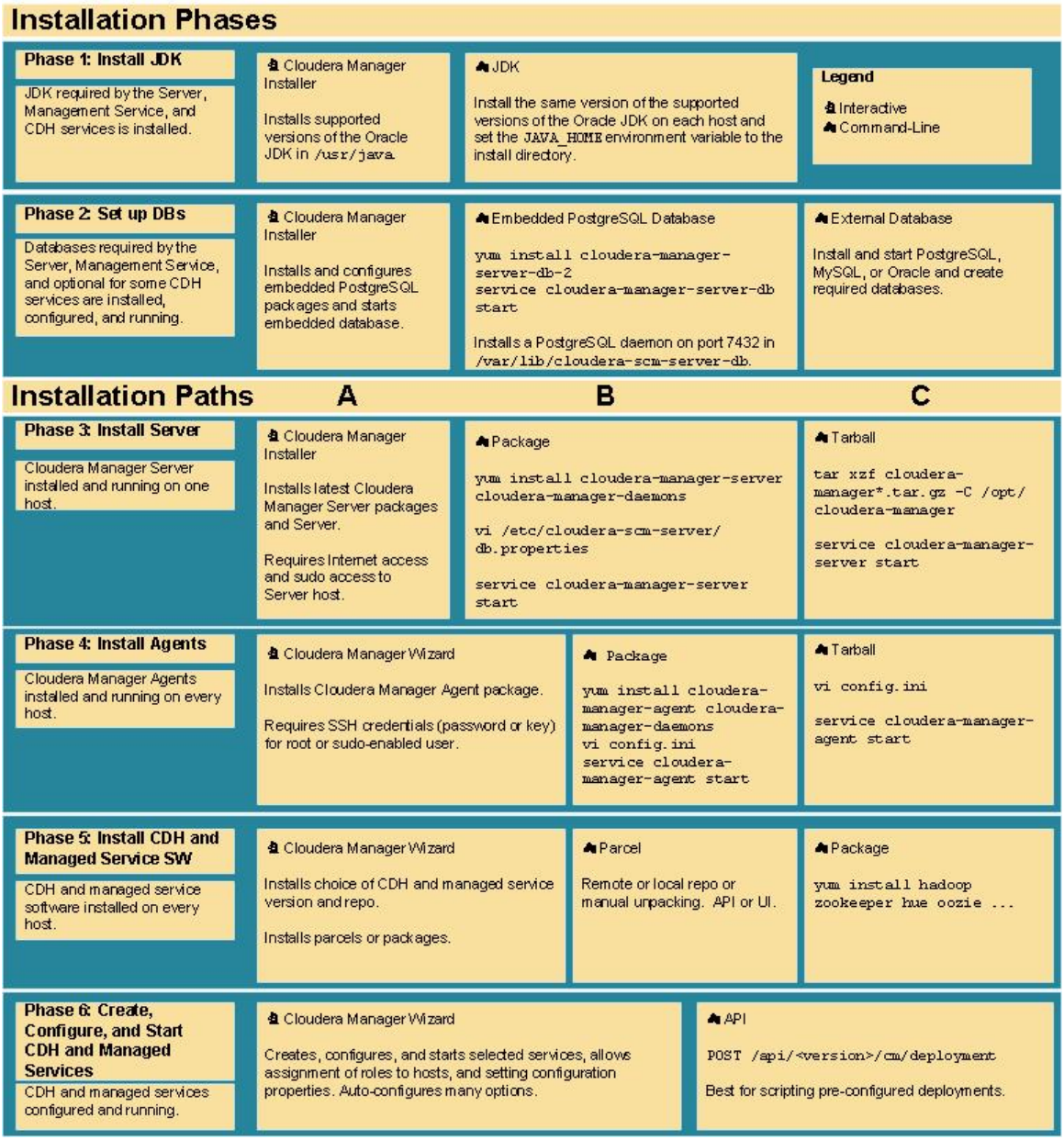
After you install Cloudera Manager and you connect to the Cloudera Manager Admin Console for the first time, you use the Cloudera Manager "first run" wizard to do the following:

- Discover cluster hosts
- Optionally install the Oracle JDK
- Optionally install CDH, managed service, and Cloudera Manager Agent software on the hosts
- Select which services to run
- Specify the mapping of service roles to hosts
- Confirm service configurations and start the services

You can choose to abort the software installation process and the Cloudera Manager wizard automatically reverts and completely rolls back the installation process for any uninstalled components. (Installation that has completed successfully on a given host is not rolled back on that host.)

Installation Phases and Paths for Cloudera Manager, CDH, and Managed Services

The following diagram illustrates the phases required to install Cloudera Manager, CDH, and managed services. Every phase is required, but there are multiple ways to accomplish each phase, depending on your organization's policies and requirements.



The six phases are grouped into three installation paths based on how the Cloudera Manager Server and database software is installed on the Cloudera Manager Server and cluster hosts. For each path you can choose to use the embedded PostgreSQL database or an external database. To review the criteria for choosing a path, see Installing Cloudera Manager, CDH, and Managed Services on page 43.

Cloudera Manager Requirements

Cloudera Manager interacts with a diversity of entities such as operating systems, databases, and browsers. This topic provides information about which major release version and minor release version of each entity is supported. In some cases, such as some browsers, a minor version may not be provided. After installing each entity, upgrade to the latest patch version and apply any other appropriate updates. The available updates may be specific to the operating system on which it is installed. For example, you might be using CentOS in your environment. You could choose 6 as the major version and 4 as the minor version. These choices would mean you would be using CentOS 6.4. After installing this operating system, apply any and all relevant CentOS 6.4 upgrades and patches.

The following sections describe various requirements for Cloudera Manager.

Supported Operating Systems

Cloudera Manager supports the following operating systems:

- RHEL-compatible systems
  - Red Hat Enterprise Linux and CentOS 5.7, 64-bit
  - Red Hat Enterprise Linux and CentOS 6.4, 64-bit
  - Red Hat Enterprise Linux and CentOS 6.4 in SE Linux Mode
  - Red Hat Enterprise Linux and CentOS 6.5, 64-bit
  - Oracle Enterprise Linux 5.6 (UEK R2), 64-bit
  - Oracle Enterprise Linux 6.4 (UEK R2), 64-bit
  - Oracle Enterprise Linux 6.5 (UEK R2, UEK R3), 64-bit
- SLES - SUSE Linux Enterprise Server 11, 64-bit. Service Pack 2 or later is required for CDH 5 and Service Pack 1 or later is required for CDH 4. To use the embedded PostgreSQL database that is installed when you follow Installation Path A - Automated Installation by Cloudera Manager, the Updates repository must be active. The SUSE Linux Enterprise Software Development Kit 11 SP1 is required on hosts running the Cloudera Manager Agents.
- Debian - Debian 7.0 and 7.1, 6.0 (deprecated), 64-bit
- Ubuntu - Ubuntu 12.04, 10.04 (deprecated), 64-bit

> **Note:**
>
> - Debian 6.0 and Ubuntu 10.04 are supported only for CDH 4.
> - Using the same version of the same operating system on all cluster hosts is strongly recommended.

Supported JDK Versions

Cloudera Manager supports Oracle JDK 7u55 and Oracle JDK 6u31, and installs them during installation and upgrade.

Supported Browsers

The Cloudera Manager Admin Console, which you use to install, configure, manage, and monitor services, supports the following browsers:

- Firefox 11 or later

- Google Chrome
- Internet Explorer 9 or later
- Safari 5 or later

Supported Databases

Cloudera Manager requires several databases. The Cloudera Manager Server stores information about configured services, role assignments, configuration history, commands, users, and running processes in a database of its own. You must also specify a database for the Activity Monitor and Reports Manager management services.

The database you choose to use must be configured to support UTF8 character set encoding. The embedded PostgreSQL database that is installed when you follow Installation Path A - Automated Installation by Cloudera Manager on page 44 automatically provides UTF8 encoding. If you install a custom database, you may need to enable UTF8 encoding. The commands for enabling UTF8 encoding are described in each database's section under Cloudera Manager and Managed Service Databases on page 21.

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

Cloudera Manager and its supporting services can use the following databases:

- MySQL - 5.0, 5.1, 5.5, and 5.6
- Oracle 11gR2
- PostgreSQL - 8.4, 9.1, and 9.2

For information about the databases supported by CDH, see CDH 4 Supported Databases and CDH 5 Supported Databases.

Supported CDH and Managed Service Versions

The following versions of CDH and managed services are supported:

- **Warning:** Cloudera Manager 5 does not support CDH 3 and you cannot upgrade Cloudera Manager 4 to Cloudera Manager 5 if you have a cluster running CDH 3. Therefore, to upgrade CDH 3 clusters to CDH 4 using Cloudera Manager you must use Cloudera Manager 4.

- **CDH 4 and CDH 5**. The latest released versions of CDH 4 and CDH 5 are strongly recommended. For information on CDH 4 requirements, see CDH 4 Requirements and Supported Versions. For information on CDH 5 requirements, see CDH 5 Requirements and Supported Versions.
- **Cloudera Impala** - Cloudera Impala is included with CDH 5. Cloudera Impala 1.2.1 with CDH 4.1.0 or later. For further information on Cloudera Impala requirements with CDH 4, see Cloudera Impala Requirements.
- **Cloudera Search** - Cloudera Search is included with CDH 5. Cloudera Search 1.2.0 with CDH 4.6.0. For further information on Cloudera Search requirements with CDH 4, see Cloudera Search Requirements.
- **Apache Spark** - 0.90 or later with CDH 4.4.0 or later.
- **Apache Accumulo** - 1.4.3 with CDH 4.3.0, 1.4.4 with CDH 4.5.0, and 1.6.0 with CDH 4.6.0.

For more information, see the Cloudera Product Compatibility Matrix.

Resource Requirements

Cloudera Manager requires resources of the following types:

- **Disk Space**
  - **Cloudera Manager Server**

  - 5 GB on the partition hosting `/var`.
  - 500 MB on the partition hosting `/usr`.
  - For parcels, the space required depends on the number of parcels you download to the Cloudera Manager Server and distribute to Agent hosts. You can download multiple parcels of the same product, of different versions and builds. If you are managing multiple clusters, there will be only one parcel of a given product/version/build/distribution downloaded on the Cloudera Manager Server—not one per cluster. In the local parcel repository on the Cloudera Manager Server the approximate sizes of the various parcels are as follows:
    - CDH 4.6 - ~700 MB per parcel, CDH 5 - ~1 GB per parcel
    - Impala - ~200 MB per parcel
    - Solr - ~ 400 MB per parcel

  - **Cloudera Management Service** - The Host Monitor and Service Monitor databases are stored on the partition hosting `/var`. Ensure that you have at least 20 GB available on this partition. For further information, see Monitoring Data Storage on page 38.
  - **Agents** - On Agent hosts each unpacked parcel requires about three times the space of the downloaded parcel on the Cloudera Manager Server. By default unpacked parcels are located in `/opt/cloudera/parcels`.

- **RAM** - 4 GB is appropriate for most cases, and is required when using Oracle databases. 2 GB may be sufficient for non-Oracle deployments involving fewer than 100 hosts. However, if you want to run the Cloudera Manager Server on a machine with 2 GB of RAM, you must tune down its maximum heap size (by modifying `-Xmx` in `/etc/default/cloudera-scm-server`). Otherwise the kernel may kill the Server for consuming too much RAM.
- **Python** - Cloudera Manager uses Python. All supported operating systems contain a Python version 2.4 or higher. Cloudera Manager and CDH 4 require at least Python 2.4, but Hue in CDH 5 requires Python 2.6 or 2.7.

Networking and Security 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 host names and addresses across all hosts. A properly formatted `/etc/hosts` 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
  ```

  The `/etc/hosts` file *must not* have duplicate IP addresses.
- 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 your 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 once the installation is complete. For further information, see Permission Requirements on page 18.
- The Cloudera Manager Agent runs as root so that it can make sure the required directories are created and that processes and files are owned by the appropriate user (for example, the `hdfs` and `mapred` users).
- No blocking by Security-Enhanced Linux (SELinux).

- Disable Ipv6 on all hosts.
- No blocking by iptables or firewalls; make sure port 7180 is open because it is the port used to access Cloudera Manager after installation. Cloudera Manager communicates using specific ports, which must be open. See Configuring Ports for Cloudera Manager on page 135.
- For RedHat and CentOS, make sure the `/etc/sysconfig/network` file on each system contains the hostname you have just set (or verified) for that system.
- Cloudera Manager and CDH use several user accounts and groups to complete their tasks. The set of user accounts and groups varies according to which components you choose to install. Do not delete these accounts or groups and do not modify their permissions and rights. Ensure no existing systems obstruct the functioning of these accounts and groups. For example, if you have scripts that delete user accounts not in a white-list, add these accounts to the list of permitted accounts. Cloudera Manager, CDH, and managed services create and use the following accounts and groups:

Account	Type	Product
`cloudera-scm`	User and group	Cloudera Manager
`flume`	User and group	CDH 4, CDH 5
`hadoop`	Group	CDH 4, CDH 5
`hbase`	User and group	CDH 4, CDH 5
`hdfs`	User and group. Must also be a member of the `hadoop` group.	CDH 4, CDH 5
`hive`	User and group	CDH 4, CDH 5
`httpfs`	User and group	CDH 4, CDH 5
`hue`	User and group	CDH 4, CDH 5
`impala`	User and group. Must also be member of the `hdfs` and `hive` groups.	CDH 4.1 or later, CDH 5
`llama`	User and group	CDH 5
`mapred`	User and group. Must also be a member of the `hadoop` group.	CDH 4, CDH 5
`oozie`	User and group	CDH 4, CDH 5
`solr`	User and group	CDH 4.3 and later, CDH 5
`spark`	User and group	Spark, CDH 5
`sqoop`	User and group	CDH 4, CDH 5
`sqoop2`	User. Must be member of the `sqoop` group.	CDH 4.2 and later, CDH 5
`yarn`	User and group	CDH 4, CDH 5
`zookeeper`	User and group	CDH 4, CDH 5

Permission Requirements

The following sections describe the permission requirements for package-based installation and upgrades of CDH with and without Cloudera Manager. The permission requirements are not controlled by Cloudera but result from standard UNIX system requirements for the installation and management of packages and running services.

Permission Requirements for Package-Based CDH Installation with Cloudera Manager

> **Important:** Unless otherwise noted, when root and/or sudo access is required, using another system (such as PowerBroker) that provides root/sudo privileges is acceptable.

- Installation of Cloudera Manager (via `cloudera-manager-installer.bin`) requires root and/or sudo access on a single host.
- Manual start/stop/restart of the Cloudera Manager Server (that is, logging onto the host running Cloudera Manager and executing: `service cloudera-scm-server` *`action`*) requires the use of root and/or sudo.
- A running instance of Cloudera Manager Server does not require root and/or sudo access, as the Server is run under the user `cloudera-scm`
- Installation of CDH components through Cloudera Manager requires the use of one of the following, as configured during the initial installation of Cloudera Manager:
  - Direct access to root user via the root password.
  - Direct access to root user using a SSH key file.
  - Passwordless sudo access for a specific user. This is the same requirement as the installation of CDH components on individual hosts, which is a requirement of the UNIX system in general.

  Using another system (such as PowerBroker) that provides root/sudo privileges is *not* acceptable.
- Cloudera Manager uses a process called the Cloudera Manager Agent on each host that is being managed. Installation of the Cloudera Manager Agent through Cloudera Manager requires the use of one of the following, as configured during the initial installation of Cloudera Manager:
  - Direct access to root user via the root password.
  - Direct access to root user using a SSH key file.
  - Passwordless sudo access for a specific user. This is the same requirement as the installation of CDH components on individual hosts, which is a requirement of the UNIX system in general.

  Using another system (such as PowerBroker) that provides root/sudo privileges is *not* acceptable.
- The Cloudera Manager Agent requires access to the root user account at runtime. This is achieved via three scenarios:
  - During Cloudera Manager and CDH installation on a given host, the Agent is automatically started upon a successful installation. It is then started via one of the following, as configured during the initial installation of Cloudera Manager:
    - Direct access to root user via the root password
    - Direct access to root user using a SSH key file
    - Passwordless sudo access for a specific user

    Using another system (such as PowerBroker) that provides root/sudo privileges is *not* acceptable.
  - Via automatic startup during system boot, via init.
  - Manual start/stop/restart of the Agent process requires root and/or sudo access. This permission requirement is to ensure that services managed by the Cloudera Manager Agent on any given host assume the appropriate user (that is, the HDFS service assumes the `hdfs` user) for correct privileges. Any action request for a CDH service managed within Cloudera Manager *does not* require root and/or sudo access, as the action(s) are handled by the Cloudera Manager Agent which is already running under the root user.

Permission Requirements for Package-Based CDH Installation without Cloudera Manager

- Installation of CDH products requires root and/or sudo access for the installation of any RPM based package during the time of installation and service startup/shut down:
  - Passwordless SSH under the root user is not required for the installation (SSH root keys)
- Upgrading previously installed CDH packages requires root and/or sudo access to be completed:
  - Passwordless SSH under the root user is not required for the upgrade process (SSH root keys)

- Cloudera recommends passwordless SSH as root (SSH root keys) for simplicity of manually installing and/or upgrading hosts within a CDH ready cluster for the following reasons:
  - Scripts can be created to assist in CDH package management across the cluster
  - Scripts can be created to assist in configuration management across the cluster
- Any changes to the CDH package, including RPM upgrades, configuration changes that require CDH service restarts, or adding CDH services require the use of root and/or sudo access to restart any host impacted by this change, which could lead to a restart of a given service on each host in the cluster.
- Start/stop/restart actions against a CDH service require the use of root and/or sudo per UNIX standards.

Cloudera Manager and Managed Service Databases

Cloudera Manager uses databases to store information about the Cloudera Manager configuration, as well as information such as the health of the system or task progress. To facilitate rapid completion of simple installations, the Cloudera Manager can install and configure an embedded PostgreSQL database as part of the Cloudera Manager installation process. This automatically installed database is referred to as an embedded PostgreSQL database. In addition, some CDH services use databases and are automatically configured to use a default database. If you plan to use the embedded and default databases provided during the Cloudera Manager installation, see Installation Path A - Automated Installation by Cloudera Manager on page 44.

While the embedded database is a useful option for getting started quickly, Cloudera Manager also allows you to opt to use your own PostgreSQL, MySQL, or Oracle database for the Cloudera Manager Server and services that use databases. To learn more about database options or if you are unsure whether or not using the embedded database is right for your environment, continue with the following sections.

What Databases Must Be Installed

The Cloudera Manager Server, Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server all require databases:

- Cloudera Manager - Contains all the information about what services you have configured, their role assignments, all configuration history, commands, users, and running processes. This is a relatively small database (<100 MB), and is the most important to back up. A monitoring database contains monitoring information about service and host status. In large clusters, this database can grow large.
- Activity Monitor - Contains information about past activities. In large clusters, this database can grow large.
- Reports Manager - Keeps track of disk utilization and processing activities over time. Medium-sized.
- Hive Metastore - Contains Hive metadata. Relatively small.
- Sentry Server - Contains authorization metadata. Relatively small.
- Cloudera Navigator Audit Server - Contains auditing information. In large clusters, this database can grow large.

The Host Monitor and Service Monitor have an internal datastore. Configuring an Activity Monitor database is only necessary if there's a MapReduce service in the deployment.

Cloudera Manager provides three install paths:

- Path A automatically installs an embedded PostgreSQL database to meet the requirements of the services. This path reduces the number of installation tasks you must complete, as well as the number of choices to make. In Path A you can also optionally choose to create external databases for Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server.
- Path B and Path C require you to create databases for the Cloudera Manager Server, Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server.

Using an external database requires more input and intervention as you either install databases or gather information about existing databases. These paths also provides greater flexibility in choosing database types and configurations.

Cloudera Manager supports deploying different types of databases in a single environment, but doing so may create unexpected complications. Cloudera recommends choosing one of the supported database providers to use for all of the Cloudera databases.

In most cases, you should install databases and services on the same host. For example, if you create the database for Activity Monitor on `myhost1`, then you should typically assign the Activity Monitor role to `myhost1`. You will assign the Activity Monitor and Reports Manager roles in the Cloudera Manager wizard during the install or upgrade process. After completing the install or upgrade process, you can also modify role assignments in the Management services pages of Cloudera Manager. While it is true that database location is changeable,

before beginning an installation or upgrade, you should decide which hosts you will use. The JDBC connector for your database *must* be installed on the hosts where you assign the Activity Monitor and Reports Manager roles.

It is possible to install the database and services on different hosts. Separating databases from services is more likely to occur in larger deployments and in cases where more sophisticated database administrators actively choose to establish such a configuration. For example, databases and services might be separated if your environment includes Oracle databases that will be separately managed by Oracle database administrators.

Setting up the Cloudera Manager Server Database

The Cloudera Manager Server database stores information about service and host configurations. You can use an embedded PostgreSQL database or an external database.

Installing and Starting the Cloudera Manager Server Embedded Database

If you are using Installation Path B - Manual Installation Using Cloudera Manager Packages on page 50 for a demonstration or proof of concept deployment, and you want to use an embedded PostgreSQL database for the Cloudera Management Server, use this procedure to install and start the database:

1. Install the embedded PostgreSQL database packages:
   - **Red Hat-compatible, if you have a yum repo configured:**

     ```
     $ sudo yum install cloudera-manager-server-db-2
     ```
   - **Red Hat-compatible, if you're transferring RPMs manually:**

     ```
     $ sudo yum --nogpgcheck localinstall cloudera-manager-server-db-2.noarch.rpm
     ```
   - **SLES:**

     ```
     $ sudo zypper install cloudera-manager-server-db-2
     ```
   - **Debian/Ubuntu**

     ```
     $ sudo apt-get install cloudera-manager-server-db-2
     ```
2. Start the PostgreSQL database:

   ```
   $ sudo service cloudera-scm-server-db start
   ```

Preparing an Cloudera Manager Server External Database

Before performing these steps, install and configure a database as described in MySQL Database on page 30, Oracle Database on page 35, or External PostgreSQL Database on page 27.

1. Run the `scm_prepare_database.sh` script:
   - Installer or package install

     ```
     /usr/share/cmf/schema/scm_prepare_database.sh
     ```
   - Tarball install

     ```
     <tarball root>/share/cmf/schema/scm_prepare_database.sh
     ```

   on the host where the Cloudera Manager Server package is installed. The script prepares the database by:

- Creating the Cloudera Manager Server database configuration file.
- Creating a database for the Cloudera Manager Server to use. This is optional and is only completed if options are specified.
- Setting up a user account for the Cloudera Manager Server. This is optional and is only completed if options are specified.

2. Remove the embedded PostgreSQL properties file:
   - Installer or package install

     ```
     /etc/cloudera-scm-server/db.mgmt.properties
     ```

   - Tarball install

     ```
     <tarball root>/etc/cloudera-scm-server/db.mgmt.properties
     ```

   if it exists.

scm_prepare_database.sh Syntax

```
scm_prepare_database.sh database-type [options] database-name username password
```

- **Note:** You can also run `scm_prepare_database.sh` without options to see the syntax.

Table 1: Required Parameters

Parameter	Description
`database-type`	One of the supported database types: ▪ MySQL - `mysql` ▪ Oracle - `oracle` ▪ PostgreSQL - `postgresql`
`database-name`	The name of the Cloudera Manager Server database you want to create or use.
`username`	The username for the Cloudera Manager Server database you want to create or use.
`password`	The password for the Cloudera Manager Server database you want to create or use. If you don't specify the password on the command line, the script will prompt you to enter it.

Table 2: Options

Option	Description
`-h or --host`	The IP address or hostname of the host where the database is installed. The default is to use the local host.
`-P or --port`	The port number to use to connect to the database. The default port is 3306 for MySQL, 5432 for PostgreSQL, and 1521 for Oracle. This option is used for a remote connection only.
`-u or --user`	The admin username for the database application. For -u, there should not be a space between the option and the provided value. If this option is supplied, the script will create a user and database for the Cloudera Manager Server; otherwise, it will use the existing user and database you created previously.

Option	Description
`-p or --password`	The admin password for the database application. The default is no password. For -p, there should not be a space between the option and the provided value.
`--scm-host`	The hostname where the Cloudera Manager Server is installed. Omit if the Cloudera Manager server and the database are installed on the same host.
`--config-path`	The path to the Cloudera Manager Server configuration files. The default is `/etc/cloudera-scm-server`.
`--schema-path`	The path to the Cloudera Manager schema files. The default is `/usr/share/cmf/schema` (the location of the script).
`-f`	The script will not stop if an error is encountered.
`-? or --help`	Display help.

Example 1: Running the script when MySQL is installed on another host

This example explains how to run the script on the Cloudera Manager Server host (myhost2) and create and use a temporary MySQL user account to connect to MySQL remotely on the MySQL host (myhost1).

1. On myhost1's MySQL prompt, create a temporary user who can connect from myhost2:

```
mysql> grant all on *.* to 'temp'@'%' identified by 'temp' with grant option;
Query OK, 0 rows affected (0.00 sec)
```

2. On the Cloudera Manager Server host (myhost2), run the script:

```
$ sudo /usr/share/cmf/schema/scm_prepare_database.sh mysql -h
myhost1.sf.cloudera.com -utemp -ptemp --scm-host myhost2.sf.cloudera.com scm scm
scm
Looking for MySQL binary
Looking for schema files in /usr/share/cmf/schema
Verifying that we can write to /etc/cloudera-scm-server
Creating SCM configuration file in /etc/cloudera-scm-server
Executing: /usr/java/jdk1.6.0_31/bin/java -cp
/usr/share/java/mysql-connector-java.jar:/usr/share/cmf/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!
```

3. On myhost1, delete the temporary user:

```
mysql> drop user 'temp'@'%';
Query OK, 0 rows affected (0.00 sec)
```

Example 2: Running the script to configure Oracle

```
[root@rhel55-6 ~]# /usr/share/cmf/schema/scm_prepare_database.sh -h cm-oracle.example.com
 oracle orcl sample_user sample_pass
Verifying that we can write to /etc/cloudera-scm-server
Creating SCM configuration file in /etc/cloudera-scm-server
Executing: /usr/java/jdk1.6.0_31/bin/java -cp
/usr/share/java/mysql-connector-java.jar:/usr/share/cmf/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 when PostgreSQL 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.

```
$ /usr/share/cmf/schema/scm_prepare_database.sh postgresql scm scm scm
```

External Databases for Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server

You can configure Cloudera Manager to use an external database for Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server. If you choose this option, you must create the databases *before* you run the Cloudera Manager installation wizard. For more information, see the instructions in MySQL Database on page 30, Oracle Database on page 35, or External PostgreSQL Database on page 27.

External Databases for Hue, and Oozie

Hue and Oozie are automatically configured with databases, but you can configure these services to use external databases after Cloudera Manager is installed.

Configuring an External Database for Hue

By default Hue is configured to use the SQLite database. If you want to use an external database for Hue, see Using an External Database for Hue.

Configuring an External Database for Oozie

By default Oozie is configured to use the Derby database. If you want to use an external database for Oozie, see Using an External Database for Oozie.

Embedded PostgreSQL Database

Installing and Starting the Embedded PostgreSQL Database

If you are using Installation Path B - Manual Installation Using Cloudera Manager Packages on page 50 for a demonstration or proof of concept deployment, and you want to use an embedded PostgreSQL database for the Cloudera Management Server, use this procedure to install and start the database:

1. Install the embedded PostgreSQL database packages:
   - **Red Hat-compatible, if you have a yum repo configured:**

     ```
     $ sudo yum install cloudera-manager-server-db-2
     ```

   - **Red Hat-compatible, if you're transferring RPMs manually:**

     ```
     $ sudo yum --nogpgcheck localinstall cloudera-manager-server-db-2.noarch.rpm
     ```

   - **SLES:**

     ```
     $ sudo zypper install cloudera-manager-server-db-2
     ```

   - **Debian/Ubuntu**

     ```
     $ sudo apt-get install cloudera-manager-server-db-2
     ```

2. Start the PostgreSQL database:

```
$ sudo service cloudera-scm-server-db start
```

Stopping the Embedded PostgreSQL Database

1. 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. Stop the Cloudera Management Service.
3. Stop the Cloudera Manager Server.
4. Stop the Cloudera Manager Server database:

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

Changing Embedded PostgreSQL Database Passwords

The embedded PostgreSQL database has predefined user accounts and passwords. To change passwords associated with the embedded PostgreSQL database accounts retrieve the user name or password, as well as other database information as follows:

- The Cloudera Manager service connects to the database using the `scm` account. Information about this account is stored in the `db.properties` file.
- The root account for the database is the `cloudera-scm` account. Information about this account is stored in the `generated_password.txt` file.

To find information about the PostgreSQL database user account that the SCM service 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 find information about the root account for the database, read 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.
```

Once you have gathered passwords, you can change the passwords for users, if desired.

External PostgreSQL Database

If you want to use an external PostgreSQL database, follow these procedures.

Installing the External PostgreSQL Server

> **Note:**
> - If you already have a PostgreSQL database set up, you can skip to the section Configuring and Starting the PostgreSQL Server on page 27 to verify that your PostgreSQL configurations meet the requirements for Cloudera Manager.
> - It is important that the data directory, which by default is `/var/lib/postgresql/data/`, is on a partition that has sufficient free space.

1. Use one or more of the following commands to set the locale:

```
export LANGUAGE=en_US.UTF-8
export LANG=en_US.UTF-8
export LC_ALL=en_US.UTF-8
locale-gen en_US.UTF-8
dpkg-reconfigure locales
```

2. Install PostgreSQL packages:

   - **Red Hat**

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

   - **SLES**

```
$ sudo zypper install postgresql91-server
```

> **Note:** This command will install PostgreSQL 9.1. If you want to install a different version, you can use `zypper search postgresql` to search for available versions. You should install version 8.4 or higher.

   - **Debian/Ubuntu**

```
$ sudo apt-get install postgresql
```

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 Name (FQDN) of the hosts hosting the management roles. If you do not make these changes, the management processes will not be able to connect to and use the database on which they depend.

1. Initialize the external PostgreSQL database. For some versions of PostgreSQL, this is done automatically the first time that you start the PostgreSQL server. In this case, issue the command:

```
$ sudo service postgresql start
```

   In other versions, you must explicitly initialize the database using:

```
$ sudo service postgresql initdb
```

See the PostgreSQL documentation for more details.

2. Enable MD5 authentication. Edit `pg_hba.conf`, which is usually found in `/var/lib/pgsql/data` or `/etc/postgresql/8.4/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.

   - Small clusters - For small to mid-sized clusters, consider the following suggestions as a starting point for settings. If resources are especially 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 Cloudera Manager is running on the same host as other roles, the following values may be acceptable:
     - `shared_buffers` - 256MB
     - `wal_buffers` - 8MB
     - `checkpoint_segments` - 16
     - `checkpoint_completion_target` - 0.9
   - Large clusters - may contain up to 1000 hosts. For large clusters consider the following suggestions as a starting point for settings.
     - `max_connection` - For large clusters, each database is typically hosted on a different host. The general rule is to allow each database on a host 100 maximum connections and then add 50 extra connections. As a result, in the normal case for large clusters, configure each of the five hosts that hosts a single database for 150 connections. You may have to increase the system resources available to PostgreSQL, as described at Connection Settings.
     - `shared_buffers` - 1024MB. 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` - 16MB. 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 16MB works well in most case.
     - `checkpoint_segments` - 128. The PostgreSQL Tuning Guide recommends values between 32 and 256 for write-intensive systems, such as this one.
     - `checkpoint_completion_target` - 0.9. This setting is only available in PostgreSQL 8.3 and later. These versions are highly recommended.

4. Configure the PostgreSQL server to start at boot.

   - **Red Hat**

```
$ sudo /sbin/chkconfig postgresql on
$ sudo /sbin/chkconfig --list postgresql
postgresql          0:off   1:off   2:on    3:on    4:on    5:on    6:off
```

- **SLES**

```
$ sudo chkconfig --add postgresql
```

- **Debian/Ubuntu**

```
$ sudo chkconfig postgresql on
```

5. Start or restart the PostgreSQL database:

```
$ sudo service postgresql restart
```

Creating Databases for Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server

Create databases and user accounts for components that require databases:

- If you are not using the Cloudera Manager installer, the Cloudera Manager Server.
- Cloudera Management Service roles:
  - Activity Monitor (if using the MapReduce service)
  - Reports Manager
- Each Hive Metastore
- Sentry Server
- Cloudera Navigator Audit Server

You can create these databases on the host where the Cloudera Manager Server will run, or on any other hosts in the cluster. For performance reasons, you should typically install each database on the host on which the service runs, as determined by the roles you will assign during installation or upgrade. In larger deployments or in cases where database administrators are managing the databases the services will use, databases may be separated from services, but do not undertake such an implementation lightly.

The database must be configured to support UTF-8 character set encoding.

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

1. Connect to PostgreSQL:

```
$ sudo -u postgres psql
```

2. If you are not using the Cloudera Manager installer, create a database for the Cloudera Manager Server. The database name, user name, and password can be anything you want. Be sure to note the names chosen, as you will need to supply them later when running the `scm_prepare_database.sh` script.

```
postgres=# CREATE ROLE scm LOGIN PASSWORD 'scm';
postgres=# CREATE DATABASE scm OWNER scm ENCODING 'UTF8';
```

3. Create databases for Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server:

```
postgres=# CREATE ROLE user LOGIN PASSWORD 'password';
postgres=# CREATE DATABASE databaseName OWNER user ENCODING 'UTF8';
```

where *user*, *password*, and *databaseName* can be anything you want. The examples shown match the default names provided in the Cloudera Manager configuration settings:

Role	Database	User	Password
Activity Monitor	amon	amon	amon_password
Reports Manager	rman	rman	rman_password
Hive Metastore Server	metastore	hive	hive_password
Sentry Server	sentry	sentry	sentry_password
Cloudera Navigator Audit Server	nav	nav	nav_password

For PostgreSQL 8.2.23 or later, also do:

```
postgres=# ALTER DATABASE Metastore SET standard_conforming_strings = off;
```

MySQL Database

If you want to use an MySQL database, follow these procedures.

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 30 to verify that your MySQL configurations meet the requirements for Cloudera Manager.
  - It is important that the `datadir` directory, which, by default, is `/var/lib/mysql`, is on a partition that has sufficient free space.

1. Install the MySQL database.

OS	Command
RHEL	`$ sudo yum install mysql-server`

SLES	`$ sudo zypper install mysql`
	`$ sudo zypper install libmysqlclient_r15`
	- **Note:** Some SLES systems encounter errors when using the preceding `zypper install` command. For more information on resolving this issue, see the Novell Knowledgebase topic, error running chkconfig.

Ubuntu and Debian	`$ sudo apt-get install mysql-server`

After issuing the command to install MySQL, you may need to respond to prompts to confirm that you do want to complete the installation.

Configuring and Starting the MySQL Server

1. Determine the version of MySQL.
2. Stop the MySQL server if it is running.

OS	Command
RHEL	`$ sudo service mysqld stop`

OS	Command
SLES, Ubuntu, and Debian	`$ sudo service mysql stop`

3. 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.
4. Determine the location of the option file, `my.cnf`.
5. Update `my.cnf` so that it conforms to the following requirements:

   - **Important:**
     - To prevent deadlocks, Cloudera Manager requires the isolation level to be set 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 are very conservative with regards to buffer sizes and memory usage. Cloudera Management Service roles need high write throughput as, based on cluster size, they may insert a lot of records in the database. Therefore 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. Clusters with fewer than 50 hosts can be considered small clusters and clusters with more than 50 hosts can be considered large clusters:
       - Small clusters - 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 storage volume.
         - 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, Cloudera Navigator, and Hive Metastore), set the maximum connections to 550.
       - Large clusters - do not store more than one database on the same host. In such a case, use a separate host for each database/host pair. The hosts need not be reserved exclusively for databases, but each database should be on a separate host.

   Here is a typical option file:

   ```
   [mysqld]
   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              = 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

   # log_bin should be on a disk with enough free space
   # NOTE: replace '/x/home/mysql/logs/binary' below with
   ```

```
#       an appropriate path for your system.
log_bin=/x/home/mysql/logs/binary/mysql_binary_log

# For MySQL version 5.1.8 or later. Comment out binlog_format for older versions.
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
```

6. 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.
7. Ensure the MySQL server starts at boot.

OS	Command

RHEL	`$ sudo /sbin/chkconfig mysqld on`
	`$ sudo /sbin/chkconfig --list mysqld`
	`mysqld 0:off 1:off 2:on 3:on 4:on 5:on 6:off`

SLES	`$ sudo chkconfig --add mysql`
Ubuntu and Debian	`$ sudo chkconfig mysql on` ▪ **Note:** `chkconfig` may not be available on recent Ubuntu releases. In such cases, you may need to use Upstart to configure MySQL to start automatically when the system boots. See the Ubuntu documentation or the Upstart Cookbook for more information.

8. Start the MySQL server:

OS	Command
RHEL	`$ sudo service mysqld start`
SLES, Ubuntu, and Debian	`$ sudo service mysql start`

9. Set the MySQL root password. In the following procedure, your current `root` password is blank. Press the **Enter** key when you're prompted for the root password.

```
$ 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 Connector

Install the JDBC connector on the Cloudera Manager Server host, as well as hosts to which you assign the Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server roles.

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

Cloudera recommends that you assign all roles that require databases on the same host and install the connector on that host. While putting all such roles on the same host is recommended, it is not required. You could install a role, such as Activity Monitor on one host and other roles on a separate host. In such a case you would install the JDBC connector on each host running roles that access the database.

OS	Command

RHEL 5 and 6	1. Download the MySQL JDBC connector from http://www.mysql.com/downloads/connector/j/5.1.html.
	2. Extract the JDBC driver JAR file from the downloaded file; for example:
	`tar zxvf mysql-connector-java-5.1.31.tar.gz`
	3. Add the JDBC driver, renamed, to the relevant server; for example:
	`$ sudo cp mysql-connector-java-5.1.31/mysql-connector-java-5.1.31-bin.jar /usr/share/java/mysql-connector-java.jar`
	If the target directory does not yet exist on this host, you can create it before copying the JAR file; for example:
	`$ sudo mkdir -p /usr/share/java/`
	`$ sudo cp mysql-connector-java-5.1.31/mysql-connector-java-5.1.31-bin.jar /usr/share/java/mysql-connector-java.jar`
	Note: Do not use the `yum install` command to install the MySQL connector package, because it installs the openJDK, and then uses Linux alternatives command to set the system JDK to be the openJDK.

SLES	`$ sudo zypper install mysql-connector-java`
Ubuntu or Debian	`$ sudo apt-get install libmysql-java`

Creating Databases for Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server

Create databases and user accounts for components that require databases:

- If you are not using the Cloudera Manager installer, the Cloudera Manager Server.
- Cloudera Management Service roles:

  - Activity Monitor (if using the MapReduce service)
  - Reports Manager
- Each Hive Metastore
- Sentry Server
- Cloudera Navigator Audit Server

You can create these databases on the host where the Cloudera Manager Server will run, or on any other hosts in the cluster. For performance reasons, you should typically install each database on the host on which the service runs, as determined by the roles you will assign during installation or upgrade. In larger deployments or in cases where database administrators are managing the databases the services will use, databases may be separated from services, but do not undertake such an implementation lightly.

The database must be configured to support UTF-8 character set encoding.

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

1. Log into MySQL as the root user:

   ```
   $ mysql -u root -p
   Enter password:
   ```

2. Create databases for the Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server:

   ```
   mysql> create database database DEFAULT CHARACTER SET utf8;
   Query OK, 1 row affected (0.00 sec)

   mysql> grant all on database.* TO 'user'@'%' IDENTIFIED BY 'password';
   Query OK, 0 rows affected (0.00 sec)
   ```

   where *database*, *user*, and *password* can be anything you want. The examples shown match the default names provided in the Cloudera Manager configuration settings:

Role	Database	User	Password
Activity Monitor	amon	amon	amon_password
Reports Manager	rman	rman	rman_password
Hive Metastore Server	metastore	hive	hive_password
Sentry Server	sentry	sentry	sentry_password
Cloudera Navigator Audit Server	nav	nav	nav_password

Backing Up MySQL Databases

To back up the MySQL database, run the `mysqldump` command on the MySQL host, as follows:

```
$ mysqldump -hhostname -uusername -ppassword database > /tmp/database-backup.sql
```

For example, to back up the Activity Monitor database amon created in Creating Databases for Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server on page 33, on the local host as the root user, with the password `amon_password`:

```
$ mysqldump -pamon_password amon > /tmp/amon-backup.sql
```

To back up the sample Activity Monitor database amon on remote host myhost.example.com as the root user, with the password amon_password:

```
$ mysqldump -hmyhost.example.com -uroot -pcloudera amon > /tmp/amon-backup.sql
```

Oracle Database

If you want to use an Oracle database, follow these procedures.

Collecting Oracle Database Information

Installing, configuring, and maintaining an Oracle database should be completed by your organization's database administrator. In preparation for configuring Cloudera Manager to work with Oracle databases, gather the following information from your Oracle DBA:

- Host Name - The DNS name or the IP address of the host where the Oracle database is installed.
- SID - the name of the database that will store Cloudera Manager information. This database could contain schema that would store information for the Cloudera Manager Server, Activity Monitor, Reports Manager, and Cloudera Navigator.
- User name - a user name for each schema that is storing information. This means you might have four unique usernames for the four schema.
- Password - a password corresponding to each user name.

You will use the Oracle database information that you have gathered to configure the external database to work with the Cloudera Manager Server.

Configuring the Oracle Server

Adjust 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 Cloudera Manager configuration database and to smaller clusters. Many factors contribute to whether to reconfigure 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 Hadoop cluster.
- Create more redo log groups and spread the redo log members across separate disks/LUNs.
- Increase the size of redo log members to be at least 1 gigabyte.

Modify 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 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, Cloudera Navigator, 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 two databases, you 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;
```

Ensure Your Oracle Database Supports UTF8

The database you use must be configured to support UTF8 character set encoding. One way your DBA might implement UTF8 character set encoding in Oracle databases is using the dbca utility. In such a case, when creating a database, the `characterSet AL32UTF8` option might be used to specify proper encoding. Consult with your DBA to ensure UTF8 encoding is properly configured.

Having collected information about your Oracle database, installed the Oracle JDBC, considered having database settings adjusted, and ensured UTF-8 encoding is enabled, proceed to Installing Cloudera Manager, CDH, and Managed Services on page 43.

Installing the Oracle JDBC Connector

You must install the JDBC connector on the Cloudera Manager Server host, as well as hosts to which you assign the Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server server roles.

Cloudera recommends that you assign all roles that require databases on the same host and install the connector on that host. While putting all such roles on the same host is recommended, it is not required. You could install a role, such as Activity Monitor on one host and other roles on a separate host. In such a case you would install the JDBC connector on each host running roles that access the database.

1. Download and install the `ojdbc6.jar` file, which contains the JDBC driver. There are different versions of the `ojdbc6.jar` file. You must download the version that is designed for:
   - Java 6
   - The Oracle database version used in your environment For example, for an environment using Oracle 11g R2, the jar file can be downloaded from http://www.oracle.com/technetwork/database/enterprise-edition/jdbc-112010-090769.html.
2. Copy the appropriate JDBC JAR file to `/usr/share/java/oracle-connector-java.jar` for use with the Cloudera Manager databases (for example, for the Activity Monitor, and so on), and for use with Hive.

   ```
   $ mkdir /usr/share/java (if necessary)
   $ cp /tmp/ojdbc6.jar /usr/share/java/oracle-connector-java.jar
   ```

Creating Databases for Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server

Create databases and user accounts for components that require databases:

- If you are not using the Cloudera Manager installer, the Cloudera Manager Server.

- Cloudera Management Service roles:
  - Activity Monitor (if using the MapReduce service)
  - Reports Manager
- Each Hive Metastore
- Sentry Server
- Cloudera Navigator Audit Server

You can create these databases on the host where the Cloudera Manager Server will run, or on any other hosts in the cluster. For performance reasons, you should typically install each database on the host on which the service runs, as determined by the roles you will assign during installation or upgrade. In larger deployments or in cases where database administrators are managing the databases the services will use, databases may be separated from services, but do not undertake such an implementation lightly.

The database must be configured to support UTF-8 character set encoding.

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

Backing up Databases

Cloudera recommends that you periodically back up the databases that Cloudera Manager uses to store configuration, monitoring, and reporting data and for managed services that require a database:

- Cloudera Manager - Contains all the information about what services you have configured, their role assignments, all configuration history, commands, users, and running processes. This is a relatively small database (<100 MB), and is the most important to back up. A monitoring database contains monitoring information about service and host status. In large clusters, this database can grow large.
- Activity Monitor - Contains information about past activities. In large clusters, this database can grow large.
- Reports Manager - Keeps track of disk utilization and processing activities over time. Medium-sized.
- Hive Metastore - Contains Hive metadata. Relatively small.
- Sentry Server - Contains authorization metadata. Relatively small.
- Cloudera Navigator Audit Server - Contains auditing information. In large clusters, this database can grow large.

Backing Up PostgreSQL Databases

The procedure for backing up a PostgreSQL database is the same whether you are using an embedded or external database:

1. Log in to the host where the Cloudera Manager Server is installed.
2. Run the following command as root:

```
cat /etc/cloudera-scm-server/db.properties.
The db.properties file contains:
# Auto-generated by scm_prepare_database.sh
# Mon Jul 27 22:36:36 PDT 2011
com.cloudera.cmf.db.type=postgresql
com.cloudera.cmf.db.host=host:7432
com.cloudera.cmf.db.name=scm
com.cloudera.cmf.db.user=scm
com.cloudera.cmf.db.password=NnYfWIjlbk
```

3. Run the following command as root using the parameters from the preceding step:

```
# pg_dump -h host -p 7432 -U scm > /tmp/scm_server_db_backup.$(date +%Y%m%d)
```

4. Enter the password specified for the `com.cloudera.cmf.db.password` property on the last line of the `db.properties file`. If you are using the embedded database, Cloudera Manager generated the password

for you during installation. If you are using an external database, enter the appropriate information for your database.

Backing Up MySQL Databases

To back up the MySQL database, run the `mysqldump` command on the MySQL host, as follows:

```
$ mysqldump -hhostname -uusername -ppassword database > /tmp/database-backup.sql
```

For example, to back up the Activity Monitor database amon created in Creating Databases for Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server on page 33, on the local host as the root user, with the password `amon_password`:

```
$ mysqldump -pamon_password amon > /tmp/amon-backup.sql
```

To back up the sample Activity Monitor database amon on remote host `myhost.example.com` as the root user, with the password `amon_password`:

```
$ mysqldump -hmyhost.example.com -uroot -pcloudera amon > /tmp/amon-backup.sql
```

Backing Up Oracle Databases

For Oracle, work with your database administrator to ensure databases are properly backed up.

Monitoring Data Storage

The Service Monitor and Host Monitor roles in the Cloudera Management Service store time series data, health data, Impala query metadata, and YARN application metadata. This section describes the process for migrating monitoring data and how to configure disk and memory properties to accommodate the requirements of these roles.

Monitoring Data Migration During Cloudera Manager Upgrade

The Cloudera Manager upgrade process automatically migrates data from existing databases to the local datastore. The upgrade process occurs only once for Host Monitor and Service Monitor, though it can be spread across multiple runs of Host Monitor and Service Monitor if they are restarted before it completes. Resource usage (CPU, memory, and disk) by Host Monitor and Service Monitor will be higher than normal during the process.

You can monitor the progress of migrating data from a Cloudera Manager 4 database to the Cloudera Manager 5 datastore in the Host Monitor and Service Monitor logs. Log statements starting with `LDBTimeSeriesDataMigrationTool` identify the upgrade process. The important statements are: `Starting DB migration` when migration is first started and `Migration progress: {} total, {} migrated, {} errors` as progress is reported. Progress is reported with partition counts, so it'll be something like `3 total, 0 migrated, 0 errors` to start, up to `3 total, 3 migrated, 0 errors` at the end.

After migration completes, the migrated data is summarized in statements such as `Running the LDBTimeSeriesRollupManager at {}, forMigratedData={}` with table names. At this point, the external database will never again be used by Host Monitor and Service Monitor and the database configurations can be removed (connection information, username, password, etc.).

Service Monitor Storage Configuration

The Service Monitor stores time series data and health data, Impala query metadata, and YARN application metadata.

By default, the data is stored in `/var/lib/cloudera-service-monitor/` on the Service Monitor host. This can be changed by modifying the **Service Monitor Storage Directory** configuration (`firehose.storage.base.directory`). To change this configuration on an active system, see Moving Monitoring Data on an Active Cluster on page 40.

You can also control how much disk space to reserve for the different classes of data the Service Monitor stores by changing the following configuration options:

- Time-series metrics and health data: Time-Series Storage (`firehose_time_series_storage_bytes` - 10 GB default)
- Impala query metadata: Impala Storage (`firehose_impala_storage_bytes` - 1 GB default)
- YARN application metadata: YARN Storage (`firehose_yarn_storage_bytes` - 1 GB default)

See Data Granularity and Time-Series Metric Data on page 39 for an explanation of how metric data is stored within Cloudera Manager and for the impact the storage limits have on data retention.

The default values are fairly small, so you should examine disk usage after several days of activity to determine how much space is needed. Do this by visiting the **Disk Usage** tab on the Service Monitor page. This page shows the current disk space consumed and its rate of growth, both broken down by the type of data stored. For example, it allows you to compare the space consumed by raw metric data versus daily summaries of that data.

Host Monitor Storage Configuration

The Host Monitor stores time series data and health data.

By default, the data is stored in `/var/lib/cloudera-host-monitor/` on the Host Monitor's host. This can be changed by modifying the **Host Monitor Storage Directory** configuration (`firehose.storage.base.directory`). To change this configuration on an active system see Moving Monitoring Data on an Active Cluster on page 40.

You can control how much disk space to reserve for Host Monitor data by changing the following configuration option:

- Time-series metrics and health data: Time Series Storage (`firehose_time_series_storage_bytes` - 10 GB default)

See the next section for an explanation of how metric data is stored within Cloudera Manager and for the impact these limits have on data retention.

The default value is fairly small so we encourage you to examine disk usage after several days of activity to determine how much space they need. You can do this by visiting the **Disk Usage** tab on the Host Monitor page. This page shows the current disk space consumed and its rate of growth, both broken down by the type of data stored. For example, it allows you to compare the space consumed by raw metric data versus daily summaries of that data.

Data Granularity and Time-Series Metric Data

The Service Monitor and Host Monitor store metric data store time-series metric data in a variety of ways. When the data is first received it is written as is to the metric store. Over time, the raw data is summarized to and stored at various data granularities. For example, after ten minutes a single ten-minute summary point is written containing the average of the metric over the period as well as the minimum, the maximum, the standard deviation, and a variety of other statistics. This process is repeated to produce hourly, six-hourly, daily, and weekly summaries. This data summarization system is only for metric data. Impala query monitoring and YARN application monitoring do not have a similar system. For those systems, when the storage limit is reached, the oldest stored records are deleted.

The Service Monitor and Host Monitor internally manage the amount of their overall storage space to dedicate to each data granularity level. When the limit for a particular level is reached, the oldest data points at that level are deleted. Note that metric data for that time period remains available at the lower granularity levels. That is, when an hourly point for a particular time is deleted to free up space, a daily point still exists covering that hour. Since each of these data granularities consumes significantly less storage than the previous summary level,

lower granularity levels can be retained for longer periods of time. In particular, given a reasonable amount of storage, weekly points can normally be retained indefinitely.

Some features, notably detailed display of health results, depend on the presence of raw data. Health history is maintained by the event store dictated by its retention policies.

Moving Monitoring Data on an Active Cluster

There are two ways to change where monitoring data is stored on a cluster: basic and advanced.

Basic: Changing the Configured Directory

1. Stop the Service or Host Monitor.
2. If you want to save your old monitoring data then copy the current directory to the new directory.
3. Update the **Storage Directory** configuration option (firehose.storage.base.directory) on corresponding role's configuration page.
4. Start the Service or Host Monitor.

Advanced: High Performance

For the best performance, and especially for a large cluster, we recommend putting the Host and Service Monitor storage directories on their own dedicated spindles. In most cases that will provide sufficient performance, but if you need additional performance you can divide your data even further. Though this cannot be configured directly with Cloudera Manager, it can be done using symbolic links.

For example if all your Service Monitor data is located in `/data/1/service_monitor` and you want to separate your Impala data from your time series data you could do the following:

1. Stop the Service Monitor.
2. Move the original Impala data in `/data/1/service_monitor/impala` to the new directory, for example `/data/2/impala_data`.
3. Create a symbolic link from `/data/1/service_monitor/impala` to `/data/2/impala_data` with the following command:

```
ln -s /data/2/impala_data /data/1/service_monitor/impala
```

4. Start the Service Monitor.

Host Monitor and Service Monitor Memory Configuration

There are two memory-related configuration options: Java heap size and non-Java memory size. The memory required or recommended for both of these configuration options depends on the size of the cluster. In addition to the memory configured, the Host and Service Monitor will also take advantage of the Linux page cache. Having memory free for use as page cache on the Service and Host Monitor hosts will improve performance.

Table 3: Small Clusters: No More Than 10 Hosts

	Required	Recommended
Java Heap Size	256 MB	512 MB
Non-Java Memory	768 MB	1.5 GB

Table 4: Medium Clusters: Between 11 and 100 Hosts

	Required	Recommended
Java Heap Size	1 GB	2 GB

	Required	Recommended
Non-Java Memory	2 GB	4 GB

Table 5: Large Clusters: More Than 100 Hosts

	Required	Recommended
Java Heap Size	2 GB	4 GB
Non-Java Memory	6 GB	12 GB

Installing Cloudera Manager, CDH, and Managed Services

A Cloudera Manager deployment consists of many software components: Cloudera Manager Server and Agent software, supporting database software, and CDH and managed service software. This section describes the three main paths for creating a new Cloudera Manager deployment and the criteria for choosing an installation path. If your cluster already has an installation of a previous version of Cloudera Manager, follow the instructions in Upgrading Cloudera Manager.

Choosing an Installation Path

The Cloudera Manager installation paths share some common phases, but the variant aspects of each path support different user and cluster host requirements:

- **Demonstration and proof of concept deployments** - There are two installation options:
  - Installation Path A - Automated Installation by Cloudera Manager on page 44 - Cloudera Manager automates the installation of the Oracle JDK, Cloudera Manager Server, embedded PostgreSQL database, and Cloudera Manager Agent packages, and configures databases for the Cloudera Manager Server and Hive Metastore and optionally for Cloudera Management Service roles. This path is recommended for demonstration and proof of concept deployments, but is *not recommended* for production deployments because its not intended to scale and may require database migration as your cluster grows. To use this method, server and cluster hosts must satisfy the following requirements:
    - Provide the ability to log in to the Cloudera Manager Server host using a root account or an account that has password-less sudo permission.
    - Allow the Cloudera Manager Server host to have uniform SSH access on the same port to all hosts. See Networking and Security Requirements on page 17 for further information.
    - All hosts must have access to standard package repositories and either `archive.cloudera.com` or a local repository with the necessary installation files.
  - Installation Path B - Manual Installation Using Cloudera Manager Packages on page 50 - you install the Oracle JDK and Cloudera Manager Server, and embedded PostgreSQL database packages on the Cloudera Manager Server host. You have two options for installing Oracle JDK, Cloudera Manager Agent, CDH, and managed service software on cluster hosts: manually install it yourself or use Cloudera Manager to automate installation. However, in order for Cloudera Manager to automate installation of Cloudera Manager Agent packages or CDH and managed service packages, cluster hosts must satisfy the following requirements:
    - Allow the Cloudera Manager Server host to have uniform SSH access on the same port to all hosts. See Networking and Security Requirements on page 17 for further information.
    - All hosts must have access to standard package repositories and either `archive.cloudera.com` or a local repository with the necessary installation files.
- **Production deployments** - require you to first manually install and configure a production database for the Cloudera Manager Server and Hive Metastore. There are two installation options:
  - Installation Path B - Manual Installation Using Cloudera Manager Packages on page 50 - you install the Oracle JDK and Cloudera Manager Server packages on the Cloudera Manager Server host. You have two options for installing Oracle JDK, Cloudera Manager Agent, CDH, and managed service software on cluster hosts: manually install it yourself or use Cloudera Manager to automate installation. However, in order for Cloudera Manager to automate installation of Cloudera Manager Agent packages or CDH and managed service packages, cluster hosts must satisfy the following requirements:
    - Allow the Cloudera Manager Server host to have uniform SSH access on the same port to all hosts. See Networking and Security Requirements on page 17 for further information.
    - All hosts must have access to standard package repositories and either `archive.cloudera.com` or a local repository with the necessary installation files.

– Installation Path C - Manual Installation Using Cloudera Manager Tarballs on page 65 - you install the Oracle JDK, Cloudera Manager Server, and Cloudera Manager Agent software as tarballs and use Cloudera Manager to automate installation of CDH and managed service software as parcels.

Installation Path A - Automated Installation by Cloudera Manager

Before proceeding with this path for a new installation, review Choosing an Installation Path on page 43. If you are upgrading an Cloudera Manager existing installation, see Upgrading Cloudera Manager.

The general steps in the procedure for Installation Path A follow.

Before You Begin

In certain circumstances you may need to perform optional installation and configuration steps.

Install and Configure External Databases

If you intend to use an external database for services or Cloudera Management Service roles, install and configure it following the instructions in External Databases for Activity Monitor, Reports Manager, Hive Metastore, Sentry Server, and Cloudera Navigator Audit Server on page 25.

(CDH 5 only) On RHEL and CentOS 5, Install Python 2.6 or 2.7

Python 2.6 or 2.7 is required to run Hue. RHEL 5 and CentOS 5, in particular, require the EPEL repository package.

In order to install packages from the EPEL repository, first download the appropriate repository rpm packages to your machine and then install Python using `yum`. For example, use the following commands for RHEL 5 or CentOS 5:

```
$ su -c 'rpm -Uvh
http://download.fedoraproject.org/pub/epel/5/i386/epel-release-5-4.noarch.rpm'
...
$ yum install python26
```

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 Debian/Ubuntu systems. If your hosts access the Internet through an HTTP proxy, you can configure `yum`, `zypper`, or `apt-get`, system-wide, to access archive.cloudera.com through a proxy. To do so, modify the system configuration on the Cloudera Manager Server host and on every cluster host as follows:

OS	File	Property
RHEL-compatible	`/etc/yum.conf`	`proxy=http://server:port/`
SLES	`/root/.curlrc`	`--proxy=http://server:port/`
Ubuntu or Debian	`/etc/apt/apt.conf`	`Acquire::http::Proxy "http://server:port";`

Download and Run the Cloudera Manager Server Installer

1. Download the Cloudera Manager installer binary from Cloudera Manager 5.1.3 Downloads to the cluster host where you want to install the Cloudera Manager Server.
   a. Click **Download Cloudera Express** or **Download Cloudera Enterprise**. See Cloudera Express and Cloudera Enterprise Features.
   b. Optionally register and click **Submit** or click the Just take me to the **download page** link. The `cloudera-manager-installer.bin` file downloads.

2. Change `cloudera-manager-installer.bin` to have executable permission.

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

3. Run the Cloudera Manager Server installer:
   - Install Cloudera Manager packages from the Internet - `sudo ./cloudera-manager-installer.bin`
   - Install Cloudera Manager packages from a local repository - `sudo ./cloudera-manager-installer.bin --skip_repo_package=1`
4. Read the Cloudera Manager README and then press **Return** or **Enter** to choose **Next**.
5. Read the Cloudera Manager Express License and then press **Return** or **Enter** to choose **Next**. Use the arrow keys and press **Return** or **Enter** to choose **Yes** to confirm you accept the license.
6. Read the Oracle Binary Code License Agreement and then press **Return** or **Enter** to choose **Next**.
7. Use the arrow keys and press **Return** or **Enter** to choose **Yes** to confirm you accept the Oracle Binary Code License Agreement. The following occurs:
   a. The installer installs the Oracle JDK and the Cloudera Manager repository files.
   b. The installer installs the Cloudera Manager Server and embedded PostgreSQL packages.
   c. The installer starts the Cloudera Manager Server and embedded PostgreSQL database.
8. When the installation completes, the complete URL provided for the Cloudera Manager Admin Console, including the port number, which is 7180 by default. Press **Return** or **Enter** to choose **OK** to continue.
9. Press **Return** or **Enter** to choose **OK** to exit the installer.

> **Note:** If the installation is interrupted for some reason, you may need to clean up before you can re-run it. See Uninstalling Cloudera Manager and Managed Software on page 125.

Start the Cloudera Manager Admin Console

The Cloudera Manager Server URL takes the following form `http://Server host:port`, where *Server host* is the fully-qualified domain name or IP address of the host where the Cloudera Manager Server is installed and *port* is the port configured for the Cloudera Manager Server. The default port is 7180.

1. Wait several minutes for the Cloudera Manager Server to complete its startup. To observe the startup process you can perform `tail -f /var/log/cloudera-scm-server/cloudera-scm-server.log` on the Cloudera Manager Server host. If the Cloudera Manager Server does not start, see Troubleshooting Installation and Upgrade Problems on page 131.
2. In a web browser, enter `http://Server host:7180`, where *Server host* is the fully-qualified domain name or IP address of the host where you installed the Cloudera Manager Server. The login screen for Cloudera Manager Admin Console displays.
3. Log into Cloudera Manager Admin Console. The default credentials are: **Username:** `admin` **Password:** `admin`. 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. While 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.

Use the Cloudera Manager Wizard for Software Installation and Configuration

The following instructions describe how to use the Cloudera Manager installation wizard to do an initial installation and configuration. The wizard lets you:

- Select the version of Cloudera Manager you want to install
- Find the cluster hosts you specify via hostname and IP address ranges
- Connect to each host with SSH to install the Cloudera Manager Agent and other components

- Optionally (Cloudera Manager 5.1.3) installs the Oracle JDK on the cluster hosts. If you choose not to have the JDK installed, you must install it on all clusters according to the following instructions prior to running the wizard:
  - CDH 5 - (CDH 5) Java Development Kit Installation.
  - CDH 4 - (CDH 4) Java Development Kit Installation.
- Install CDH and managed service packages or parcels
- Configure CDH and managed services automatically and start the services

> - **Important:** All hosts in the cluster must have some way to access installation files via one of the following methods:
>   - Internet access to allow the wizard to install software packages or parcels from `archive.cloudera.com`.
>   - A custom internal repository that the host(s) can access. For example, for a Red Hat host, you could set up a Yum repository. See Creating and Using a Package Repository on page 98 for more information about this option.

Choose Cloudera Manager Edition and Hosts

1. Choose which edition to install:
   - Cloudera Express, which does not require a license, but provides a somewhat limited set of features.
   - Cloudera Enterprise Data Hub Edition Trial, which does not require a license, but expires after 60 days and cannot be renewed
   - Cloudera Enterprise with one of the following license types:
     - Basic Edition
     - Flex Edition
     - Data Hub Edition

   If you choose Cloudera Express or Cloudera Enterprise Data Hub Edition Trial, you can elect to upgrade the license at a later time. See Managing Licenses.
2. If you have elected Cloudera Enterprise, install a license:
   a. Click **Upload License**.
   b. Click the document icon to the left of the **Select a License File** text field.
   c. Navigate to the location of your license file, click the file, and click **Open**.
   d. Click **Upload**.

   Click **Continue** to proceed with the installation.
3. Information is displayed indicating what the CDH installation includes. At this point, you can access online Help or the Support Portal if you wish. Click **Continue** to proceed with the installation.
4. To enable Cloudera Manager to automatically discover hosts on which to install CDH and managed services, enter the cluster hostnames or IP addresses. You can also specify hostname and IP address ranges. For example:

Range Definition	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].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

You can specify multiple addresses and address ranges by separating them by 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. The scan results will include all addresses scanned, but only scans that reach hosts running SSH will be selected for inclusion in your cluster by default. If you don't know the IP addresses of all of the hosts, you can enter an address range that spans over unused addresses and then deselect the hosts that do not exist (and are not discovered) later in this procedure. However, keep in mind that wider ranges will require more time to scan.

5. Click **Search**. Cloudera Manager identifies the hosts on your cluster to allow you to configure them for services. 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**. To find additional hosts, click **New Search**, add the host names or IP addresses and click **Search** again. 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. Common causes of loss of connectivity are firewalls and interference from SELinux.
6. Verify that the number of hosts shown matches the number of hosts where you want to install services. Deselect host entries that do not exist and deselect the hosts where you do not want to install services. Click **Continue**. The Select Repository page displays.

Choose Software Installation Method and Install Software

1. Select the repository type to use for the installation: parcels or packages.
   - **Use Parcels**:
     1. Choose the parcels to install. The choices you see depend on the repositories you have chosen – a repository may contain multiple parcels. Only the parcels for the latest supported service versions are configured by default.

        You can add additional parcels for previous versions by specifying custom repositories. For example, you can find the locations of the previous CDH 4 parcels at `http://archive.cloudera.com/cdh4/parcels/`. Or, if you are installing CDH 4.3 and want to use Sentry for Policy File-Based Hive Authorization, you can add the Sentry parcel using this mechanism.

        1. To specify the parcel directory, local parcel repository, add a parcel repository, or specify the properties of a proxy server through which parcels are downloaded, click the **More Options** button and do one or more of the following:
           - **Parcel Directory** and **Local Parcel Repository Path** - Specify the location of parcels on cluster hosts and the Cloudera Manager Server host.
           - **Parcel Repository** - In the **Remote Parcel Repository URLs** field, click the + button and enter the URL of the repository. The URL you specify is added to the list of repositories listed in the Configuring Server Parcel Settings on page 84 page and a parcel is added to the list of parcels on the Select Repository page. If you have multiple repositories configured, you will see all the unique parcels contained in all your repositories.
           - **Proxy Server** - Specify the properties of a proxy server.
        2. Click **OK**.

   - **Use Packages**:
     1. Select the major release of CDH to install.
     2. Select the specific release of CDH to install.
     3. Select the specific releases of Impala and Solr to install, assuming you have selected an appropriate CDH version. You can choose either the latest version or use a custom repository. Choose **None** if you do not want to install that service.

2. Select the release of Cloudera Manager Agent to install. You can choose either the version that matches the Cloudera Manager Server you are currently using or specify a version in a custom repository.

3. If you opted to use custom repositories for installation files, you can provide a GPG key URL that applies for all repositories.
4. Click **Continue**.
   - (Cloudera Manager 5.1.3) Leave **Install Oracle Java SE Development Kit (JDK)** checked to allow Cloudera Manager to install the JDK on each cluster host or uncheck if you plan to install it yourself.
   - If your local laws permit you to deploy unlimited strength encryption and you are running a secure cluster, check the **Install Java Unlimited Strength Encryption Policy Files** checkbox.

   Click **Continue**.
5. Specify SSH login properties:
   a. Select **root** or enter the user name for an account that has password-less sudo permission.
   b. Select an authentication method:
      - If you choose to use password authentication, enter and confirm the password.
      - If you choose to use public-key authentication provide a passphrase and path to the required key files.
   c. You can choose to specify an alternate SSH port. The default value is 22.
   d. You can specify the maximum number of host installations to run at once. The default value is 10.

   Click **Continue**. Cloudera Manager performs the following:

   - **Parcels** - installs the Oracle JDK and the Cloudera Manager Agent packages and starts the Agent. Click **Continue**. During the parcel installation, progress is indicated for the two phases of the parcel installation process (Download and Distribution) in a separate progress bars. If you are installing multiple parcels you will see progress bars for each parcel. When the **Continue** button appears at the bottom of the screen, the installation process is completed.
   - **Packages** - configures package repositories, installs the Oracle JDK, CDH and managed service and the Cloudera Manager Agent packages, and starts the Agent. When the **Continue** button appears at the bottom of the screen, the installation process is completed. If the installation has completed successfully on some hosts but failed on others, you can click **Continue** if you want to skip installation on the failed hosts and continue to the next screen to start configuring services on the successful hosts.

   While packages are being installed, the status of installation on each host is displayed. You can click the **Details** link for individual hosts to view detailed information about the installation and error messages if installation fails on any hosts. If you click the **Abort Installation** button while installation is in progress, it will halt any pending or in-progress installations and roll back any in-progress installations to a clean state. The **Abort Installation** button does not affect host installations that have already completed successfully or already failed.
6. Click **Continue**. The Host Inspector runs to validate the installation, and provides a summary of what it finds, including all the versions of the installed components. If the validation is successful, click **Finish**. The Cluster Setup page displays.

Add Services

1. In the first page of the Add Services wizard you choose the combination of services to install and whether to install Cloudera Navigator:
   - Click the radio button next to the combination of services to install:

CDH 4	CDH 5
▪ **Core Hadoop** - HDFS, MapReduce, ZooKeeper, Oozie, Hive, and Hue ▪ **Core with HBase** ▪ **Core with Impala** ▪ **All Services** - HDFS, MapReduce, ZooKeeper, HBase, Impala, Oozie, Hive, Hue, and Sqoop	▪ **Core Hadoop** - HDFS, YARN (includes MapReduce 2), ZooKeeper, Oozie, Hive, Hue, and Sqoop ▪ **Core with HBase** ▪ **Core with Impala** ▪ **Core with Search** ▪ **Core with Spark**

CDH 4	CDH 5
▪ **Custom Services** - Any combination of services.	▪ **All Services** - HDFS, YARN (includes MapReduce 2), ZooKeeper, Oozie, Hive, Hue, Sqoop, HBase, Impala, Solr, Spark, and Key-Value Store Indexer ▪ **Custom Services** - Any combination of services.

As you select the services, keep the following in mind:

– Some services depend on other services; for example, HBase requires HDFS and ZooKeeper. Cloudera Manager tracks dependencies and installs the correct combination of services.
– In a CDH 4 cluster, the MapReduce service is the default MapReduce computation framework. Choose **Custom Services** to install YARN or use the Add Service functionality to add YARN after installation completes.

> ▪ **Important:** You can create a YARN service in a CDH 4 cluster, but it is not considered production ready.

– In a CDH 5 cluster, the YARN service is the default MapReduce computation framework. Choose **Custom Services** to install MapReduce or use the Add Service functionality to add MapReduce after installation completes.

> ▪ **Important:** In CDH 5 the MapReduce service has been deprecated. However, the MapReduce service is fully supported for backward compatibility through the CDH 5 life cycle.

– The Flume service can be added only after your cluster has been set up.

▪ If you have chosen Data Hub Edition Trial or Cloudera Enterprise, optionally check the **Include Cloudera Navigator** checkbox to enable Cloudera Navigator. See the Cloudera Navigator Documentation.

Click **Continue**. The Customize Role Assignments page displays.

2. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign role instances to hosts of your choosing, if desired.

   Click a field below a role to display a dialog containing a pageable list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

   The following shortcuts for specifying hostname patterns are supported:

   ▪ Range of hostnames (without the domain portion)

Range Definition	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].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

   ▪ IP addresses
   ▪ Rack name

   Click the **View By Host** button for an overview of the role assignment by hostname ranges.

3. When you are satisfied with the assignments, click **Continue**. The Database Setup page displays.
4. Configure database settings:

a. Choose the database type:
   - Leave the default setting of **Use Embedded Database** to have Cloudera Manager create and configure required databases. Make a note of the auto-generated passwords.
   - Select **Use Custom Databases** to specify external databases.
     1. Enter the database host, database type, database name, username, and password for the database that you created when you set up the database.

b. Click **Test Connection** to confirm that Cloudera Manager can communicate with the database using the information you have supplied. If the test succeeds in all cases, click **Continue**; otherwise check and correct the information you have provided for the database and then try the test again. (For some servers, if you are using the embedded database, you will see a message saying the database will be created at a later step in the installation process.) The Review Changes page displays.

5. Review the configuration changes to be applied. Confirm the settings entered for file system paths. The file paths required vary based on the services to be installed.

   - **Warning:** DataNode data directories should not be placed on NAS devices.

   Click **Continue**. The wizard starts the services.
6. When all of the services are started, click **Continue**. You will see a success message indicating that your cluster has been successfully started.
7. Click **Finish** to proceed to the Home Page.

Configure Cluster CDH Version for Package Installs

If you have installed CDH as a package, after an install or upgrade make sure that the cluster CDH version matches the package CDH version, using the procedure in Configuring the CDH Version for a Cluster in *Managing Clusters with Cloudera Manager*. If the cluster CDH version does not match the package CDH version, Cloudera Manager will incorrectly enable and disable service features based on the cluster's configured CDH version.

Change the Default Administrator Password

As soon as possible after running the wizard and beginning to use Cloudera Manager, change the default administrator password:

1. Right-click the logged-in username at the far right of the top navigation bar and select **Change Password**.
2. Enter the current password, and a new password twice and then click **Update**.

Test the Installation

You can test the installation following the instructions in Testing the Installation on page 123.

Installation Path B - Manual Installation Using Cloudera Manager Packages

Before proceeding with this path for a new installation, review Choosing an Installation Path on page 43. If you are upgrading an Cloudera Manager existing installation, see Upgrading Cloudera Manager.

To install the Cloudera Manager Server using packages, follow the instructions in this section. You can also use Puppet or Chef to install the packages. The general steps in the procedure for Installation Path B follow.

Before You Begin

Install and Configure Databases

Cloudera Manager Server, Cloudera Management Service, and the Hive Metastore data is stored in a database. Install and configure required databases following the instructions in Cloudera Manager and Managed Service Databases on page 21.

(CDH 5 only) On RHEL and CentOS 5, Install Python 2.6 or 2.7

Python 2.6 or 2.7 is required to run Hue. RHEL 5 and CentOS 5, in particular, require the EPEL repository package.

In order to install packages from the EPEL repository, first download the appropriate repository rpm packages to your machine and then install Python using `yum`. For example, use the following commands for RHEL 5 or CentOS 5:

```
$ su -c 'rpm -Uvh
http://download.fedoraproject.org/pub/epel/5/i386/epel-release-5-4.noarch.rpm'
...
$ yum install python26
```

Establish Your Cloudera Manager Repository Strategy

Cloudera recommends installing products using package management tools such as `yum` for Red Hat compatible systems, `zypper` for SLES, and `apt-get` for Debian/Ubuntu. These tools depend on access to repositories to install software. For example, Cloudera maintains Internet-accessible repositories for CDH and Cloudera Manager installation files. Strategies for installing Cloudera Manager include:

- Standard Cloudera repositories. For this method, ensure you have added the required repository information to your systems. For Cloudera Manager repository locations and client repository files, see Cloudera Manager Version and Download Information.
- Internally hosted repositories. You might use internal repositories for environments where hosts do not have access to the Internet. In such a case, ensure your environment is properly prepared. For more information, see Understanding Custom Installation Solutions on page 95.

Red Hat-compatible

1. Save the appropriate Cloudera Manager repo file (`cloudera-manager.repo`) for your system:

OS Version	Repo URL
Red Hat/CentOS/Oracle 5	http://archive.cloudera.com/cm5/redhat/5/x86_64/cm/cloudera-manager.repo
Red Hat/CentOS 6	http://archive.cloudera.com/cm5/redhat/6/x86_64/cm/cloudera-manager.repo

2. Copy the repo file to the `/etc/yum.repos.d/` directory.

SLES

1. Run the following command:

```
$ sudo zypper addrepo -f
http://archive.cloudera.com/cm5/sles/11/x86_64/cm/cloudera-manager.repo
```

2. Update your system package index by running:

```
$ sudo zypper refresh
```

Ubuntu or Debian

1. Save the appropriate Cloudera Manager list file (`cloudera.list`) for your system:

OS Version	Repo URL
Ubuntu Precise (12.04)	http://archive.cloudera.com/cm5/ubuntu/precise/amd64/cm/cloudera.list
Ubuntu Lucid (10.04)	http://archive.cloudera.com/cm5/ubuntu/lucid/amd64/cm/cloudera.list
Debian Wheezy (7.0 and 7.1)	http://archive.cloudera.com/cm5/debian/wheezy/amd64/cm/cloudera.list
Debian Wheezy (6.0)	http://archive.cloudera.com/cm5/debian/squeeze/amd64/cm/cloudera.list

2. Copy the content of that file and append it to the content of the `cloudera.list` in the `/etc/apt/sources.list.d/` directory.
3. Update your system package index by running:

```
$ sudo apt-get update
```

Install the Oracle JDK

Install the Oracle Java Development Kit (JDK) on the Cloudera Manager Server host.

The JDK is included in the Cloudera Manager 5 repositories. Once you have the repo or list file in the correct place, you can install the JDK as follows:

OS	Command
RHEL	`$ sudo yum install oracle-j2sdk1.7`
SLES	`$ sudo zypper install oracle-j2sdk1.7`
Ubuntu or Debian	`$ sudo apt-get install oracle-j2sdk1.7`

Install the Cloudera Manager Server Packages

Install the Cloudera Manager Server packages either on the host where the database is installed, or on a host that has access to the database. This host need not be a host in the cluster that you want to manage with Cloudera Manager. On the Cloudera Manager Server host, type the following commands to install the Cloudera Manager packages.

OS	Command
RHEL, if you have a yum repo configured	`$ sudo yum install cloudera-manager-daemons cloudera-manager-server`
RHEL,if you're manually transferring RPMs	`$ sudo yum --nogpgcheck localinstall cloudera-manager-daemons-*.rpm` `$ sudo yum --nogpgcheck localinstall cloudera-manager-server-*.rpm`
SLES	`$ sudo zypper install cloudera-manager-daemons cloudera-manager-server`
Ubuntu or Debian	`$ sudo apt-get install cloudera-manager-daemons cloudera-manager-server`

Set up a Database for the Cloudera Manager Server

Set up the Cloudera Manager Server database as described in Setting up the Cloudera Manager Server Database on page 22.

(Optional) Install Cloudera Manager Agent, CDH, and Managed Service Software

You can have Cloudera Manager install Cloudera Manager Agent packages or manually install the packages yourself. Similarly, you can allow Cloudera Manager to install CDH and managed service software or manually install the software yourself.

If you choose to have Cloudera Manager install the software (in Choose Software Installation Method and Install Software on page 61), you must satisfy the requirements described in Choosing an Installation Path on page 43. If you satisfy the requirements and choose to have Cloudera Manager install software, you can go to Start the Cloudera Manager Server on page 59. Otherwise, proceed with the following sections.

Install the Oracle JDK

Install the Oracle JDK on the cluster hosts. Cloudera Manager 5 can manage both CDH 5 and CDH 4, and the required JDK version varies accordingly:

- CDH 5 - (CDH 5) Java Development Kit Installation.
- CDH 4 - (CDH 4) Java Development Kit Installation.

Install Cloudera Manager Agent Packages

If you to manually install the packages yourself, on every Cloudera Manager Agent host (including those that will run one or more of the Cloudera Management Service roles: Service Monitor, Activity Monitor, Event Server, Alert Publisher, Reports Manager) do the following:

1. Use one of the following commands to install the Cloudera Manager Agent packages:

OS	Command
RHEL, if you have a yum repo configured:	`$ sudo yum install cloudera-manager-agent cloudera-manager-daemons`
RHEL, if you're manually transferring RPMs:	`$ sudo yum --nogpgcheck localinstall cloudera-manager-agent-package.*.x86_64.rpm cloudera-manager-daemons`
SLES	`$ sudo zypper install cloudera-manager-agent cloudera-manager-daemons`
Ubuntu or Debian	`$ sudo apt-get install cloudera-manager-agent cloudera-manager-daemons`

2. On every Cloudera Manager Agent 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
`server_host`	Name of host where the Cloudera Manager Server is running.
`server_port`	Port on host where the Cloudera Manager Server is running.

   For more information on Agent configuration options, see Agent Configuration File.

Install CDH and Managed Service Packages

For more information about manually installing CDH packages, see CDH 4 Installation Guide or CDH 5 Installation Guide.

1. Choose a repository strategy:
   - Standard Cloudera repositories. For this method, ensure you have added the required repository information to your systems.

- Internally hosted repositories. You might use internal repositories for environments where hosts do not have access to the Internet. In such a case, ensure your environment is properly prepared. For more information, see Understanding Custom Installation Solutions on page 95.

2. Install packages:

<table>
<tr><th>CDH Version</th><th>Procedure</th></tr>
<tr><td>CDH 5</td><td>

- **Red Hat**

1. Download and install the "1-click Install" package

a. Download the CDH 5 "1-click Install" package.

Click the entry in the table below that matches your Red Hat or CentOS system, choose **Save File**, and save the file to a directory to which you have write access (it can be your home directory).

<table>
<tr><th>OS Version</th><th>Click this Link</th></tr>
<tr><td>Red Hat/CentOS/Oracle 5</td><td>Red Hat/CentOS/Oracle 5 link</td></tr>
<tr><td>Red Hat/CentOS/Oracle 6</td><td>Red Hat/CentOS/Oracle 6 link</td></tr>
</table>

b. Install the RPM:

- **Red Hat/CentOS/Oracle 5**

```
$ sudo yum --nogpgcheck localinstall cloudera-cdh-5-0.x86_64.rpm
```

- **Red Hat/CentOS/Oracle 6**

```
$ sudo yum --nogpgcheck localinstall cloudera-cdh-5-0.x86_64.rpm
```

2. (Optionally) add a repository key:

- **Red Hat/CentOS/Oracle 5**

```
$ sudo rpm --import
http://archive.cloudera.com/cdh5/redhat/5/x86_64/cdh/RPM-GPG-KEY-cloudera
```

- **Red Hat/CentOS/Oracle 6**

```
$ sudo rpm --import
http://archive.cloudera.com/cdh5/redhat/6/x86_64/cdh/RPM-GPG-KEY-cloudera
```

3. Install the CDH packages:

```
$ sudo yum clean all
$ sudo yum install avro-tools crunch flume-ng hadoop-hdfs-fuse
hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat
hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms
hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell
kite llama mahout oozie pig pig-udf-datafu search sentry
solr-mapreduce spark-python sqoop sqoop2 whirr
```

</td></tr>
</table>

<table>
<tr><th>CDH Version</th><th>Procedure</th></tr>
<tr><td></td><td>

<blockquote>
Note: Installing these packages will also install all the other CDH packages that are needed for a full CDH 5 installation.
</blockquote>

- **SLES**

1. Download and install the "1-click Install" package.

   a. Download the CDH 5 "1-click Install" package.

   Click this link, choose **Save File**, and save it to a directory to which you have write access (it can be your home directory).

   b. Install the RPM:

```
$ sudo rpm -i cloudera-cdh-5-0.x86_64.rpm
```

   c. Update your system package index by running:

```
$ sudo zypper refresh
```

2. (Optionally) add a repository key:

```
$ sudo rpm --import
http://archive.cloudera.com/cdh5/sles/11/x86_64/cdh/RPM-GPG-KEY-cloudera
```

3. Install the CDH packages:

```
$ sudo zypper clean --all
$ sudo zypper install avro-tools crunch flume-ng hadoop-hdfs-fuse
hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat
hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms
hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell
kite llama mahout oozie pig pig-udf-datafu search sentry
solr-mapreduce spark-python sqoop sqoop2 whirr
```

<blockquote>
Note: Installing these packages will also install all the other CDH packages that are needed for a full CDH 5 installation.
</blockquote>

- **Ubuntu and Debian**

1. Download and install the "1-click Install" package

   a. Download the CDH 5 "1-click Install" package:

OS Version	Click this Link
Wheezy	Wheezy link
Precise	Precise link

   b. Install the package. Do one of the following:

   - Choose **Open with** in the download window to use the package manager.
   - Choose **Save File**, save the package to a directory to which you have write access (it can be your home directory) and install it from the command line, for example:

```
sudo dpkg -i cdh5-repository_1.0_all.deb
```

</td></tr>
</table>

<table>
<tr><th>CDH Version</th><th>Procedure</th></tr>
<tr><td></td><td>
2. (Optionally) add a repository key:

Debian Wheezy
<pre>$ curl -s
http://archive.cloudera.com/cdh5/debian/wheezy/amd64/cdh/archive.key
 | sudo apt-key add -</pre>
Ubuntu Precise
<pre>$ curl -s
http://archive.cloudera.com/cdh5/ubuntu/precise/amd64/cdh/archive.key
 | sudo apt-key add -</pre>

3. Install the CDH packages:
<pre>$ sudo apt-get update
$ sudo apt-get install avro-tools crunch flume-ng hadoop-hdfs-fuse
hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat
hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms
hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell
kite llama mahout oozie pig pig-udf-datafu search sentry
solr-mapreduce spark-python sqoop sqoop2 whirr</pre>

Note: Installing these packages will also install all the other CDH packages that are needed for a full CDH 5 installation.

</td></tr>
<tr><td>CDH 4, Impala, and Solr</td><td>

Red Hat-compatible

1. Click the entry in the table at CDH Download Information that matches your Red Hat or CentOS system.
2. Navigate to the repo file (<code>cloudera-cdh4.repo</code>) for your system and save it in the <code>/etc/yum.repos.d/</code> directory.
3. Optionally add a repository key:

Red Hat/CentOS/Oracle 5
<pre>$ sudo rpm --import
http://archive.cloudera.com/cdh4/redhat/5/x86_64/cdh/RPM-GPG-KEY-cloudera</pre>
Red Hat/CentOS 6
<pre>$ sudo rpm --import
http://archive.cloudera.com/cdh4/redhat/6/x86_64/cdh/RPM-GPG-KEY-cloudera</pre>

4. Install packages on every host in your cluster:
a. Install CDH 4 packages:
<pre>$ sudo yum -y install bigtop-utils bigtop-jsvc bigtop-tomcat hadoop
 hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn
hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie
 oozie-client pig zookeeper</pre>
</td></tr>
</table>

CDH Version	Procedure

b. To install the hue-common package and all Hue applications on the Hue host, install the hue meta-package:

```
$ sudo yum install hue
```

5. (**Requires CDH 4.2 or later**) Install Impala

   a. Click the entry in the table at Cloudera Impala Version and Download Information that matches your Red Hat or CentOS system.
   b. Navigate to the repo file for your system and save it in the `/etc/yum.repos.d/` directory.
   c. Install Impala and the Impala Shell on Impala machines:

```
$ sudo yum -y install impala impala-shell
```

6. (**Requires CDH 4.3 or later**) Install Search

   a. Click the entry in the table at Cloudera Search Version and Download Information that matches your Red Hat or CentOS system.
   b. Navigate to the repo file for your system and save it in the `/etc/yum.repos.d/` directory.
   c. Install the Solr Server on machines where you want Cloudera Search.

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

- SLES

1. Run the following command:

```
$ sudo zypper addrepo -f
http://archive.cloudera.com/cdh4/sles/11/x86_64/cdh/cloudera-cdh4.repo
```

2. Update your system package index by running:

```
$ sudo zypper refresh
```

3. Optionally add a repository key:

```
$ sudo rpm --import
http://archive.cloudera.com/cdh4/sles/11/x86_64/cdh/RPM-GPG-KEY-cloudera
```

4. Install packages on every host in your cluster:

   a. Install CDH 4 packages:

```
$ sudo zypper install bigtop-utils bigtop-jsvc bigtop-tomcat hadoop
 hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn
hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie
 oozie-client pig zookeeper
```

   b. To install the hue-common package and all Hue applications on the Hue host, install the hue meta-package:

```
$ sudo zypper install hue
```

CDH Version	Procedure

c. (**Requires CDH 4.2 or later**) Install Impala

a. Run the following command:

```
$ sudo zypper addrepo -f
http://archive.cloudera.com/impala/sles/11/x86_64/impala/cloudera-impala.repo
```

b. Install Impala and the Impala Shell on Impala machines:

```
$ sudo zypper install impala impala-shell
```

d. (**Requires CDH 4.3 or later**) Install Search

a. Run the following command:

```
$ sudo zypper addrepo -f
http://archive.cloudera.com/search/sles/11/x86_64/search/cloudera-search.repo
```

b. Install the Solr Server on machines where you want Cloudera Search.

```
$ sudo zypper install solr-server
```

- **Ubuntu or Debian**

1. Click the entry in the table at CDH Version and Packaging Information that matches your Ubuntu or Debian system.
2. Navigate to the list file (`cloudera.list`) for your system and save it in the `/etc/apt/sources.list.d/` directory. For example, to install CDH 4 for 64-bit Ubuntu Lucid, your `cloudera.list` file should look like:

```
deb [arch=amd64]
http://archive.cloudera.com/cdh4/ubuntu/lucid/amd64/cdh lucid-cdh4
contrib
deb-src http://archive.cloudera.com/cdh4/ubuntu/lucid/amd64/cdh
lucid-cdh4 contrib
```

3. Optionally add a repository key:

- **Ubuntu Lucid**

```
$ curl -s
http://archive.cloudera.com/cdh4/ubuntu/lucid/amd64/cdh/archive.key
  | sudo apt-key add -
```

- **Ubuntu Precise**

```
$ curl -s
http://archive.cloudera.com/cdh4/ubuntu/precise/amd64/cdh/archive.key
  | sudo apt-key add -
```

- **Debian Squeeze**

```
$ curl -s
http://archive.cloudera.com/cdh4/debian/squeeze/amd64/cdh/archive.key
  | sudo apt-key add -
```

4. Install packages on every host in your cluster:

CDH Version	Procedure

	a. Install CDH 4 packages:
	`$ sudo apt-get install bigtop-utils bigtop-jsvc bigtop-tomcat hadoop hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie oozie-client pig zookeeper`
	b. To install the hue-common package and all Hue applications on the Hue host, install the hue meta-package:
	`$ sudo apt-get install hue`
	c. (**Requires CDH 4.2 or later**) Install Impala
	a. Click the entry in the table at Cloudera Impala Version and Download Information and that matches your Ubuntu or Debian system.
	b. Navigate to the list file for your system and save it in the `/etc/apt/sources.list.d/` directory.
	c. Install Impala and the Impala Shell on Impala machines:
	`$ sudo apt-get install impala impala-shell`
	d. (**Requires CDH 4.3 or later**) Install Search
	a. Click the entry in the table at Cloudera Search Version and Download Information that matches your Ubuntu or Debian system.
	b. Install Solr Server on machines where you want Cloudera Search:
	`$ sudo apt-get install solr-server`

Start the Cloudera Manager Server

- **Important:** When you start the Cloudera Manager Server and Agents, Cloudera Manager assumes you are not already running HDFS and MapReduce. If these services are running:
  1. Shut down HDFS and MapReduce. See Stopping Services (for CDH 4) or Stopping Services (for CDH 5) for the commands to stop these services.
  2. Configure the init scripts to *not* start on boot, use commands similar to those shown in Configuring init to Start Core Hadoop System Services or Configuring init to Start Core Hadoop System Services but *disable* the start on boot (for example, `$ sudo chkconfig hadoop-hdfs-namenode off`).

  Contact Cloudera Support for help converting your existing Hadoop configurations for use with Cloudera Manager.

1. Run this command on the Cloudera Manager Server host:

```
$ sudo service cloudera-scm-server start
```

If the Cloudera Manager Server does not start, see Troubleshooting Installation and Upgrade Problems on page 131.

(Optional) Start the Cloudera Manager Agents

If you installed the Cloudera Manager Agent packages in Install Cloudera Manager Agent Packages on page 53, run this command on each Agent host:

```
$ sudo service cloudera-scm-agent start
```

When the Agent starts up, it contacts the Cloudera Manager Server. If there is a communication failure between a Cloudera Manager Agent and Cloudera Manager Server, see Troubleshooting Installation and Upgrade Problems on page 131.

When the Agent hosts reboot, `cloudera-scm-agent` starts automatically.

Start the Cloudera Manager Admin Console

The Cloudera Manager Server URL takes the following form `http://Server host:port`, where *Server host* is the fully-qualified domain name or IP address of the host where the Cloudera Manager Server is installed and *port* is the port configured for the Cloudera Manager Server. The default port is 7180.

1. Wait several minutes for the Cloudera Manager Server to complete its startup. To observe the startup process you can perform `tail -f /var/log/cloudera-scm-server/cloudera-scm-server.log` on the Cloudera Manager Server host. If the Cloudera Manager Server does not start, see Troubleshooting Installation and Upgrade Problems on page 131.
2. In a web browser, enter `http://Server host:7180`, where *Server host* is the fully-qualified domain name or IP address of the host where you installed the Cloudera Manager Server. The login screen for Cloudera Manager Admin Console displays.
3. Log into Cloudera Manager Admin Console. The default credentials are: **Username:** `admin` **Password:** `admin`. 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. While 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.

Choose Cloudera Manager Edition and Hosts

The following instructions describe how to use the Cloudera Manager wizard to choose which edition of Cloudera Manager you are using and which hosts will run CDH and managed services.

1. When you start the Cloudera Manager Admin Console, the install wizard starts up. Click **Continue** to get started.
2. Choose which edition to install:
   - Cloudera Express, which does not require a license, but provides a somewhat limited set of features.
   - Cloudera Enterprise Data Hub Edition Trial, which does not require a license, but expires after 60 days and cannot be renewed
   - Cloudera Enterprise with one of the following license types:
     - Basic Edition
     - Flex Edition
     - Data Hub Edition

   If you choose Cloudera Express or Cloudera Enterprise Data Hub Edition Trial, you can elect to upgrade the license at a later time. See Managing Licenses.
3. If you have elected Cloudera Enterprise, install a license:
   a. Click **Upload License**.
   b. Click the document icon to the left of the **Select a License File** text field.
   c. Navigate to the location of your license file, click the file, and click **Open**.
   d. Click **Upload**.

Click **Continue** to proceed with the installation.

4. Click **Continue** in the next screen. The **Specify Hosts** page displays.
5. Do one of the following:
   - If you installed Cloudera Agent packages in Install Cloudera Manager Agent Packages on page 53, choose from among hosts with the packages installed:
     1. Click the **Currently Managed Hosts** tab.
     2. Choose the hosts to add to the cluster.
   - Search for and choose hosts:
     1. To enable Cloudera Manager to automatically discover hosts on which to install CDH and managed services, enter the cluster hostnames or IP addresses. You can also specify hostname and IP address ranges. For example:

Range Definition	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].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

     You can specify multiple addresses and address ranges by separating them by 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. The scan results will include all addresses scanned, but only scans that reach hosts running SSH will be selected for inclusion in your cluster by default. If you don't know the IP addresses of all of the hosts, you can enter an address range that spans over unused addresses and then deselect the hosts that do not exist (and are not discovered) later in this procedure. However, keep in mind that wider ranges will require more time to scan.

     2. Click **Search**. Cloudera Manager identifies the hosts on your cluster to allow you to configure them for services. 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**. To find additional hosts, click **New Search**, add the host names or IP addresses and click **Search** again. 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. Common causes of loss of connectivity are firewalls and interference from SELinux.
     3. Verify that the number of hosts shown matches the number of hosts where you want to install services. Deselect host entries that do not exist and deselect the hosts where you do not want to install services. Click **Continue**. The Select Repository page displays.

6. Click **Continue**. The **Select Repository** page displays.

Choose Software Installation Method and Install Software

The following instructions describe how to use the Cloudera Manager wizard to install Cloudera Manager Agent, CDH, and managed service software.

1. Select how CDH and managed service software is installed: packages or parcels:
   - **Use Packages** - If you *did not* install packages in Install CDH and Managed Service Packages on page 53, click the package versions to install. Otherwise, select the CDH version (CDH 4 or CDH 5) that matches the packages that you installed manually.
   - **Use Parcels**

1. Choose the parcels to install. The choices you see depend on the repositories you have chosen – a repository may contain multiple parcels. Only the parcels for the latest supported service versions are configured by default.

   You can add additional parcels for previous versions by specifying custom repositories. For example, you can find the locations of the previous CDH 4 parcels at `http://archive.cloudera.com/cdh4/parcels/`. Or, if you are installing CDH 4.3 and want to use Sentry for Policy File-Based Hive Authorization, you can add the Sentry parcel using this mechanism.

   1. To specify the parcel directory, local parcel repository, add a parcel repository, or specify the properties of a proxy server through which parcels are downloaded, click the **More Options** button and do one or more of the following:
      - **Parcel Directory** and **Local Parcel Repository Path** - Specify the location of parcels on cluster hosts and the Cloudera Manager Server host.
      - **Parcel Repository** - In the **Remote Parcel Repository URLs** field, click the + button and enter the URL of the repository. The URL you specify is added to the list of repositories listed in the Configuring Server Parcel Settings on page 84 page and a parcel is added to the list of parcels on the Select Repository page. If you have multiple repositories configured, you will see all the unique parcels contained in all your repositories.
      - **Proxy Server** - Specify the properties of a proxy server.
   2. Click **OK**.

2. If you *did not* install Cloudera Manager Agent packages in Install Cloudera Manager Agent Packages on page 53, do the following:
   a. Select the release of Cloudera Manager Agent to install. You can choose either the version that matches the Cloudera Manager Server you are currently using or specify a version in a custom repository.
   b. If you opted to use custom repositories for installation files, you can provide a GPG key URL that applies for all repositories.
3. Click **Continue**.
   - (Cloudera Manager 5.1.3) Leave **Install Oracle Java SE Development Kit (JDK)** checked to allow Cloudera Manager to install the JDK on each cluster host or uncheck if you plan to install it yourself.
   - If your local laws permit you to deploy unlimited strength encryption and you are running a secure cluster, check the **Install Java Unlimited Strength Encryption Policy Files** checkbox.

   Click **Continue**.
4. If your local laws permit you to deploy unlimited strength encryption and you are running a secure cluster, check the **Install Java Unlimited Strength Encryption Policy Files** checkbox.
5. If you chose to have Cloudera Manager install packages, specify host installation properties:
   a. Select **root** or enter the user name for an account that has password-less sudo permission.
   b. Select an authentication method:
      - If you choose to use password authentication, enter and confirm the password.
      - If you choose to use public-key authentication provide a passphrase and path to the required key files.
   c. You can choose to specify an alternate SSH port. The default value is 22.
   d. You can specify the maximum number of host installations to run at once. The default value is 10.
6. Click **Continue**. If you *did not* install packages in (Optional) Install Cloudera Manager Agent, CDH, and Managed Service Software on page 53, Cloudera Manager installs the Oracle JDK, Cloudera Manager Agent,packages and CDH and managed service packages or parcels. During the parcel installation, progress is indicated for the two phases of the parcel installation process (Download and Distribution) in a separate progress bars. If you are installing multiple parcels you will see progress bars for each parcel. When the **Continue** button appears at the bottom of the screen, the installation process is completed. Click **Continue**.

7. Click **Continue**. The Host Inspector runs to validate the installation, and provides a summary of what it finds, including all the versions of the installed components. If the validation is successful, click **Finish**. The Cluster Setup page displays.

Add Services

The following instructions describe how to use the Cloudera Manager wizard to configure and start CDH and managed services.

1. In the first page of the Add Services wizard you choose the combination of services to install and whether to install Cloudera Navigator:
   - Click the radio button next to the combination of services to install:

CDH 4	CDH 5
▪ **Core Hadoop** - HDFS, MapReduce, ZooKeeper, Oozie, Hive, and Hue ▪ **Core with HBase** ▪ **Core with Impala** ▪ **All Services** - HDFS, MapReduce, ZooKeeper, HBase, Impala, Oozie, Hive, Hue, and Sqoop ▪ **Custom Services** - Any combination of services.	▪ **Core Hadoop** - HDFS, YARN (includes MapReduce 2), ZooKeeper, Oozie, Hive, Hue, and Sqoop ▪ **Core with HBase** ▪ **Core with Impala** ▪ **Core with Search** ▪ **Core with Spark** ▪ **All Services** - HDFS, YARN (includes MapReduce 2), ZooKeeper, Oozie, Hive, Hue, Sqoop, HBase, Impala, Solr, Spark, and Key-Value Store Indexer ▪ **Custom Services** - Any combination of services.

     As you select the services, keep the following in mind:
     - Some services depend on other services; for example, HBase requires HDFS and ZooKeeper. Cloudera Manager tracks dependencies and installs the correct combination of services.
     - In a CDH 4 cluster, the MapReduce service is the default MapReduce computation framework. Choose **Custom Services** to install YARN or use the Add Service functionality to add YARN after installation completes.

       > ▪ **Important:** You can create a YARN service in a CDH 4 cluster, but it is not considered production ready.

     - In a CDH 5 cluster, the YARN service is the default MapReduce computation framework. Choose **Custom Services** to install MapReduce or use the Add Service functionality to add MapReduce after installation completes.

       > ▪ **Important:** In CDH 5 the MapReduce service has been deprecated. However, the MapReduce service is fully supported for backward compatibility through the CDH 5 life cycle.

     - The Flume service can be added only after your cluster has been set up.
   - If you have chosen Data Hub Edition Trial or Cloudera Enterprise, optionally check the **Include Cloudera Navigator** checkbox to enable Cloudera Navigator. See the Cloudera Navigator Documentation.

   Click **Continue**. The Customize Role Assignments page displays.
2. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign role instances to hosts of your choosing, if desired.

Click a field below a role to display a dialog containing a pageable list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	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].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
- Rack name

Click the **View By Host** button for an overview of the role assignment by hostname ranges.

3. When you are satisfied with the assignments, click **Continue**. The Database Setup page displays.
4. On the Database Setup page, configure settings for required databases:
   a. Enter the database host, database type, database name, username, and password for the database that you created when you set up the database.
   b. Click **Test Connection** to confirm that Cloudera Manager can communicate with the database using the information you have supplied. If the test succeeds in all cases, click **Continue**; otherwise check and correct the information you have provided for the database and then try the test again. (For some servers, if you are using the embedded database, you will see a message saying the database will be created at a later step in the installation process.) The Review Changes page displays.
5. Review the configuration changes to be applied. Confirm the settings entered for file system paths. The file paths required vary based on the services to be installed.

   - **Warning:** DataNode data directories should not be placed on NAS devices.

   Click **Continue**. The wizard starts the services.
6. When all of the services are started, click **Continue**. You will see a success message indicating that your cluster has been successfully started.
7. Click **Finish** to proceed to the Home Page.

Configure Cluster CDH Version for Package Installs

If you have installed CDH as a package, after an install or upgrade make sure that the cluster CDH version matches the package CDH version, using the procedure in Configuring the CDH Version for a Cluster in *Managing Clusters with Cloudera Manager*. If the cluster CDH version does not match the package CDH version, Cloudera Manager will incorrectly enable and disable service features based on the cluster's configured CDH version.

Change the Default Administrator Password

As soon as possible after running the wizard and beginning to use Cloudera Manager, change the default administrator password:

1. Right-click the logged-in username at the far right of the top navigation bar and select **Change Password**.
2. Enter the current password, and a new password twice and then click **Update**.

Test the Installation

You can test the installation following the instructions in Testing the Installation on page 123.

Installation Path C - Manual Installation Using Cloudera Manager Tarballs

Before proceeding with this path for a new installation, review Choosing an Installation Path on page 43. If you are upgrading an Cloudera Manager existing installation, see Upgrading Cloudera Manager.

To avoid using system packages, and to use tarballs and parcels instead, follow the instructions in this section.

> **Note:** When installing with tarballs and parcels, some services may require additional dependencies which are not provided by Cloudera. To determine these dependencies, check the logs if a service fails to start or has errors. The logs should specify whether there are missing dependencies, which you then must install manually.

Before You Begin

Install and Configure Databases

Cloudera Manager Server, Cloudera Management Service, and the Hive Metastore data is stored in a database. Install and configure required databases following the instructions in Cloudera Manager and Managed Service Databases on page 21.

(CDH 5 only) On RHEL and CentOS 5, Install Python 2.6 or 2.7

Python 2.6 or 2.7 is required to run Hue. RHEL 5 and CentOS 5, in particular, require the EPEL repository package.

In order to install packages from the EPEL repository, first download the appropriate repository rpm packages to your machine and then install Python using `yum`. For example, use the following commands for RHEL 5 or CentOS 5:

```
$ su -c 'rpm -Uvh
http://download.fedoraproject.org/pub/epel/5/i386/epel-release-5-4.noarch.rpm'
...
$ yum install python26
```

Install the Cloudera Manager Server and Agents

Tarballs contain both the Cloudera Manager Server and Cloudera Manager Agent in a single file. Download tarballs from the locations listed in Cloudera Manager Version and Download Information. Copy the tarballs and unpack them on all hosts on which you intend to install Cloudera Manager Server and Cloudera Manager Agents in a location of your choosing. If necessary, create a new directory to accommodate the files you extract from the tarball. For instance if `/opt/cloudera-manager` does not exist, create that using a command similar to:

```
$ sudo mkdir /opt/cloudera-manager
```

When you have a location to which to extract the contents of the tarball, extract the contents. For example, to copy a tar file to your home directory and extract the contents of all tar files to the `/opt/` directory, you might use a command similar to the following:

```
$ tar xzf cloudera-manager*.tar.gz -C /opt/cloudera-manager
```

The files are extracted to a subdirectory named according to the Cloudera Manager version being extracted. For example, files could extract to `/opt/cloudera-manager/cm-5.0/`. This full path is needed later and is referred to as *tarball root* directory.

Create Users

The Cloudera Manager Server and managed services need a user account to complete tasks. When installing Cloudera Manager from tarballs, you much create this user account on all hosts manually. Because Cloudera Manager Server and managed services are configured to use the user account `cloudera-scm` by default, creating a user with this name is the simplest approach. After creating such a user, it is automatically used after installation is complete.

To create a user `cloudera-scm`, use a command such as the following:

```
$ useradd --system --home=/opt/cloudera-manager/cm-5.0/run/cloudera-scm-server
--no-create-home --shell=/bin/false --comment "Cloudera SCM User" cloudera-scm
```

For the preceding useradd command, ensure the `--home` argument path matches your environment. This argument varies according to where you place the tarball and the version number varies among releases. For example, the `--home` location could be `/opt/cm-5.0/run/cloudera-scm-server`.

Configure Cloudera Manager Agents

On every Cloudera Manager Agent host, configure the Cloudera Manager Agent to point to the Cloudera Manager Server by setting the following properties in the *`tarball root`*`/etc/cloudera-scm-agent/config.ini` configuration file:

Property	Description
`server_host`	Name of host where the Cloudera Manager Server is running.
`server_port`	Port on host where the Cloudera Manager Server is running.

Custom Cloudera Manager Users and Directories

Cloudera Manager is built to use a default set of directories and user accounts. You can use the default locations and accounts, but there is also the option to change these settings. In some cases, changing these settings is required. For most installations, you can skip ahead to Configure a Database for the Cloudera Manager Server on page 67. By default, Cloudera Manager services creates directories in `/var/log` and `/var/lib`. The directories the Cloudera Manager installer attempts to create are:

- `/var/log/cloudera-scm-headlamp`
- `/var/log/cloudera-scm-firehose`
- `/var/log/cloudera-scm-alertpublisher`
- `/var/log/cloudera-scm-eventserver`
- `/var/lib/cloudera-scm-headlamp`
- `/var/lib/cloudera-scm-firehose`
- `/var/lib/cloudera-scm-alertpublisher`
- `/var/lib/cloudera-scm-eventserver`

If you are using a custom user and directory for Cloudera Manager, you must create these directories on the Cloudera Manager Server host and assign ownership of these directories to your user manually. Issues might arise if any of these directories already exist. The Cloudera Manager installer makes no changes to existing directories. In such a case, Cloudera Manager is unable to write to any existing directories for which it does not have proper permissions and services may not perform as expected.

Two ways to resolve such situations are: Changing the ownership of existing directories or specifying alternate directories for agents. You do not need to complete both procedures.

To change ownership for existing directories:

1. Change the directory owner to the Cloudera Manager user. If the Cloudera Manager user and group are `cloudera-scm` and you needed to take ownership of the headlamp log directory, you would issue a command similar to the following:

```
$ chown -R cloudera-scm:cloudera-scm /var/log/cloudera-scm-headlamp
```

2. Repeat the process of using chown to change ownership for all existing directories to the Cloudera Manager user.

To use alternate directories for services:

1. If the directories you plan to use do not exist, create them now. For example to create `/var/cm_logs/cloudera-scm-headlamp` for use by the `cloudera-scm` user, you might use the following commands:

```
mkdir /var/cm_logs/cloudera-scm-headlamp
chown cloudera-scm /var/cm_logs/cloudera-scm-headlamp
```

2. Connect to the Cloudera Manager Admin Console.
3. Under the Cloudera Managed Services, click the name of the service.
4. In the service status page, click **Configuration**.
5. In the settings page, enter a term in the **Search** field to find the settings to be change. For example, you might enter "/var" or "directory".
6. Update each value with the new locations for Cloudera Manager to use.
7. Click **Save Changes**.

Configure a Database for the Cloudera Manager Server

Set up the Cloudera Manager Server database as described in Setting up the Cloudera Manager Server Database on page 22.

Create a Parcel Repository Directory

1. Create a parcel repository directory:

```
$ mkdir -p /opt/cloudera/parcel-repo
```

2. Change the directory ownership to be the username you are using to run Cloudera Manager:

```
$ chown username:groupname /opt/cloudera/parcel-repo
```

where *username* and *groupname* are the user and group names (respectively) you are using to run Cloudera Manager. For example, if you use the default username `cloudera-scm`, you would give the command:

```
$ chown cloudera-scm:cloudera-scm /opt/cloudera/parcel-repo
```

Start the Cloudera Manager Server

- **Important:** When you start the Cloudera Manager Server and Agents, Cloudera Manager assumes you are not already running HDFS and MapReduce. If these services are running:
  1. Shut down HDFS and MapReduce. See Stopping Services (for CDH 4) or Stopping Services (for CDH 5) for the commands to stop these services.
  2. Configure the init scripts to *not* start on boot, use commands similar to those shown in Configuring init to Start Core Hadoop System Services or Configuring init to Start Core Hadoop System Services but *disable* the start on boot (for example, `$ sudo chkconfig hadoop-hdfs-namenode off`).

  Contact Cloudera Support for help converting your existing Hadoop configurations for use with Cloudera Manager.

The way in which you start the Cloudera Manager Server varies according to what account you want the server to run under:

- As root:

```
$ sudo tarball root/etc/init.d/cloudera-scm-server start
```

- As another user. If you run as another user, ensure the user you created for Cloudera Manager owns the location to which you extracted the tarball including the newly created database files. If you followed the earlier examples and created the directory `/opt/cloudera-manager` and the user `cloudera-scm`, you could use the following command to change ownership of the directory:

```
$ sudo chown -R cloudera-scm:cloudera-scm /opt/cloudera-manager
```

  Once you have established proper ownership of directory locations, you can start Cloudera Manager Server using the user account you chose. For example, you might run the Cloudera Manager Server as `cloudera-service`. In such a case there are following options:

  – Run the following command:

```
$ sudo -u user tarball root/etc/init.d/cloudera-scm-server start
```

  – Edit the configuration files so the script internally changes the user. Then run the script as root. To make this possible, complete the following steps:

    1. Remove the following line from *tarball root*`/etc/default/cloudera-scm-server`:

```
export CMF_SUDO_CMD=" "
```

       Change the user and group in *tarball root*`/etc/init.d/cloudera-scm-server` to the user you want the server to run as. For example, to run as cloudera-service, change the user and group as follows:

```
USER=cloudera-service
GROUP=cloudera-service
```

    2. Run the server script as root:

```
$ sudo tarball root/etc/init.d/cloudera-scm-server start
```

- To start the Cloudera Manager Server automatically after a reboot:
  1. On the Cloudera Manager Server host, open the `/etc/init.d/cloudera-scm-server` file and change the value of `CMF_DEFAULTS` from `${CMF_DEFAULTS:-/etc/default}` to *tarball root*`/etc/default`.

- Run the following commands on the Cloudera Manager Server host:
  - **RHEL-compatible and SLES**

```
$ cp tarball root/etc/init.d/cloudera-scm-server /etc/init.d/cloudera-scm-server
$ chkconfig cloudera-scm-server on
```

  - **Debian/Ubuntu**

```
$ cp tarball root/etc/init.d/cloudera-scm-server /etc/init.d/cloudera-scm-server
$ update-rc.d cloudera-scm-server defaults
```

If the Cloudera Manager Server does not start, see Troubleshooting Installation and Upgrade Problems on page 131.

Start the Cloudera Manager Agents

- To start the Cloudera Manager Agent, run this command on each Agent host:

```
$ sudo tarball root/etc/init.d/cloudera-scm-agent start
```

  When the Agent starts, it contacts the Cloudera Manager Server.
- To start the Cloudera Manager Agents automatically after a reboot:
  1. Run the following commands on each Agent host:
     - **RHEL-compatible and SLES**

```
$ cp tarball root/etc/init.d/cloudera-scm-agent /etc/init.d/cloudera-scm-agent
$ chkconfig cloudera-scm-agent on
```

     - **Debian/Ubuntu**

```
$ cp tarball root/etc/init.d/cloudera-scm-agent /etc/init.d/cloudera-scm-agent
$ update-rc.d cloudera-scm-agent defaults
```

  2. On each Agent, open the *tarball root*`/etc/init.d/cloudera-scm-agent` file and change the value of `CMF_DEFAULTS` from `${CMF_DEFAULTS:-/etc/default}` to *tarball root*`/etc/default`.

Start the Cloudera Manager Admin Console

The Cloudera Manager Server URL takes the following form `http://`*Server host*`:port`, where *Server host* is the fully-qualified domain name or IP address of the host where the Cloudera Manager Server is installed and *port* is the port configured for the Cloudera Manager Server. The default port is 7180.

1. Wait several minutes for the Cloudera Manager Server to complete its startup. To observe the startup process you can perform `tail -f /var/log/cloudera-scm-server/cloudera-scm-server.log` on the Cloudera Manager Server host. If the Cloudera Manager Server does not start, see Troubleshooting Installation and Upgrade Problems on page 131.
2. In a web browser, enter `http://`*Server host*`:7180`, where *Server host* is the fully-qualified domain name or IP address of the host where you installed the Cloudera Manager Server. The login screen for Cloudera Manager Admin Console displays.
3. Log into Cloudera Manager Admin Console. The default credentials are: **Username:** `admin` **Password:** `admin`. 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. While 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.

Choose Cloudera Manager Edition and Hosts

1. When you start the Cloudera Manager Admin Console, the install wizard starts up. Click **Continue** to get started.
2. Choose which edition to install:
   - Cloudera Express, which does not require a license, but provides a somewhat limited set of features.
   - Cloudera Enterprise Data Hub Edition Trial, which does not require a license, but expires after 60 days and cannot be renewed
   - Cloudera Enterprise with one of the following license types:
     - Basic Edition
     - Flex Edition
     - Data Hub Edition

   If you choose Cloudera Express or Cloudera Enterprise Data Hub Edition Trial, you can elect to upgrade the license at a later time. See Managing Licenses.
3. If you have elected Cloudera Enterprise, install a license:
   a. Click **Upload License**.
   b. Click the document icon to the left of the **Select a License File** text field.
   c. Navigate to the location of your license file, click the file, and click **Open**.
   d. Click **Upload**.

   Click **Continue** to proceed with the installation.
4. Click **Continue** in the next screen. The **Specify Hosts** page displays.
5. Click the **Currently Managed Hosts** tab.
6. Choose the hosts to add to the cluster.
7. Click **Continue**. The **Select Repository** page displays.

Choose Software Installation Method and Install Software

1. Click **Use Parcels** to install CDH and managed services using parcels and then do the following:
   a. Choose the parcels to install. The choices you see depend on the repositories you have chosen – a repository may contain multiple parcels. Only the parcels for the latest supported service versions are configured by default.

      You can add additional parcels for previous versions by specifying custom repositories. For example, you can find the locations of the previous CDH 4 parcels at `http://archive.cloudera.com/cdh4/parcels/`. Or, if you are installing CDH 4.3 and want to use Sentry for Policy File-Based Hive Authorization, you can add the Sentry parcel using this mechanism.

      1. To specify the parcel directory, local parcel repository, add a parcel repository, or specify the properties of a proxy server through which parcels are downloaded, click the **More Options** button and do one or more of the following:
         - **Parcel Directory** and **Local Parcel Repository Path** - Specify the location of parcels on cluster hosts and the Cloudera Manager Server host.
         - **Parcel Repository** - In the **Remote Parcel Repository URLs** field, click the + button and enter the URL of the repository. The URL you specify is added to the list of repositories listed in the Configuring Server Parcel Settings on page 84 page and a parcel is added to the list of parcels on the Select Repository page. If you have multiple repositories configured, you will see all the unique parcels contained in all your repositories.
         - **Proxy Server** - Specify the properties of a proxy server.
      2. Click **OK**.

b. Click **Continue**. Cloudera Manager installs the CDH and managed service parcels. During the parcel installation, progress is indicated for the two phases of the parcel installation process (Download and Distribution) in a separate progress bars. If you are installing multiple parcels you will see progress bars for each parcel. When the **Continue** button appears at the bottom of the screen, the installation process is completed. Click **Continue**.

2. Click **Continue**. The Host Inspector runs to validate the installation, and provides a summary of what it finds, including all the versions of the installed components. If the validation is successful, click **Finish**. The Cluster Setup page displays.

Add Services

The following instructions describe how to use the Cloudera Manager wizard to configure and start CDH and managed services.

1. In the first page of the Add Services wizard you choose the combination of services to install and whether to install Cloudera Navigator:
   - Click the radio button next to the combination of services to install:

CDH 4	CDH 5
▪ **Core Hadoop** - HDFS, MapReduce, ZooKeeper, Oozie, Hive, and Hue ▪ **Core with HBase** ▪ **Core with Impala** ▪ **All Services** - HDFS, MapReduce, ZooKeeper, HBase, Impala, Oozie, Hive, Hue, and Sqoop ▪ **Custom Services** - Any combination of services.	▪ **Core Hadoop** - HDFS, YARN (includes MapReduce 2), ZooKeeper, Oozie, Hive, Hue, and Sqoop ▪ **Core with HBase** ▪ **Core with Impala** ▪ **Core with Search** ▪ **Core with Spark** ▪ **All Services** - HDFS, YARN (includes MapReduce 2), ZooKeeper, Oozie, Hive, Hue, Sqoop, HBase, Impala, Solr, Spark, and Key-Value Store Indexer ▪ **Custom Services** - Any combination of services.

     As you select the services, keep the following in mind:
     - Some services depend on other services; for example, HBase requires HDFS and ZooKeeper. Cloudera Manager tracks dependencies and installs the correct combination of services.
     - In a CDH 4 cluster, the MapReduce service is the default MapReduce computation framework. Choose **Custom Services** to install YARN or use the Add Service functionality to add YARN after installation completes.

       ▪ **Important:** You can create a YARN service in a CDH 4 cluster, but it is not considered production ready.

     - In a CDH 5 cluster, the YARN service is the default MapReduce computation framework. Choose **Custom Services** to install MapReduce or use the Add Service functionality to add MapReduce after installation completes.

       ▪ **Important:** In CDH 5 the MapReduce service has been deprecated. However, the MapReduce service is fully supported for backward compatibility through the CDH 5 life cycle.

     - The Flume service can be added only after your cluster has been set up.
   - If you have chosen Data Hub Edition Trial or Cloudera Enterprise, optionally check the **Include Cloudera Navigator** checkbox to enable Cloudera Navigator. See the Cloudera Navigator Documentation.

   Click **Continue**. The Customize Role Assignments page displays.

2. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign role instances to hosts of your choosing, if desired.

   Click a field below a role to display a dialog containing a pageable list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

   The following shortcuts for specifying hostname patterns are supported:

   - Range of hostnames (without the domain portion)

Range Definition	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].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

   - IP addresses
   - Rack name

   Click the **View By Host** button for an overview of the role assignment by hostname ranges.

3. When you are satisfied with the assignments, click **Continue**. The Database Setup page displays.
4. On the Database Setup page, configure settings for required databases:
   a. Enter the database host, database type, database name, username, and password for the database that you created when you set up the database.
   b. Click **Test Connection** to confirm that Cloudera Manager can communicate with the database using the information you have supplied. If the test succeeds in all cases, click **Continue**; otherwise check and correct the information you have provided for the database and then try the test again. (For some servers, if you are using the embedded database, you will see a message saying the database will be created at a later step in the installation process.) The Review Changes page displays.
5. Review the configuration changes to be applied. Confirm the settings entered for file system paths. The file paths required vary based on the services to be installed.

   > - **Warning:** DataNode data directories should not be placed on NAS devices.

   Click **Continue**. The wizard starts the services.
6. When all of the services are started, click **Continue**. You will see a success message indicating that your cluster has been successfully started.
7. Click **Finish** to proceed to the Home Page.

(Optional) Change the Cloudera Manager User

After configuring your services, the installation wizard attempts to automatically start the Cloudera Management Service under the assumption that it will run using `cloudera-scm`. If you configured this service to run using a user other than `cloudera-scm`, then the Cloudera Management Service roles do not start automatically. In such a case, change the service configuration to use the user account that you selected:

1. Connect to the Cloudera Manager Admin Console.
2. Do one of the following:
   - Select **Clusters** > **Cloudera Management Service** > **Cloudera Management Service**.
   - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.

3. Click the **Configuration** tab.
4. Use the search box to find the property to be changed. For example, you might enter "system" to find the **System User** and **System Group** properties.
5. Make any changes required to the System User and System Group to ensure Cloudera Manager uses the proper user accounts.
6. Click **Save Changes**.

After making this configuration change, manually start the Cloudera Management Service roles.

Change the Default Administrator Password

As soon as possible after running the wizard and beginning to use Cloudera Manager, change the default administrator password:

1. Right-click the logged-in username at the far right of the top navigation bar and select **Change Password**.
2. Enter the current password, and a new password twice and then click **Update**.

Test the Installation

You can test the installation following the instructions in Testing the Installation on page 123.

Installing Impala

Cloudera Impala is included with CDH 5. To use Cloudera Impala with CDH 4, you must install both CDH and Impala on the hosts that will run Impala.

> **Note:**
>
> - See Supported CDH and Managed Service Versions on page 16 for supported versions.
> - Before proceeding, review the installation options described in Choosing an Installation Path on page 43.

Installing Impala after Upgrading Cloudera Manager

If you have just upgraded Cloudera Manager from a version that did not support Impala, the Impala software is not installed automatically. (Upgrading Cloudera Manager does not automatically upgrade CDH or other managed services). You can add Impala using parcels; go to the **Hosts** tab, and select the **Parcels** tab. You should see at least one Impala parcel available for download. See Parcels on page 77 for detailed instructions on using parcels to install or upgrade Impala. If you do not see any Impala parcels available, click the **Edit Settings** button on the **Parcels** page to go to the Parcel configuration settings and verify that the Impala parcel repo URL (http://archive.cloudera.com/impala/parcels/latest/) has been configured in the **Parcels** configuration page. See Parcel Configuration Settings on page 84 for more details.

Post Installation Configuration

See The Impala Service in *Managing Clusters with Cloudera Manager* for instructions on configuring the Impala service.

Installing Search

Cloudera Search is provided by the Solr service. The Solr service is included with CDH 5. To use Cloudera Search with CDH 4, you must install both CDH and Search on the hosts that will run Search.

- **Note:**
  - See Supported CDH and Managed Service Versions on page 16 for supported versions.
  - Before proceeding, review the installation options described in Choosing an Installation Path on page 43.

Installing Search after Upgrading Cloudera Manager

If you have just upgraded Cloudera Manager from a version that did not support Search, the Search software is not installed automatically. (Upgrading Cloudera Manager does not automatically upgrade CDH or other managed services). You can add Search using parcels; go to the **Hosts** tab, and select the **Parcels** tab. You should see at least one Solr parcel available for download. See Parcels on page 77 for detailed instructions on using parcels to install or upgrade Solr. If you do not see any Solr parcels available, click the **Edit Settings** button on the **Parcels** page to go to the Parcel configuration settings and verify that the Search parcel repo URL (http://archive.cloudera.com/search/parcels/latest/) has been configured in the **Parcels** configuration page. See Parcel Configuration Settings on page 84 for more details.

Post Installation Configuration

See The Solr Service in *Managing Clusters with Cloudera Manager* for instructions on configuring Cloudera Search.

Installing Spark

Apache Spark is included with CDH 5. To use Apache Spark with CDH 4, you must install both CDH and Spark on the hosts that will run Spark.

- **Note:**
  - See Supported CDH and Managed Service Versions on page 16 for supported versions.
  - Before proceeding, review the installation options described in Choosing an Installation Path on page 43.

Installing Spark after Upgrading Cloudera Manager

If you have just upgraded Cloudera Manager from a version that did not support Spark, the Spark software is not installed automatically. (Upgrading Cloudera Manager does not automatically upgrade CDH or other managed services).

You can add Spark using parcels; go to the **Hosts** tab, and select the **Parcels** tab. You should see at least one Spark parcel available for download. See Parcels on page 77 for detailed instructions on using parcels to install or upgrade Spark. If you do not see any Spark parcels available, click the **Edit Settings** button on the **Parcels** page to go to the Parcel configuration settings and verify that the Spark parcel repo URL (http://archive.cloudera.com/spark/parcels/latest/) has been configured in the **Parcels** configuration page. See Parcel Configuration Settings on page 84 for more details.

Post Installation Configuration

See The Spark Service in *Managing Clusters with Cloudera Manager* for instructions on adding the Spark service.

Installing GPL Extras

GPL Extras contains LZO functionality.

To install the GPL Extras parcel:

1. Add the appropriate repository to the Cloudera Manager list of parcel repositories. The public repositories can be found at:
   - **CDH 5** - http://archive.cloudera.com/gplextras5/parcels/latest
   - **CDH 4** - http://archive.cloudera.com/gplextras/parcels/latest

   If you are using LZO with Impala, you must choose a specific version of the GPL Extras parcel for the Impala version according to the following table:

Impala Version	Version Directory	GPL Extras Parcel Version
CDH 5.*x.y*	`5.x.y/`	GPLEXTRAS-5.*x.y*
1.4.0	`0.4.15.85/`	HADOOP_LZO-0.4.15-1.gplextras.p0.85
1.3.1	`0.4.15.64/`	HADOOP_LZO-0.4.15-1.gplextras.p0.64
1.2.4	`0.4.15.58/`	HADOOP_LZO-0.4.15-1.gplextras.p0.58
1.2.3	`0.4.15.39/`	HADOOP_LZO-0.4.15-1.gplextras.p0.39
1.2.2	`0.4.15.37/`	HADOOP_LZO-0.4.15-1.gplextras.p0.37
1.2.1	`0.4.15.33/`	HADOOP_LZO-0.4.15-1.gplextras.p0.33

   To create the repository URL, append the version directory to the URL (CDH 4) `http://archive.cloudera.com/gplextras/parcels/` or (CDH 5) `http://archive.cloudera.com/gplextras5/parcels/` respectively. For example: http://archive.cloudera.com/gplextras5/parcels/5.0.2.
2. Download, distribute, and activate the parcel.

Managing Software Distribution

A major function of Cloudera Manager is to distribute and activate software in your cluster. Cloudera Manager supports two software distribution formats: packages and parcels.

A ***package*** is a binary distribution format that contains compiled code and meta-information such as a package description, version, and dependencies. Package management systems evaluate this meta-information to allow package searches, perform upgrades to a newer version, and ensure that all dependencies of a package are fulfilled. Cloudera Manager uses the native "system package manager" for each supported OS.

A ***parcel*** is a binary distribution format containing the program files, along with additional metadata used by Cloudera Manager. There are a few notable differences between parcels and packages:

- Parcels are self-contained and installed in a versioned directory. This means that multiple versions of a given parcel can be installed side-by-side. You can then designate one of these installed versions as the active one. With traditional packages, only one package can be installed at a time so there's no distinction between what's installed and what's active.
- Parcels can be installed at any location in the filesystem. By default, parcels are installed in `/opt/cloudera/parcels`.

Parcels

Required Role: Administrator

A ***parcel*** is a binary distribution format containing the program files, along with additional metadata used by Cloudera Manager. There are a few notable differences between parcels and packages:

- Parcels are self-contained and installed in a versioned directory. This means that multiple versions of a given parcel can be installed side-by-side. You can then designate one of these installed versions as the active one. With traditional packages, only one package can be installed at a time so there's no distinction between what's installed and what's active.
- Parcels can be installed at any location in the filesystem. By default, parcels are installed in `/opt/cloudera/parcels`.

Parcels are available for CDH 4.1.3 or later, and for Impala, Search, Spark, and Accumulo.

Advantages of Parcels

As a consequence of their unique properties, parcels offer a number of advantages over packages:

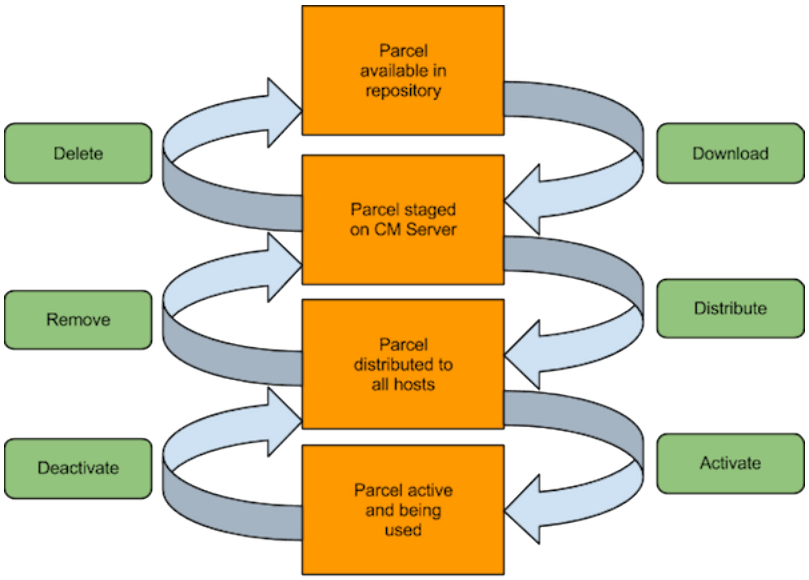
- **CDH is distributed as a single object** - In contrast to having a separate package for each part of CDH, when using parcels there is just a single object to install. This is especially useful when managing a cluster that isn't connected to the Internet.
- **Internal consistency** - All CDH components are matched so there isn't a danger of different parts coming from different versions of CDH.
- **Installation outside of `/usr`** - In some environments, Hadoop administrators do not have privileges to install system packages. In the past, these administrators had to fall back to CDH tarballs, which deprived them of a lot of infrastructure that packages provide. With parcels, administrators can install to `/opt` or anywhere else without having to step through all the additional manual steps of regular tarballs.

> - **Note:** With parcel software distribution, the path to the CDH libraries is `/opt/cloudera/parcels/CDH/lib` instead of the usual `/usr/lib`. You should not link `/usr/lib/` elements to parcel deployed paths, as such links may confuse scripts that distinguish between the two paths.

- **Installation of CDH without sudo** - Parcel installation is handled by the Cloudera Manager Agent running as root so it's possible to install CDH without needing sudo.
- **Decouples distribution from activation** - Due to side-by-side install capabilities, it is possible to stage a new version of CDH across the cluster in advance of switching over to it. This allows the longest running part of an upgrade to be done ahead of time without affecting cluster operations, consequently reducing the downtime associated with upgrade.
- **Rolling upgrades** - These are only possible with parcels, due to their side-by-side nature. Packages require shutting down the old process, upgrading the package, and then starting the new process. This can be hard to recover from in the event of errors and requires extensive integration with the package management system to function seamlessly. When a new version is staged side-by-side, switching to a new minor version is simply a matter of changing which version of CDH is used when restarting each process. It then becomes practical to do upgrades with rolling restarts, where service roles are restarted in the right order to switch over to the new version with minimal service interruption. Your cluster can continue to run on the existing installed components while you stage a new version across your cluster, without impacting your current operations. Note that major version upgrades (for example, CDH 4 to CDH 5) require full service restarts due to the substantial changes between the versions. Finally, you can upgrade individual parcels, or multiple parcels at the same time.
- **Easy downgrades** - Reverting back to an older minor version can be as simple as upgrading. Note that some CDH components may require explicit additional steps due to schema upgrades.
- **Upgrade management** - Cloudera Manager can fully manage all the steps involved in a CDH version upgrade. In contrast, with packages, Cloudera Manager can only help with initial installation.
- **Distributing additional components** - Parcels are not limited to CDH. Cloudera Impala, Cloudera Search, LZO, and add-on service parcels are also available.
- **Compatibility with other distribution tools** - If there are specific reasons to use other tools for download and/or distribution, you can do so, and Cloudera Manager will work alongside your other tools. For example, you can handle distribution with Puppet. Or, you can download the parcel to Cloudera Manager Server manually (perhaps because your cluster has no Internet connectivity) and then have Cloudera Manager distribute the parcel to the cluster.

Parcel Life Cycle

To enable upgrades and additions with minimal disruption, parcels participate in six phases: download, distribute, activate: deactivate, remove, and delete.

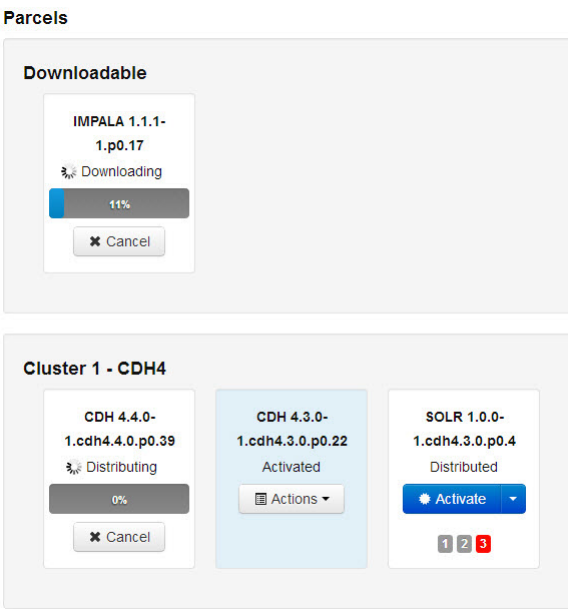


- **Downloading** a parcel copies the appropriate software to a local parcel repository on the Cloudera Manager Server, where it is available for distribution to the other hosts in any of your clusters managed by this Cloudera Manager Server. You can have multiple parcels for a given product downloaded to your Cloudera Manager Server. Once a parcel has been downloaded to the Server, it will be available for distribution on all clusters managed by the Server. A downloaded parcel will appear in the cluster-specific section for every cluster managed by this Cloudera Manager Server.
- **Distributing** a parcel copies the parcel to the member hosts of a cluster and unpacks it. Distributing a parcel does not actually upgrade the components running on your cluster; the current services continue to run unchanged. You can have multiple parcels distributed on your cluster.

  > **Note:** The distribute process does not require Internet access; rather the Cloudera Manager Agent on each cluster member downloads the parcels from the local parcel repository on the Cloudera Manager Server.

- **Activating** a parcel causes the Cloudera Manager to link to the new components, ready to run the new version *upon the next restart*. Activation does not automatically stop the current services or perform a restart — you have the option to restart the service(s) after activation, or you can allow the system administrator to determine the appropriate time to perform those operations.
- **Deactivating** a parcel causes Cloudera Manager to unlink from the parcel components. A parcel cannot be deactivated while it is still in use on one or more hosts.
- **Removing** a parcel causes Cloudera Manager to remove the parcel components from the hosts.
- **Deleting** a parcel causes Cloudera Manager to remove the parcel components from the local parcel repository.

For example, the following screenshot:



shows:

- One activated CDH parcel
- One SOLR parcel distributed and ready to activate
- One Impala parcel being downloaded
- One CDH parcel being distributed

Cloudera Manager detects when new parcels are available. The parcel indicator in the Admin Console navigation bar (5) indicates how many parcels are eligible for downloading or distribution. For example, CDH parcels older than the active one do not contribute to the count if you are already using the latest version. If no parcels

are eligible, or if all parcels have been activated, then the indicator will not have a number badge. You can configure Cloudera Manager to download and distribute parcels automatically, if desired.

> **Important:** If you plan to upgrade CDH you should follow the instructions in Upgrading CDH and Managed Services. There are additional steps that must be performed in order to successfully upgrade.

Parcel Locations

The default location for the local parcel directory on the Cloudera Manager Server host is `/opt/cloudera/parcel-repo`. To change this location, follow the instructions in Configuring Server Parcel Settings on page 84.

The default location for the distributed parcels on the managed hosts is `/opt/cloudera/parcels`. To change this location, set the `parcel_dir` property in `/etc/cloudera-scm-agent/config.ini` file of the Cloudera Manager Agent and restart the Cloudera Manager Agent or by following the instructions in Configuring the Host Parcel Directory on page 85.

> **Note:** With parcel software distribution, the path to the CDH libraries is `/opt/cloudera/parcels/CDH/lib` instead of the usual `/usr/lib`. You should not link `/usr/lib/` elements to parcel deployed paths, as such links may confuse scripts that distinguish between the two paths.

Managing Parcels

Through the Parcels interface in Cloudera Manager, you can determine what software versions are running across your clusters. You access the Parcels page by doing one of the following:

- Clicking the parcel indicator in the Admin Console navigation bar ( )
- Clicking the **Hosts** in the top navigation bar, then the **Parcels** tab.

The Parcels page is divided into several sections. The top section, labeled **Downloadable**, shows you all the parcels that are available for download from the configured parcel repositories.

Below the Downloadable section, each cluster managed by this Cloudera Manager Server has a section that shows the parcels that have been downloaded, distributed, or activated on that cluster.

When you download a parcel, it appears under every cluster, if you are managing more than one. However, this just indicates that the parcel is available for distribution on those clusters — in fact there is only one copy of the downloaded parcel, residing on the Cloudera Manager Server. Only after you distribute the parcel to a cluster will copies of it be placed on the hosts in that cluster.

Downloading a Parcel

1. Click the parcel indicator in the top navigation bar. This takes you to the **Hosts** page, **Parcels** tab. By default, any parcels available for download are shown in the **Available Remotely** section of the Parcels page. Parcels available for download will display a **Download** button.

   If the parcel you want is not shown here — for example, you want to upgrade to version of CDH that is not the most current version — you can make additional remote parcel repositories available through the Administration Settings page. You can also configure the location of the local parcel repository and other settings. See Parcel Configuration Settings on page 84.

2. Click **Download** to initiate the download of the parcel from the remote parcel repository to your local repository.

When the parcel has been downloaded, the button label changes to Distribute.

- **Note:** The parcel download is done at the Cloudera Manager Server, so with multiple clusters, the downloaded parcels are shown as available to *all* clusters managed by the Cloudera Manager Server. However, distribution (to a specific cluster's member hosts) must be selected on a cluster-by-cluster basis.

Distributing a Parcel

Parcels that have been downloaded can be distributed to the hosts in your cluster, available for activation.

From the Parcels tab, click the **Distribute** button for the parcel you want to distribute. This starts the distribution process to the hosts in the cluster.

Distribution does not require Internet access; rather the Cloudera Manager Agent on each cluster member downloads the parcel from the local parcel repository hosted on the Cloudera Manager Server.

If you have a large number of hosts to which the parcels should be distributed, you can control how many concurrent uploads Cloudera Manager will perform. You can configure this setting on the **Administration** page, **Properties** tab under the Parcels section.

You can delete a parcel that is ready to be distributed; click the triangle at the right end of the Distribute button to access the Delete command. This will delete the downloaded parcel from the local parcel repository.

Distributing parcels to the hosts in the cluster does not affect the current running services.

Activating a Parcel

Parcels that have been distributed to the hosts in a cluster are ready to be activated.

1. From the Parcels tab, click the **Activate** button for the parcel you want to activate. This will update Cloudera Manager to point to the new software, ready to be run the next time a service is restarted.
2. A pop-up warns you that your currently running process will not be affected until you restart, and gives you the option to perform a restart. If you do not want to restart at this time, click **Close**.

If you elect not to restart services as part of the Activation process, you can instead go to the **Clusters** tab and restart your services at a later time. Until you restart services, the current software will continue to run. This allows you to restart your services at a time that is convenient based on your maintenance schedules or other considerations.

Activating a new parcel also deactivates the previously active parcel (if any) for the product you've just upgraded. However, until you restart the services, the previously active parcel will have the link **Still in use** and you will not be able to remove the parcel until it is no longer being used.

- **Note:** Under some situations, such as doing a major release upgrade (for example, CDH 4 to CDH 5) additional upgrade steps may be necessary. In this case, instead of **Activate**, the button may instead say **Upgrade**. This indicates that there may be additional steps involved in the upgrade.

Deactivating a Parcel

You can deactivate an active parcel; this will update Cloudera Manager to point to the previous software version, ready to be run the next time a service is restarted. To deactivate a parcel, click **Actions** on an activated parcel and select **Deactivate**.

To use the previous version of the software, go to the **Clusters** tab and restart your services.

- **Note:** If you did your original installation from parcels, and there is only one version of your software installed (that is, no packages, and no previous parcels have been activated and started) then when you attempt to restart after deactivating the current version, your roles will be stopped but will not be able to restart.

Removing a Parcel

To remove a parcel, click the down arrow to the right of an **Activate** button and select **Remove from Hosts**.

Deleting a Parcel

To delete a parcel, click the down arrow to the right of a **Distribute** button and select **Delete**.

Troubleshooting

If you experience an error while performing parcel operations, click on the red 'X' icons on the parcel page to display a message that will identify the source of the error.

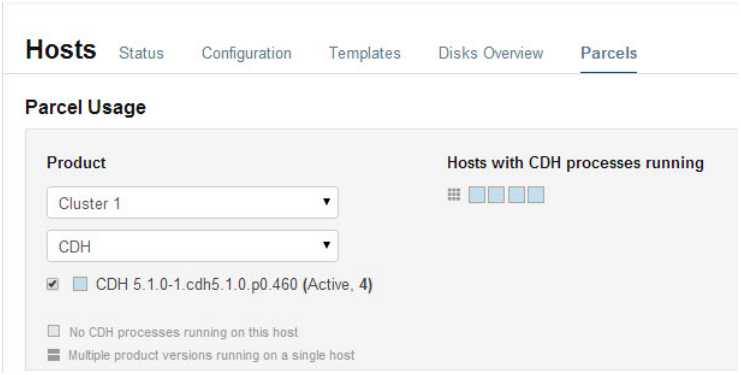
If you have a parcel distributing but never completing, make sure you have enough free space in the parcel download directories, as Cloudera Manager will retry to downloading and unpacking parcels even if there is insufficient space.

Viewing Parcel Usage

The **Parcel Usage** page shows you which parcels are in current use in your clusters. This is particularly useful in a large deployment where it may be difficult to keep track of what versions are installed across the cluster, especially if some hosts were not available when you performed an installation or upgrade, or were added later. To display the Parcel Usage page:

1. Do one of the following:
   - Click  in the top navigation bar
   - Click **Hosts** in the top navigation bar and click the **Parcels** tab.
2. Click the **Parcel Usage** button.

This page only shows the usage of parcels, not components that were installed as packages. If you select a cluster running packages (for example, a CDH 4 cluster) the cluster is not displayed, and instead you will see a message indicating the cluster is not running parcels. If you have individual hosts running components installed as packages, they will appear as "empty."



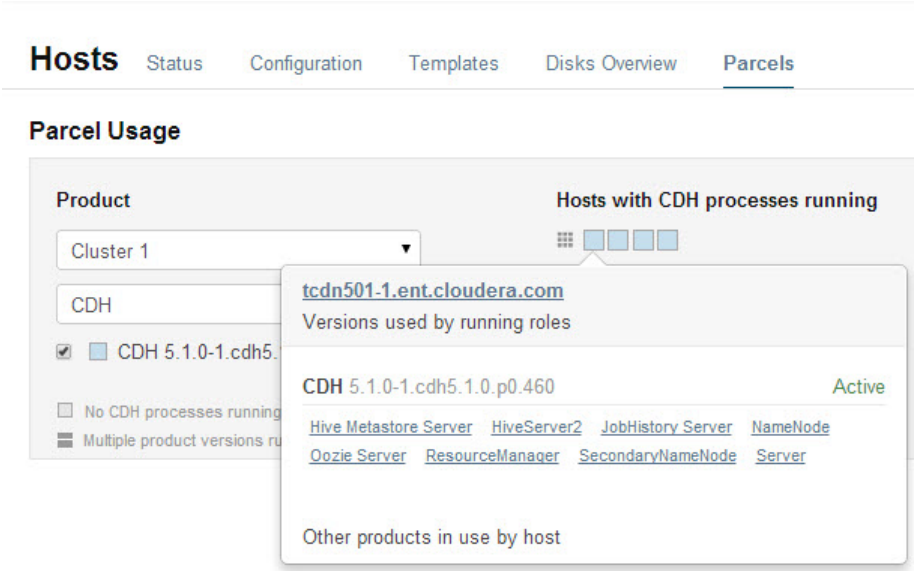
You can view parcel usage by cluster, or by product (CDH, SOLR, IMPALA, SPARK, or ACCUMULO).

You can also view just the hosts running only the active parcels, or just hosts running older parcels (not the currently active parcels) or both.

The "host map" at the right shows each host in the cluster with the status of the parcels on that host. If the host is actually running the processes from the currently activated parcels, the host is indicated in blue. A black square indicates that a parcel has been activated, but that all the running processes are from an earlier version of the software. This can happen, for example, if you have not restarted a service or role after activating a new parcel.

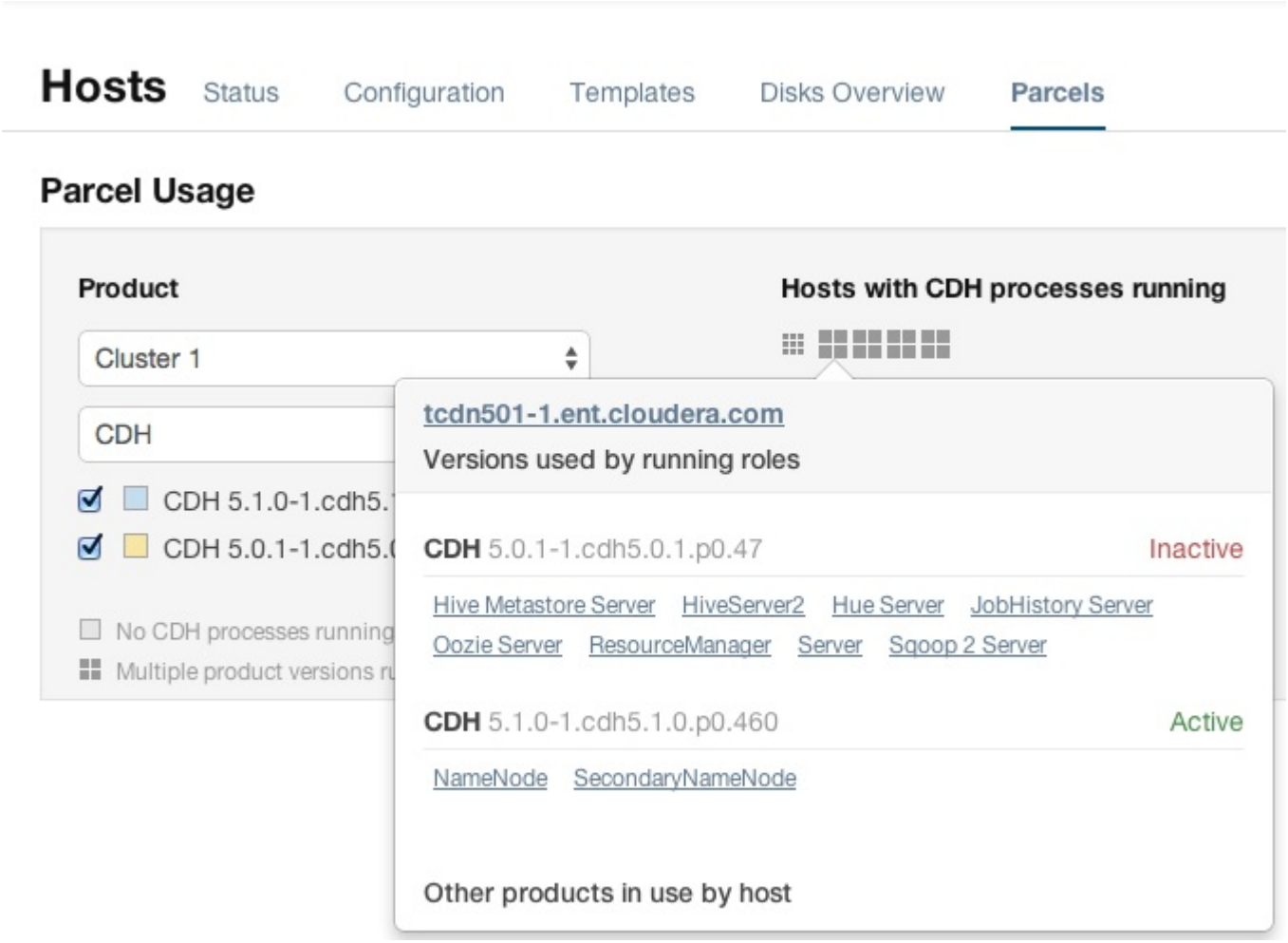
Move the cursor over the  icon to see the rack to which the hosts are assigned. Hosts on different racks are displayed in separate rows.

To view the exact versions of the software running on a given host, you can click on the square representing the host. This pops up a display showing the parcel versions installed on that host.



For CDH 4.4, Impala 1.1.1, and Solr 0.9.3 or later, it will list the roles running on the selected host that are part of the listed parcel. Clicking a role takes you to the Cloudera Manager page for that role. It also shows whether the parcel is Active or not.

If a host is running a mix of software versions, the square representing the host is shown by a four-square icon ⊞. When you move the cursor over that host, both the active and inactive components are shown. For example, in the image below the older CDH parcel has been deactivated but only the HDFS service has been restarted.



Parcel Configuration Settings

Configuring Server Parcel Settings

1. Do one of the following to open the parcel settings page:
   - 1. Click in the top navigation bar
     2. Click the **Edit Settings** button.
   - 1. Select **Administration** > **Settings**.
     2. Click the **Parcels** category.
   - 1. Click the **Hosts** tab.
     2. Click the **Configuration** tab.
     3. Click the **Parcels** category.
     4. Click the **Edit Settings** button.
2. Specify a property:
   - **Local Parcel Repository Path** defines the path on the Cloudera Manager Server host where downloaded parcels are stored.
   - **Remote Parcel Repository URLs** is a list of repositories that Cloudera Manager should check for parcels. Initially this points to the latest released CDH 4, CDH 5, Impala, and Solr repositories but you can add your own repository locations to the list. You can use this mechanism to add Cloudera repositories that are

not listed by default, such as older versions of CDH, or the Sentry parcel for CDH 4.3. You can also use this to add your own custom repositories. The locations of the Cloudera parcel repositories are `http://archive.cloudera.com/product/parcels/version`, where *product* is cdh4, cdh5, gplextras5, impala, search, and sentry, and *version* is a specific product version or `latest`.

To add a parcel repository:

1. In the **Remote Parcel Repository URLs** list, click ✚ to open an additional row.
2. Enter the path to the repository.

3. Click **Save Changes**.

You can also:

- Set the frequency with which Cloudera Manager will check for new parcels.
- Configure a proxy to access to the remote repositories.
- Configure whether downloads and distribution of parcels should occur automatically whenever new ones are detected. If automatic downloading/distribution are not enabled (the default), you must go to the **Parcels** page to initiate these actions.
- Control which products can be downloaded if automatic downloading is enabled.
- Control whether to retain downloaded parcels.
- Control whether to retain old parcel version and how many parcel versions to retain

You can configure the bandwidth limits and the number of concurrent uploads, to tune the load that parcel distribution puts on your network. The defaults are up to 50 concurrent parcel uploads and 50 MiB/s aggregate bandwidth.

- The concurrent upload count (**Maximum Parcel Uploads**) doesn't matter, theoretically, if all hosts have the same speed Ethernet. In general, 50 concurrent uploads is an acceptable setting in most cases. However, in a scenario where the server has more bandwidth (say 10Gbe while the normal hosts are using 1Gbe), then the count is important to maximize bandwidth, and would need to be at least the difference in speeds (10x in this case).
- The bandwidth limit (**Parcel Distribution Rate Limit**) should be your Ethernet speed (in MiB/seconds) divided by approximately 16. You can use a higher limit if you have QoS set up to prevent starving other services, or if you are willing accept a higher risk of higher bandwidth load.

Configuring a Proxy Server

To configure a proxy server through which parcels are downloaded, follow the instructions in Configuring Network Settings.

Configuring the Host Parcel Directory

To configure the location of distributed parcels:

1. Click **Hosts** in the top navigation bar.
2. Click the **Configuration** tab.
3. Configure the value of the **Parcel Directory** property. The setting of the `parcel_dir` property in the Cloudera Manager Agent configuration file overrides this setting.
4. Click **Save Changes** to commit the changes.

Migrating from Packages to Parcels

Required Role: Administrator

Managing software distribution using parcels offers many advantages over packages. To migrate from packages to the *same version* parcel, perform the following steps. To upgrade to a different version, see Upgrading CDH and Managed Services.

Download, Distribute, and Activate Parcels

1. In the Cloudera Manager Admin Console, click the Parcels indicator in the top navigation bar (or).
2. Click **Download** for the version that matches the CDH or service version of the currently installed packages. If the parcel you want is not shown here—for example, if you want to use a version of CDH that is not the most current version—you can add parcel repositories through the Parcel Configuration Settings on page 84 page:
   - **CDH 4**
     - CDH - `http://archive.cloudera.com/cdh4/parcels/`
     - Impala - `http://archive.cloudera.com/impala/parcels/`
     - Search `http://archive.cloudera.com/search/parcels/`
     - Spark - `http://archive.cloudera.com/spark/parcels/`
     - GPL Extras - `http://archive.cloudera.com/gplextras/parcels/`
   - **CDH 5** - Impala, Spark, and Search are included in the CDH parcel.
     - CDH - `http://archive.cloudera.com/cdh5/parcels/`
     - GPL Extras - `http://archive.cloudera.com/gplextras5/parcels/`
   - **Other services**
     - Accumulo - `http://archive.cloudera.com/accumulo/parcels/`
     - Sqoop connectors - `http://archive.cloudera.com/sqoop-connectors/parcels/`

   If your Cloudera Manager Server does not have Internet access, you can obtain the required parcel file(s) and put them into a repository. See Creating and Using a Parcel Repository on page 97 for more details.
3. When the download has completed, click **Distribute** for the version you downloaded.
4. When the parcel has been distributed and unpacked, the button will change to say **Activate**.
5. Click **Activate**.

Restart the Cluster and Deploy Client Configuration

1. Restart the cluster:
   a. On the Home page, click to the right of the cluster name and select **Restart**.
   b. Click **Restart** that appears in the next screen to confirm. The **Command Details** window shows the progress of stopping services.

      When **All services successfully started** appears, the task is complete and you can close the **Command Details** window.

   You can optionally perform a rolling restart.
2. Redeploy client configurations:
   a. On the Home page, click to the right of the cluster name and select **Deploy Client Configuration**.
   b. Click **Deploy Client Configuration**.

Uninstall Packages

1. Uninstall the CDH packages on each host:
   - **Not including Impala and Search**

Operating System	Command
RHEL	`$ sudo yum remove bigtop-utils bigtop-jsvc bigtop-tomcat hue-common sqoop2-client`
SLES	`$ sudo zypper remove bigtop-utils bigtop-jsvc bigtop-tomcat hue-common sqoop2-client`
Ubuntu or Debian	`$ sudo apt-get purge bigtop-utils bigtop-jsvc bigtop-tomcat hue-common sqoop2-client`

- **Including Impala and Search**

Operating System	Command
RHEL	`$ sudo yum remove 'bigtop-*' hue-common impala-shell solr-server sqoop2-client`
SLES	`$ sudo zypper remove 'bigtop-*' hue-common impala-shell solr-server sqoop2-client`
Ubuntu or Debian	`$ sudo apt-get purge 'bigtop-*' hue-common impala-shell solr-server sqoop2-client`

Restart Cloudera Manager Agents

Restart all the Cloudera Manager Agents to force an update of the symlinks to point to the newly installed components. On each host run:

```
$ sudo service cloudera-scm-agent restart
```

Update Applications to Reference Parcel Paths

With parcel software distribution, the path to the CDH libraries is `/opt/cloudera/parcels/CDH/lib` instead of the usual `/usr/lib`. You should not link `/usr/lib/` elements to parcel deployed paths, as such links may confuse scripts that distinguish between the two paths. Instead you should update your applications to reference the new library locations.

Migrating from Parcels to Packages

Required Role: Administrator

To migrate from a parcel to the *same version* packages, perform the following steps. To upgrade to a different version, see Upgrading CDH and Managed Services.

Install Packages

Install CDH and Managed Service Packages

For more information about manually installing CDH packages, see CDH 4 Installation Guide or CDH 5 Installation Guide.

1. Choose a repository strategy:
   - Standard Cloudera repositories. For this method, ensure you have added the required repository information to your systems.
   - Internally hosted repositories. You might use internal repositories for environments where hosts do not have access to the Internet. In such a case, ensure your environment is properly prepared. For more information, see Understanding Custom Installation Solutions on page 95.

2. Install packages:

<table>
<tr><th>CDH Version</th><th>Procedure</th></tr>
<tr><td>CDH 5</td><td>

- **Red Hat**

1. Download and install the "1-click Install" package

a. Download the CDH 5 "1-click Install" package.

Click the entry in the table below that matches your Red Hat or CentOS system, choose **Save File**, and save the file to a directory to which you have write access (it can be your home directory).

<table>
<tr><th>OS Version</th><th>Click this Link</th></tr>
<tr><td>Red Hat/CentOS/Oracle 5</td><td>Red Hat/CentOS/Oracle 5 link</td></tr>
<tr><td>Red Hat/CentOS/Oracle 6</td><td>Red Hat/CentOS/Oracle 6 link</td></tr>
</table>

b. Install the RPM:

- **Red Hat/CentOS/Oracle 5**

```
$ sudo yum --nogpgcheck localinstall cloudera-cdh-5-0.x86_64.rpm
```

- **Red Hat/CentOS/Oracle 6**

```
$ sudo yum --nogpgcheck localinstall cloudera-cdh-5-0.x86_64.rpm
```

2. (Optionally) add a repository key:

- **Red Hat/CentOS/Oracle 5**

```
$ sudo rpm --import
http://archive.cloudera.com/cdh5/redhat/5/x86_64/cdh/RPM-GPG-KEY-cloudera
```

- **Red Hat/CentOS/Oracle 6**

```
$ sudo rpm --import
http://archive.cloudera.com/cdh5/redhat/6/x86_64/cdh/RPM-GPG-KEY-cloudera
```

3. Install the CDH packages:

```
$ sudo yum clean all
$ sudo yum install avro-tools crunch flume-ng hadoop-hdfs-fuse
hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat
hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms
hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell
kite llama mahout oozie pig pig-udf-datafu search sentry
solr-mapreduce spark-python sqoop sqoop2 whirr
```

Note: Installing these packages will also install all the other CDH packages that are needed for a full CDH 5 installation.

</td></tr>
</table>

<table>
<tr><th>CDH Version</th><th>Procedure</th></tr>
<tr><td></td><td>

- SLES

1. Download and install the "1-click Install" package.

   a. Download the CDH 5 "1-click Install" package.

   Click this link, choose **Save File**, and save it to a directory to which you have write access (it can be your home directory).

   b. Install the RPM:

```
$ sudo rpm -i cloudera-cdh-5-0.x86_64.rpm
```

   c. Update your system package index by running:

```
$ sudo zypper refresh
```

2. (Optionally) add a repository key:

```
$ sudo rpm --import
http://archive.cloudera.com/cdh5/sles/11/x86_64/cdh/RPM-GPG-KEY-cloudera
```

3. Install the CDH packages:

```
$ sudo zypper clean --all
$ sudo zypper install avro-tools crunch flume-ng hadoop-hdfs-fuse
hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat
hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms
hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell
kite llama mahout oozie pig pig-udf-datafu search sentry
solr-mapreduce spark-python sqoop sqoop2 whirr
```

   - **Note:** Installing these packages will also install all the other CDH packages that are needed for a full CDH 5 installation.

- **Ubuntu and Debian**

1. Download and install the "1-click Install" package

   a. Download the CDH 5 "1-click Install" package:

OS Version	Click this Link
Wheezy	Wheezy link
Precise	Precise link

   b. Install the package. Do one of the following:

   - Choose **Open with** in the download window to use the package manager.
   - Choose **Save File**, save the package to a directory to which you have write access (it can be your home directory) and install it from the command line, for example:

```
sudo dpkg -i cdh5-repository_1.0_all.deb
```

2. (Optionally) add a repository key:

</td></tr>
</table>

<table>
<tr><th>CDH Version</th><th>Procedure</th></tr>
<tr><td></td><td>

Debian Wheezy
<pre>$ curl -s
http://archive.cloudera.com/cdh5/debian/wheezy/amd64/cdh/archive.key
 | sudo apt-key add -</pre>
Ubuntu Precise
<pre>$ curl -s
http://archive.cloudera.com/cdh5/ubuntu/precise/amd64/cdh/archive.key
 | sudo apt-key add -</pre>

3. Install the CDH packages:
<pre>$ sudo apt-get update
$ sudo apt-get install avro-tools crunch flume-ng hadoop-hdfs-fuse
hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat
hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms
hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell
kite llama mahout oozie pig pig-udf-datafu search sentry
solr-mapreduce spark-python sqoop sqoop2 whirr</pre>

Note: Installing these packages will also install all the other CDH packages that are needed for a full CDH 5 installation.

</td></tr>
<tr><td>CDH 4, Impala, and Solr</td><td>

Red Hat-compatible

1. Click the entry in the table at CDH Download Information that matches your Red Hat or CentOS system.
2. Navigate to the repo file (<tt>cloudera-cdh4.repo</tt>) for your system and save it in the <tt>/etc/yum.repos.d/</tt> directory.
3. Optionally add a repository key:

Red Hat/CentOS/Oracle 5
<pre>$ sudo rpm --import
http://archive.cloudera.com/cdh4/redhat/5/x86_64/cdh/RPM-GPG-KEY-cloudera</pre>
Red Hat/CentOS 6
<pre>$ sudo rpm --import
http://archive.cloudera.com/cdh4/redhat/6/x86_64/cdh/RPM-GPG-KEY-cloudera</pre>

4. Install packages on every host in your cluster:
a. Install CDH 4 packages:
<pre>$ sudo yum -y install bigtop-utils bigtop-jsvc bigtop-tomcat hadoop
 hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn
hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie
 oozie-client pig zookeeper</pre>
b. To install the <tt>hue-common</tt> package and all Hue applications on the Hue host, install the hue meta-package:
<pre>$ sudo yum install hue</pre>
</td></tr>
</table>

CDH Version	Procedure

	5. (**Requires CDH 4.2 or later**) Install Impala
	a. Click the entry in the table at Cloudera Impala Version and Download Information that matches your Red Hat or CentOS system.
	b. Navigate to the repo file for your system and save it in the `/etc/yum.repos.d/` directory.
	c. Install Impala and the Impala Shell on Impala machines:
	`$ sudo yum -y install impala impala-shell`
	6. (**Requires CDH 4.3 or later**) Install Search
	a. Click the entry in the table at Cloudera Search Version and Download Information that matches your Red Hat or CentOS system.
	b. Navigate to the repo file for your system and save it in the `/etc/yum.repos.d/` directory.
	c. Install the Solr Server on machines where you want Cloudera Search.
	`$ sudo yum -y install solr-server`
	▪ SLES
	1. Run the following command:
	`$ sudo zypper addrepo -f http://archive.cloudera.com/cdh4/sles/11/x86_64/cdh/cloudera-cdh4.repo`
	2. Update your system package index by running:
	`$ sudo zypper refresh`
	3. Optionally add a repository key:
	`$ sudo rpm --import http://archive.cloudera.com/cdh4/sles/11/x86_64/cdh/RPM-GPG-KEY-cloudera`
	4. Install packages on every host in your cluster:
	a. Install CDH 4 packages:
	`$ sudo zypper install bigtop-utils bigtop-jsvc bigtop-tomcat hadoop hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie oozie-client pig zookeeper`
	b. To install the hue-common package and all Hue applications on the Hue host, install the hue meta-package:
	`$ sudo zypper install hue`
	c. (**Requires CDH 4.2 or later**) Install Impala

CDH Version	Procedure

a. Run the following command:

```
$ sudo zypper addrepo -f
http://archive.cloudera.com/impala/sles/11/x86_64/impala/cloudera-impala.repo
```

b. Install Impala and the Impala Shell on Impala machines:

```
$ sudo zypper install impala impala-shell
```

d. (**Requires CDH 4.3 or later**) Install Search

a. Run the following command:

```
$ sudo zypper addrepo -f
http://archive.cloudera.com/search/sles/11/x86_64/search/cloudera-search.repo
```

b. Install the Solr Server on machines where you want Cloudera Search.

```
$ sudo zypper install solr-server
```

- **Ubuntu or Debian**

1. Click the entry in the table at CDH Version and Packaging Information that matches your Ubuntu or Debian system.
2. Navigate to the list file (`cloudera.list`) for your system and save it in the `/etc/apt/sources.list.d/` directory. For example, to install CDH 4 for 64-bit Ubuntu Lucid, your `cloudera.list` file should look like:

```
deb [arch=amd64]
http://archive.cloudera.com/cdh4/ubuntu/lucid/amd64/cdh lucid-cdh4
contrib
deb-src http://archive.cloudera.com/cdh4/ubuntu/lucid/amd64/cdh
lucid-cdh4 contrib
```

3. Optionally add a repository key:
   - **Ubuntu Lucid**

```
$ curl -s
http://archive.cloudera.com/cdh4/ubuntu/lucid/amd64/cdh/archive.key
  | sudo apt-key add -
```

   - **Ubuntu Precise**

```
$ curl -s
http://archive.cloudera.com/cdh4/ubuntu/precise/amd64/cdh/archive.key
  | sudo apt-key add -
```

   - **Debian Squeeze**

```
$ curl -s
http://archive.cloudera.com/cdh4/debian/squeeze/amd64/cdh/archive.key
  | sudo apt-key add -
```

4. Install packages on every host in your cluster:

CDH Version	Procedure

	a. Install CDH 4 packages:
	`$ sudo apt-get install bigtop-utils bigtop-jsvc bigtop-tomcat hadoop hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie oozie-client pig zookeeper`
	b. To install the hue-common package and all Hue applications on the Hue host, install the hue meta-package:
	`$ sudo apt-get install hue`
	c. (**Requires CDH 4.2 or later**) Install Impala
	a. Click the entry in the table at Cloudera Impala Version and Download Information and that matches your Ubuntu or Debian system.
	b. Navigate to the list file for your system and save it in the `/etc/apt/sources.list.d/` directory.
	c. Install Impala and the Impala Shell on Impala machines:
	`$ sudo apt-get install impala impala-shell`
	d. (**Requires CDH 4.3 or later**) Install Search
	a. Click the entry in the table at Cloudera Search Version and Download Information that matches your Ubuntu or Debian system.
	b. Install Solr Server on machines where you want Cloudera Search:
	`$ sudo apt-get install solr-server`

Deactivate Parcels

When you deactivate a parcel, Cloudera Manager points to the installed packages, ready to be run the next time a service is restarted. To deactivate parcels,

1. Go to the Parcels page by doing one of the following:
   - Clicking the parcel indicator in the Admin Console navigation bar ( )
   - Clicking the **Hosts** in the top navigation bar, then the **Parcels** tab.
2. Click **Actions** on the activated CDH and managed service parcels and select **Deactivate**.

Restart the Cluster

1. On the Home page, click to the right of the cluster name and select **Restart**.
2. Click **Restart** that appears in the next screen to confirm. The **Command Details** window shows the progress of stopping services.

   When **All services successfully started** appears, the task is complete and you can close the **Command Details** window.

You can optionally perform a rolling restart.

Remove and Delete Parcels

Removing a Parcel

To remove a parcel, click the down arrow to the right of an **Activate** button and select **Remove from Hosts**.

Deleting a Parcel

To delete a parcel, click the down arrow to the right of a **Distribute** button and select **Delete**.

Understanding Custom Installation Solutions

Cloudera hosts two types of software repositories that you can use to install products such as Cloudera Manager or CDH—parcel repositories and RHEL and SLES RPM and Debian/Ubuntu package repositories.

These repositories are effective solutions in most cases, but custom installation solutions are sometimes required. Using the software repositories requires client access over the Internet and results in the installation of the latest version of products. An alternate solution is required if:

- You need to install older product versions. For example, in a CDH cluster, all hosts must run the same CDH 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 are unable to reach the Cloudera repository. (For a parcel installation, only the Cloudera Manager Server needs Internet access, but for a package installation, all cluster members need access to the Cloudera repository). Some organizations choose to partition parts of their network from outside access. Isolating segments of a network can provide greater assurance that valuable data is not compromised by individuals out of maliciousness or for personal gain. In such a case, the isolated computers are unable to access Cloudera repositories for new installations or upgrades.

In both of these cases, using a custom repository solution allows you to meet the needs of your organization, whether that means installing older versions of Cloudera software or installing any version of Cloudera software on hosts that are disconnected from the Internet.

Understanding 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. Parcels are available for CDH 4.1.3 and onwards. For detailed information about parcels, see Parcels on page 77.

If your Cloudera Manager Server does not have Internet access, you can obtain the required parcel files and put them into a parcel repository. See Creating and Using a Parcel Repository on page 97.

Understanding Package Management

Before getting into the details of how to configure a custom package management solution in your environment, it can be useful to have more information about:

- Package management tools
- Package repositories

Package Management Tools

Packages (`rpm` or deb files) help ensure that installations complete successfully by encoding each package's dependencies. That means that if you request the installation of a solution, all required elements can be 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` (Debian/Ubuntu) are tools that can find and install any required packages. For example, for RHEL, you might enter `yum install hadoop-0.20-hive`. `yum` would inform you that the hive package requires `hadoop-0.20` and offers to complete that installation for you. `zypper` and `apt-get` provide similar functionality.

Package Repositories

Package management tools operate on package repositories.

Repository Configuration Files

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

- RedHat/CentOS `yum` - `/etc/yum.repos.d`
- SLES `zypper` - `/etc/zypp/zypper.conf`
- Debian/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:

```
[user@localhost ~]$ ls -l /etc/yum.repos.d/
total 24
-rw-r--r-- 1 root root 2245 Apr 25  2010 CentOS-Base.repo
-rw-r--r-- 1 root root  626 Apr 25  2010 CentOS-Media.repo
```

The `.repo` files contain pointers to one or many repositories. There are similar pointers inside configuration files for `zypper` and `apt-get`. In the following snippet 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=$basearch&repo=os
#baseurl=http://mirror.centos.org/centos/$releasever/os/$basearch/
gpgcheck=1
gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-CentOS-5

#released updates
[updates]
name=CentOS-$releasever - Updates
mirrorlist=http://mirrorlist.centos.org/?release=$releasever&arch=$basearch&repo=updates
#baseurl=http://mirror.centos.org/centos/$releasever/updates/$basearch/
gpgcheck=1
gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-CentOS-5
# ...
```

Listing Repositories

You can list the repositories you have enabled. The command varies according to operating system:

- RedHat/CentOS - `yum repolist`
- SLES - `zypper repos`
- Debian/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 what you might find on a CentOS system in `repolist`:

```
[root@localhost yum.repos.d]$ yum repolist
Loaded plugins: fastestmirror
Loading mirror speeds from cached hostfile
 * addons: mirror.san.fastserv.com
 * base: centos.eecs.wsu.edu
 * extras: mirrors.ecvps.com
 * updates: mirror.5ninesolutions.com
repo id                       repo name                                   status
addons                        CentOS-5 - Addons                           enabled:
```

```
 0
base                                CentOS-5 - Base                                enabled: 3,434
extras                               CentOS-5 - Extras                               enabled:
296
updates                             CentOS-5 - Updates                             enabled: 1,137
repolist: 4,867
```

Creating and Using a Parcel Repository

This topic describes how to create a repository and then how to direct hosts in your environment to use that repository. To create a repository, you simply put the parcel files you want to host in one directory. Then publish the resulting repository on a website.

Install a Web Server

The repository is typically hosted using HTTP on a host inside your network. If you already have a web server in your organization, you can move the repository directory, which will include both the RPMs and the `repodata/` subdirectory, to a location hosted by the web server. An easy web server to install is the Apache HTTPD. If you are able to use an existing web server, then note the URL and skip to Download Parcel and Publish Files on page 97.

Installing Apache HTTPD

You may need to respond to some prompts to confirm you want to complete the installation.

OS	Command
RHEL	`[root@localhost yum.repos.d]$ yum install httpd`
SLES	`[root@localhost zypp]$ zypper install httpd`
Ubuntu or Debian	`[root@localhost apt]$ apt-get install httpd`

Starting Apache HTTPD

OS	Command

RHEL	`[root@localhost tmp]$  service httpd start`
	`Starting httpd:  [  OK`
	`]`

SLES	`[root@localhost tmp]$ service apache2 start`
	`Starting httpd:  [  OK`
	`]`

Ubuntu or Debian	`[root@localhost tmp]$ service apache2 start`
	`Starting httpd:  [  OK`
	`]`

Download Parcel and Publish Files

1. Download the parcel and `manifest.json` files for your OS distribution from
   - **CDH 4**
     - CDH - `http://archive.cloudera.com/cdh4/parcels/`
     - Impala - `http://archive.cloudera.com/impala/parcels/`
     - Search `http://archive.cloudera.com/search/parcels/`
     - Spark - `http://archive.cloudera.com/spark/parcels/`
     - GPL Extras - `http://archive.cloudera.com/gplextras/parcels/`

- **CDH 5** - Impala, Spark, and Search are included in the CDH parcel.
  - CDH - `http://archive.cloudera.com/cdh5/parcels/`
  - GPL Extras - `http://archive.cloudera.com/gplextras5/parcels/`
- **Other services**
  - Accumulo - `http://archive.cloudera.com/accumulo/parcels/`
  - Sqoop connectors - `http://archive.cloudera.com/sqoop-connectors/parcels/`

2. Move the parcel and `manifest.json` files to the web server directory, and modify file permissions. For example, you might use the following commands:

```
[root@localhost tmp]$ mkdir /var/www/html/cdh4.6
[root@localhost tmp]$ mv CDH-4.6.0-1.cdh4.6.0.p0.26-lucid.parcel
/var/www/html/cdh4.6
[root@localhost tmp]$ mv manifest.json /var/www/html/cdh4.6
[root@localhost tmp]$ chmod -R ugo+rX /var/www/html/cdh4.6
```

After moving the files and changing permissions, visit `http://hostname:80/cdh4.6/` to verify that you can access the parcel. Apache may have been configured to not show indexes, which is also acceptable.

Configure the Cloudera Manager Server to Use the Parcel URL

1. Do one of the following to open the parcel settings page:
   - 1. Click in the top navigation bar
     2. Click the **Edit Settings** button.
   - 1. Select **Administration** > **Settings**.
     2. Click the **Parcels** category.
   - 1. Click the **Hosts** tab.
     2. Click the **Configuration** tab.
     3. Click the **Parcels** category.
     4. Click the **Edit Settings** button.
2. In the **Remote Parcel Repository URLs** list, click + to open an additional row.
3. Enter the path to the parcel. For example, `http://hostname:80/cdh4.6/`.
4. Click **Save Changes** to commit the changes.

Creating and Using a Package Repository

This topic describes how to create a package repository and then how to direct hosts in your environment to use that repository. To create a repository, you simply put the repo files you want to host in one directory. Then publish the resulting repository on a website.

Install a Web Server

The repository is typically hosted using HTTP on a host inside your network. If you already have a web server in your organization, you can move the repository directory, which will include both the RPMs and the `repodata/` subdirectory, to some a location hosted by the web server. An easy web server to install is the Apache HTTPD. If you are able to use an existing web server, then note the URL and skip to Download Tarball and Publish Repository Files on page 99.

Installing Apache HTTPD

You may need to respond to some prompts to confirm you want to complete the installation.

OS	Command
RHEL	`[root@localhost yum.repos.d]$ yum install httpd`
SLES	`[root@localhost zypp]$ zypper install httpd`
Ubuntu or Debian	`[root@localhost apt]$ apt-get install httpd`

Starting Apache HTTPD

OS	Command
RHEL	`[root@localhost tmp]$  service httpd start` `Starting httpd:  [  OK  ]`
SLES	`[root@localhost tmp]$ service apache2 start` `Starting httpd:  [  OK  ]`
Ubuntu or Debian	`[root@localhost tmp]$ service apache2 start` `Starting httpd:  [  OK  ]`

Download Tarball and Publish Repository Files

1. Download the tarball for your OS distribution from the repo as tarball archive.
2. Unpack the tarball, move the files to the web server directory, and modify file permissions. For example, you might use the following commands:

```
[root@localhost tmp]$ gunzip cm5.0.0-centos6.tar.gz
[root@localhost tmp]$ tar xvf cm5.0.0-centos6.tar
[root@localhost tmp]$ mv cm /var/www/html
[root@localhost tmp]$ chmod -R ugo+rX /var/www/html/cm
```

After moving files and changing permissions, visit `http://<hostname>:80/cm` to verify that you see an index of files. Apache may have been configured to not show indexes, which is also acceptable.

Modify Clients to Find Repository

Having established the repository, modify the clients so they find the repository.

OS	Command

RHEL	Create files on client systems with the following information and format, where *hostname* is the name of the web server you created in Install a Web Server on page 98:
	`[myrepo]`
	`name=myrepo`
	`baseurl=http://hostname/cm/5`
	`enabled=1`
	`gpgcheck=0`
	See `man yum.conf` for more details. Put that file into `/etc/yum.repos.d/myrepo.repo` on all of your hosts to enable them to find the packages that you are hosting.

SLES	Use the `zypper` utility to update client system repo information by issuing the following command: `$ zypper addrepo http://hostname/cm alias`
Ubuntu or Debian	Add a new `list` file to `/etc/apt/sources.list.d/` on client systems. For example, you might create the file

OS	Command

	`/etc/apt/sources.list.d/my-private-cloudera-repo.list`. In that file, create an entry to your newly created repository. For example:
	`$ cat /etc/apt/sources.list.d/my-private-cloudera-repo.list`
	`deb http://hostname/cm cloudera`
	After adding your `.list` file, ensure `apt-get` uses the latest information by issuing the following command:
	`$ sudo apt-get update`

After completing these steps, you have established the environment necessary to install a previous version of Cloudera Manager or install Cloudera Manager to hosts that are not connected to the Internet. Proceed with the installation process, being sure to target the newly created repository with your package management tool.

Installing Cloudera Manager and CDH on EC2

The following procedure leads you through setting up Cloudera Manager and CDH on a cluster of Amazon Web Services (AWS) EC2 instances.

- The Cloudera Manager installation wizard launches the EC2 version of the wizard when Cloudera Manager is started on EC2.
- The resulting installation uses an embedded PostgreSQL database; there is no option for setting up other databases.
- This wizard installs and starts all the latest Cloudera Manager-managed CDH services.

> **Note:**
> - The EC2 version of the wizard does not support Amazon Virtual Private Cloud (Amazon VPC).
> - This setup is not recommended for production use.

Step 1: Set up an AWS EC2 instance for the Cloudera Manager Server.

> **Note:** The instance on which you install the Cloudera Manager Server must conform to the requirements described in Networking and Security Requirements on page 17. In particular, SELinux and iptables must be disabled.

1. Log into the AWS console.
2. Go to **EC2**.
3. Create a security group:
   a. In the left pane, click **Security Groups**.
   b. Click **Create Security Group**.
   c. When prompted, enter a name and description, and click **OK**.
   d. Select the group you created in the list of groups.
   e. In the bottom panel, go to the **Inbound** tab.
   f. Authorize TCP ports 22, 7180, 7182, 7183, and 7432.
   g. Authorize ICMP Echo Reply.
4. Create (or import) an SSH key pair:
   a. In the left pane, click **Key Pairs**.
   b. Click **Create Key Pair**.
   c. When prompted, enter a key pair name and click OK

d. Your private key *keypair-name*.pem will be downloaded automatically. AWS does not store the private keys – if you lose this file, you won't be able to SSH into instances you provision with this key pair.

5. Launch an EC2 instance:

   a. In the left pane, click **Instances**.
   b. Click **Launch Instance**.
   c. Select the Ubuntu 12.04 AMI 64-bit or other operating system supported by Cloudera Manager. See Cloudera Manager Requirements on page 15.
   d. Choose the Instance Type. Cloudera recommends using at least **General purpose** > **m1.large** instances.
   e. In the Configure Security Group tab, use the security group and key pair you prepared in the previous steps.
   f. Look at the instance details, and copy the public hostname.
   g. SSH into the instance:

```
$ ssh -i private-key-file username@ec2-xx-xx-xx-xx.compute-1.amazonaws.com
```

   The *username* is usually "ubuntu" on Ubuntu systems, and "ec2-user" on most other Linux images on EC2.

   h. Download the Cloudera Manager installer:

```
$ wget
http://archive.cloudera.com/cm5/installer/latest/cloudera-manager-installer.bin
```

   i. Execute the installer:

```
$ sudo su
$ chmod +x cloudera-manager-installer.bin
$ ./cloudera-manager-installer.bin
```

6. When the installer finishes, navigate to `http://public hostname:7180` and log into the Cloudera Manager Admin console.
7. (optional) Configure TLS encryption. (See Configuring TLS Security for Cloudera Manager).

> **Note:**
>
> - You must upload your AWS account credentials to launch the EC2 instances in the installation wizard, and Cloudera strongly recommends configuring TLS connection.
> - If you encounter any problems, consult Troubleshooting Installation and Upgrade Problems on page 131.

Step 2: Use the Cloud Wizard to provision cloud instances and install Cloudera Manager and CDH.

1. Log into the Cloudera Manager Admin Console on your EC2 instance: `<public hostname>:7180`. The initial user name and password are `admin`.
2. Choose which edition to install:

   - Cloudera Express, which does not require a license, but provides a somewhat limited set of features.
   - Cloudera Enterprise Data Hub Edition Trial, which does not require a license, but expires after 60 days and cannot be renewed
   - Cloudera Enterprise with one of the following license types:
     - Basic Edition
     - Flex Edition
     - Data Hub Edition

If you choose Cloudera Express or Cloudera Enterprise Data Hub Edition Trial, you can elect to upgrade the license at a later time. See Managing Licenses.

3. The Welcome Page appears.

   > **Warning:** Instances provisioned on AWS EC2 by this wizard are instance store-based, so all data will be lost when an instance is stopped or terminated.

   Click **Continue**.
4. Provide the instance specifications:
   a. Choose the OS.
   b. Alternatively, you may use a custom AMI:
      - Make sure the AMI is in the same region as Cloudera Manager Server.
      - Specify the username Cloudera Manager should use to SSH in. This is usually "ubuntu" on Ubuntu systems, and "ec2-user" on most other Linux images on EC2.
   c. Choose the type of EC2 instances you want to provision. Instances not matching the minimum requirements are deliberately removed from the list. For CDH 5 hosts, select **General purpose** > **m1.large** or larger instances.
   d. Specify the number of instances you wish to provision.
   e. Specify the group name (string). This string will be included in the name of your instances and the security group and key pair, which will be created by Cloudera Manager.
5. Provide credentials:
   a. Enter the AWS access and secret key. To create new ones, follow these instructions:
      a. Go to https://console.aws.amazon.com/iam/.
      b. Click **Users**.
      c. Check the box next to the desired user, scroll down and click **Manage Access Keys**.
      d. Copy the new keys and paste them to the inputs.
   b. Choose the instance authentication method:
      a. Let Cloudera Manager create a new SSH key pair for your instances. You will be able to download the private key later to SSH into the new instances.
      b. Import and upload your own key:
         a. In the console, go to **Key Pairs**.
         b. Click **Import Key**.
         c. Select your private key file, specify the name and click **Yes, Import**.
6. Review the Installation settings:
   a. You may go back if you want to correct any information you provided in the previous steps.
   b. Once the instances are provisioned, you must terminate them if you need to modify the installation settings.
   c. Click **Start Installation**.
7. Provision new instances. Once instances are provisioned:
   a. Download the private SSH key if you chose to create one.
8. The wizard leads you through the installation steps:
   a. Install Cloudera Manager and CDH.
   b. Run the Host Inspector.
   c. Start all services.

9. When you are finished, terminate the instances through the AWS EC2 console.

Terminating EC2 Instances

> **Warning:** Cloudera Manager will only terminate instances if the installation fails. You must terminate the instances manually when you are done using the CDH cluster.

1. Sign into the AWS EC2 console.
2. In the left pane, select **Instances**.
3. Select the instances you want to terminate. You may use the string you entered as "group name" to filter the instances provisioned by Cloudera Manager.
4. From **Actions** select **Terminate**.

Using Whirr to Launch Cloudera Manager

Cloudera Manager provides an installation wizard that installs Cloudera Manager, CDH and Impala on a cluster of Amazon Web Services (AWS) EC2 instances. See Installing Cloudera Manager and CDH on EC2 on page 100. Alternatively, you can install Cloudera Manager using Whirr following the instructions here. Follow these instructions to start a cluster on Amazon Elastic Compute Cloud (EC2) running Cloudera Manager.

This method uses Whirr to start a cluster with:

- One host running the Cloudera Manager Admin Console
- A user-selectable number of hosts for the Hadoop cluster itself.

Once Whirr has started the cluster, you use Cloudera Manager in the usual way.

Step 1: Set your AWS credentials as environment variables

Run the following commands from your local host:

```
$ export AWS_ACCESS_KEY_ID=...
$ export AWS_SECRET_ACCESS_KEY=...
```

Step 2: Install Whirr

Install CDH repositories and the whirr package. For CDH 4, see the CDH 4 Installation Guide. For CDH 5, see the CDH 5 Installation Guide.

Create environment variables:

```
$ export WHIRR_HOME=/usr/lib/whirr
$ export PATH=$WHIRR_HOME/bin:$PATH
```

Step 3: Create a password-less SSH Key Pair

Create a password-less SSH Key Pair for Whirr to use:

```
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa_cm
```

Step 4: Get your Whirr-Cloudera-Manager Configuration

You can download a sample Whirr EC2 Cloudera Manager configuration as follows:

```
$ curl -O https://raw.github.com/cloudera/whirr-cm/master/cm-ec2.properties
```

To upload a Cloudera Manager License as part of the installation (Cloudera can provide this if you do not have one), place the license in a file `cm-license.txt` on the Whirr classpath (for example in `$WHIRR_HOME/conf`), using a command such as the following:

```
$ mv -v eval_acme_20120925_cloudera_enterprise_license.txt
$WHIRR_HOME/conf/cm-license.txt
```

To upload a Cloudera Manager configuration as part of the installation, place the configuration in a file called `cm-config.json` on the Whirr classpath (for example in `$WHIRR_HOME/conf`). The format of this file should match the JSON as downloaded from the Cloudera Manager UI. For example:

```
$ curl -O https://raw.github.com/cloudera/whirr-cm/master/cm-config.json
$ mv -v cm-config.json $WHIRR_HOME/conf/cm-config.json
```

Step 5: Launch a Cloudera Manager Cluster

The following command starts a cluster with five Hadoop hosts:

```
$ whirr launch-cluster --config cm-ec2.properties
```

- **Note:**
  - To change the number of hosts edit the `whirr.instance-templates` line in the `cm-ec2.properties` file. For example, to launch a cluster with 20 hosts: `whirr.instance-templates=1 cmserver,20 cmagent`
  - To add a no-op host to use as gateway host: `whirr.instance-templates=1 cmserver,20 cmagent,1 noop`

Whirr reports progress to the console as it runs. The command exits when the cluster is ready to be used.

Using the Cluster

Once the Hadoop cluster is up and running you can run jobs from any Cloudera Manager Agent host, or from a Cloudera Manager gateway host.

Using a Gateway Host (Optional)

In most cases, you will not a need a gateway host, but you may want to consider using one if you want to run jobs on a host that is not also running CDH TaskTracker and DataNode processes. In that case, edit `whirr.instance-templates` to use the noop option shown in the previous section, launch the cluster, and then follow Cloudera Manager instructions to add a gateway role on the no-op host, which you can find in the documentation for your version of Cloudera Manager, for example at Role Instances.

Then SSH to the gateway host. Now you can interact with the cluster; for example, to list files in HDFS:

```
hadoop fs -ls /tmp
```

Shutting Down the Cluster

When you want to shut down the cluster, run the following command.

- **Important:** All data and state stored on the cluster will be lost.

```
whirr destroy-cluster --config cm-ec2.properties
```

Configuring a Custom Java Home Location

Java, which Cloudera services require, may be installed at a custom location. Follow the installation instructions in:

- CDH 5 - (CDH 5) Java Development Kit Installation.
- CDH 4 - (CDH 4) Java Development Kit Installation.

If you choose to use a custom Java location, modify the host configuration to ensure the JDK can be found:

1. Open the Cloudera Manager Admin Console.
2. In the main navigation bar, click the **Hosts** tab and optionally click a specific host link.
3. Click the **Configuration** tab.
4. In the **Advanced** category, click the **Java Home Directory** property.
5. Set the property to the custom location.
6. Click **Save Changes**.
7. Restart all services.

If you don't update the configuration, Cloudera services will be unable to find this resource and will not start.

Installing Older Versions of Cloudera Manager 5

The Cloudera Manager installation solutions, such as the installer downloadable from the Cloudera Downloads website, install the most recent version of Cloudera Manager. This ensures that you install the latest features and bug fixes. While having the latest version of Cloudera Manager is valuable, in some cases it may be necessary to install previous versions.

The most common reason to install a previous version is when you want to expand an existing cluster. In this case, follow the instructions in Adding a Host to the Cluster.

You can also add a cluster to be managed by the same instance of Cloudera Manager – you do this using the **Add Cluster** feature from the Services page in the Cloudera Manager Admin Console. In this case, follow the instructions in Adding a Cluster.

You may also want to install a previous version of the Cloudera Manager server on a new cluster if, for example, you have validated a specific version and want to deploy that version on additional clusters. Installing an older version of Cloudera Manager requires several manual steps to install and configure the database and the correct version of the Cloudera Manager Server. Once these are done, you can run the Express wizard to complete the installation of Cloudera Manager and CDH.

Before You Begin

Install and Configure Databases

Cloudera Manager Server, Cloudera Management Service, and the Hive Metastore data is stored in a database. Install and configure required databases following the instructions in Cloudera Manager and Managed Service Databases on page 21.

(CDH 5 only) On RHEL and CentOS 5, Install Python 2.6 or 2.7

Python 2.6 or 2.7 is required to run Hue. RHEL 5 and CentOS 5, in particular, require the EPEL repository package.

In order to install packages from the EPEL repository, first download the appropriate repository rpm packages to your machine and then install Python using yum. For example, use the following commands for RHEL 5 or CentOS 5:

```
$ su -c 'rpm -Uvh
http://download.fedoraproject.org/pub/epel/5/i386/epel-release-5-4.noarch.rpm'
...
$ yum install python26
```

Establish Your Cloudera Manager Repository Strategy

- **Download and Edit the Repo File for RHEL-compatible OSs or SLES**
  1. Download the Cloudera Manager repo file (`cloudera-manager.repo`) for your OS version using the links provided in the Cloudera Manager Version and Download Information page. For example, for Red Hat/CentOS 6, this is found at `http://archive.cloudera.com/cm5/redhat/6/x86_64/cm/cloudera-manager.repo`
  2. Edit the file to change the baseurl to point to the specific version of Cloudera Manager you want to download. For example, if you want to install Cloudera Manager version 5.0.1, change: `baseurl=http://archive.cloudera.com/cm5/redhat/6/x86_64/cm/5/` to `baseurl=http://archive.cloudera.com/cm5/redhat/6/x86_64/cm/5.0.1/`.
  3. Save the edited file:
     - For Red Hat or CentOS, save it in `/etc/yum.repos.d/`.
     - For SLES, save it in `/etc/zypp/repos.d`.

- **Download and Edit the cloudera.list file for Debian or Apt**
  1. Download the Cloudera Manager list file (`cloudera.list`) using the links provided at Cloudera Manager Version and Download Information. For example, for for Ubuntu 10.04 (lucid), this is found at `http://archive.cloudera.com/cm5/ubuntu/lucid/amd64/cm/cloudera.list`
  2. Edit the file to change the second-to-last element to specify the version of Cloudera Manager you want to install. For example, with Ubuntu lucid, for if you want to install Cloudera Manager version 5.0.1, change: `deb http://archive.cloudera.com/cm5/ubuntu/lucid/amd64/cm lucid-cm5 contrib` to `deb http://archive.cloudera.com/cm5/ubuntu/lucid/amd64/cm lucid-cm5.0.1 contrib`.
  3. Save the edited file in the directory `/etc/apt/sources.list.d/`.

Install the Oracle JDK

Install the Oracle Java Development Kit (JDK) on the Cloudera Manager Server host.

The JDK is included in the Cloudera Manager 5 repositories. Once you have the repo or list file in the correct place, you can install the JDK as follows:

OS	Command
RHEL	`$ sudo yum install oracle-j2sdk1.7`
SLES	`$ sudo zypper install oracle-j2sdk1.7`
Ubuntu or Debian	`$ sudo apt-get install oracle-j2sdk1.7`

Install the Cloudera Manager Server Packages

Install the Cloudera Manager Server packages either on the host where the database is installed, or on a host that has access to the database. This host need not be a host in the cluster that you want to manage with Cloudera Manager. On the Cloudera Manager Server host, type the following commands to install the Cloudera Manager packages.

OS	Command
RHEL, if you have a yum repo configured	`$ sudo yum install cloudera-manager-daemons cloudera-manager-server`
RHEL,if you're manually transferring RPMs	`$ sudo yum --nogpgcheck localinstall cloudera-manager-daemons-*.rpm` `$ sudo yum --nogpgcheck localinstall cloudera-manager-server-*.rpm`
SLES	`$ sudo zypper install cloudera-manager-daemons cloudera-manager-server`
Ubuntu or Debian	`$ sudo apt-get install cloudera-manager-daemons cloudera-manager-server`

Set up a Database for the Cloudera Manager Server

Set up the Cloudera Manager Server database as described in Setting up the Cloudera Manager Server Database on page 22.

(Optional) Install Cloudera Manager Agent, CDH, and Managed Service Software

You can have Cloudera Manager install Cloudera Manager Agent packages or manually install the packages yourself. Similarly, you can allow Cloudera Manager to install CDH and managed service software or manually install the software yourself.

If you choose to have Cloudera Manager install the software (in Choose Software Installation Method and Install Software on page 61), you must satisfy the requirements described in Choosing an Installation Path on page 43. If you satisfy the requirements and choose to have Cloudera Manager install software, you can go to Start the Cloudera Manager Server on page 59. Otherwise, proceed with the following sections.

Install the Oracle JDK

Install the Oracle JDK on the cluster hosts. Cloudera Manager 5 can manage both CDH 5 and CDH 4, and the required JDK version varies accordingly:

- CDH 5 - (CDH 5) Java Development Kit Installation.
- CDH 4 - (CDH 4) Java Development Kit Installation.

Install Cloudera Manager Agent Packages

If you to manually install the packages yourself, on every Cloudera Manager Agent host (including those that will run one or more of the Cloudera Management Service roles: Service Monitor, Activity Monitor, Event Server, Alert Publisher, Reports Manager) do the following:

1. Use one of the following commands to install the Cloudera Manager Agent packages:

OS	Command
RHEL, if you have a yum repo configured:	`$ sudo yum install cloudera-manager-agent cloudera-manager-daemons`
RHEL, if you're manually transferring RPMs:	`$ sudo yum --nogpgcheck localinstall cloudera-manager-agent-package.*.x86_64.rpm cloudera-manager-daemons`
SLES	`$ sudo zypper install cloudera-manager-agent cloudera-manager-daemons`
Ubuntu or Debian	`$ sudo apt-get install cloudera-manager-agent cloudera-manager-daemons`

2. On every Cloudera Manager Agent 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
`server_host`	Name of host where the Cloudera Manager Server is running.
`server_port`	Port on host where the Cloudera Manager Server is running.

For more information on Agent configuration options, see Agent Configuration File.

Install CDH and Managed Service Packages

For more information about manually installing CDH packages, see CDH 4 Installation Guide or CDH 5 Installation Guide.

1. Choose a repository strategy:
   - Standard Cloudera repositories. For this method, ensure you have added the required repository information to your systems.
   - Internally hosted repositories. You might use internal repositories for environments where hosts do not have access to the Internet. In such a case, ensure your environment is properly prepared. For more information, see Understanding Custom Installation Solutions on page 95.
2. Install packages:

CDH Version	Procedure
CDH 5	▪ **Red Hat** **1.** Download and install the "1-click Install" package **a.** Download the CDH 5 "1-click Install" package. Click the entry in the table below that matches your Red Hat or CentOS system, choose **Save File**, and save the file to a directory to which you have write access (it can be your home directory). OS Version: Red Hat/CentOS/Oracle 5 — Click this Link: Red Hat/CentOS/Oracle 5 link OS Version: Red Hat/CentOS/Oracle 6 — Click this Link: Red Hat/CentOS/Oracle 6 link **b.** Install the RPM: ▪ **Red Hat/CentOS/Oracle 5** `$ sudo yum --nogpgcheck localinstall cloudera-cdh-5-0.x86_64.rpm` ▪ **Red Hat/CentOS/Oracle 6** `$ sudo yum --nogpgcheck localinstall cloudera-cdh-5-0.x86_64.rpm` **2.** (Optionally) add a repository key:

CDH Version	Procedure

- **Red Hat/CentOS/Oracle 5**

```
$ sudo rpm --import
http://archive.cloudera.com/cdh5/redhat/5/x86_64/cdh/RPM-GPG-KEY-cloudera
```

- **Red Hat/CentOS/Oracle 6**

```
$ sudo rpm --import
http://archive.cloudera.com/cdh5/redhat/6/x86_64/cdh/RPM-GPG-KEY-cloudera
```

3. Install the CDH packages:

```
$ sudo yum clean all
$ sudo yum install avro-tools crunch flume-ng hadoop-hdfs-fuse
hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat
hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms
hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell
kite llama mahout oozie pig pig-udf-datafu search sentry
solr-mapreduce spark-python sqoop sqoop2 whirr
```

> **Note:** Installing these packages will also install all the other CDH packages that are needed for a full CDH 5 installation.

- **SLES**

1. Download and install the "1-click Install" package.

   a. Download the CDH 5 "1-click Install" package.

   Click this link, choose **Save File**, and save it to a directory to which you have write access (it can be your home directory).

   b. Install the RPM:

```
$ sudo rpm -i cloudera-cdh-5-0.x86_64.rpm
```

   c. Update your system package index by running:

```
$ sudo zypper refresh
```

2. (Optionally) add a repository key:

```
$ sudo rpm --import
http://archive.cloudera.com/cdh5/sles/11/x86_64/cdh/RPM-GPG-KEY-cloudera
```

3. Install the CDH packages:

```
$ sudo zypper clean --all
$ sudo zypper install avro-tools crunch flume-ng hadoop-hdfs-fuse
hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat
hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms
hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell
kite llama mahout oozie pig pig-udf-datafu search sentry
solr-mapreduce spark-python sqoop sqoop2 whirr
```

> **Note:** Installing these packages will also install all the other CDH packages that are needed for a full CDH 5 installation.

<table>
<tr><th>CDH Version</th><th>Procedure</th></tr>
<tr><td></td><td>

- **Ubuntu and Debian**

1. Download and install the "1-click Install" package
   a. Download the CDH 5 "1-click Install" package:

OS Version	Click this Link
Wheezy	Wheezy link
Precise	Precise link

   b. Install the package. Do one of the following:
      - Choose **Open with** in the download window to use the package manager.
      - Choose **Save File**, save the package to a directory to which you have write access (it can be your home directory) and install it from the command line, for example:

```
sudo dpkg -i cdh5-repository_1.0_all.deb
```

2. (Optionally) add a repository key:
   - **Debian Wheezy**

```
$ curl -s
http://archive.cloudera.com/cdh5/debian/wheezy/amd64/cdh/archive.key
  | sudo apt-key add -
```

   - **Ubuntu Precise**

```
$ curl -s
http://archive.cloudera.com/cdh5/ubuntu/precise/amd64/cdh/archive.key
  | sudo apt-key add -
```

3. Install the CDH packages:

```
$ sudo apt-get update
$ sudo apt-get install avro-tools crunch flume-ng hadoop-hdfs-fuse
hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat
hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms
hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell
kite llama mahout oozie pig pig-udf-datafu search sentry
solr-mapreduce spark-python sqoop sqoop2 whirr
```

> **Note:** Installing these packages will also install all the other CDH packages that are needed for a full CDH 5 installation.

</td></tr>
<tr><td>CDH 4, Impala, and Solr</td><td>

- **Red Hat-compatible**

1. Click the entry in the table at CDH Download Information that matches your Red Hat or CentOS system.
2. Navigate to the repo file (`cloudera-cdh4.repo`) for your system and save it in the `/etc/yum.repos.d/` directory.
3. Optionally add a repository key:

</td></tr>
</table>

CDH Version	Procedure

- **Red Hat/CentOS/Oracle 5**

```
$ sudo rpm --import
http://archive.cloudera.com/cdh4/redhat/5/x86_64/cdh/RPM-GPG-KEY-cloudera
```

- **Red Hat/CentOS 6**

```
$ sudo rpm --import
http://archive.cloudera.com/cdh4/redhat/6/x86_64/cdh/RPM-GPG-KEY-cloudera
```

4. Install packages on every host in your cluster:

   a. Install CDH 4 packages:

```
$ sudo yum -y install bigtop-utils bigtop-jsvc bigtop-tomcat hadoop
 hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn
hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie
 oozie-client pig zookeeper
```

   b. To install the hue-common package and all Hue applications on the Hue host, install the hue meta-package:

```
$ sudo yum install hue
```

5. (**Requires CDH 4.2 or later**) Install Impala

   a. Click the entry in the table at Cloudera Impala Version and Download Information that matches your Red Hat or CentOS system.
   b. Navigate to the repo file for your system and save it in the `/etc/yum.repos.d/` directory.
   c. Install Impala and the Impala Shell on Impala machines:

```
$ sudo yum -y install impala impala-shell
```

6. (**Requires CDH 4.3 or later**) Install Search

   a. Click the entry in the table at Cloudera Search Version and Download Information that matches your Red Hat or CentOS system.
   b. Navigate to the repo file for your system and save it in the `/etc/yum.repos.d/` directory.
   c. Install the Solr Server on machines where you want Cloudera Search.

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

- **SLES**

1. Run the following command:

```
$ sudo zypper addrepo -f
http://archive.cloudera.com/cdh4/sles/11/x86_64/cdh/cloudera-cdh4.repo
```

2. Update your system package index by running:

```
$ sudo zypper refresh
```

CDH Version	Procedure

3. Optionally add a repository key:

```
$ sudo rpm --import
http://archive.cloudera.com/cdh4/sles/11/x86_64/cdh/RPM-GPG-KEY-cloudera
```

4. Install packages on every host in your cluster:

   a. Install CDH 4 packages:

```
$ sudo zypper install bigtop-utils bigtop-jsvc bigtop-tomcat hadoop
 hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn
hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie
 oozie-client pig zookeeper
```

   b. To install the `hue-common` package and all Hue applications on the Hue host, install the hue meta-package:

```
$ sudo zypper install hue
```

   c. (**Requires CDH 4.2 or later**) Install Impala

      a. Run the following command:

```
$ sudo zypper addrepo -f
http://archive.cloudera.com/impala/sles/11/x86_64/impala/cloudera-impala.repo
```

      b. Install Impala and the Impala Shell on Impala machines:

```
$ sudo zypper install impala impala-shell
```

   d. (**Requires CDH 4.3 or later**) Install Search

      a. Run the following command:

```
$ sudo zypper addrepo -f
http://archive.cloudera.com/search/sles/11/x86_64/search/cloudera-search.repo
```

      b. Install the Solr Server on machines where you want Cloudera Search.

```
$ sudo zypper install solr-server
```

- **Ubuntu or Debian**

1. Click the entry in the table at CDH Version and Packaging Information that matches your Ubuntu or Debian system.
2. Navigate to the list file (`cloudera.list`) for your system and save it in the `/etc/apt/sources.list.d/` directory. For example, to install CDH 4 for 64-bit Ubuntu Lucid, your `cloudera.list` file should look like:

```
deb [arch=amd64]
http://archive.cloudera.com/cdh4/ubuntu/lucid/amd64/cdh lucid-cdh4
contrib
deb-src http://archive.cloudera.com/cdh4/ubuntu/lucid/amd64/cdh
lucid-cdh4 contrib
```

3. Optionally add a repository key:

CDH Version	Procedure

- **Ubuntu Lucid**

```
$ curl -s
http://archive.cloudera.com/cdh4/ubuntu/lucid/amd64/cdh/archive.key
  | sudo apt-key add -
```

- **Ubuntu Precise**

```
$ curl -s
http://archive.cloudera.com/cdh4/ubuntu/precise/amd64/cdh/archive.key
  | sudo apt-key add -
```

- **Debian Squeeze**

```
$ curl -s
http://archive.cloudera.com/cdh4/debian/squeeze/amd64/cdh/archive.key
  | sudo apt-key add -
```

4. Install packages on every host in your cluster:

   a. Install CDH 4 packages:

```
$ sudo apt-get install bigtop-utils bigtop-jsvc bigtop-tomcat
hadoop hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn
hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie
 oozie-client pig zookeeper
```

   b. To install the hue-common package and all Hue applications on the Hue host, install the hue meta-package:

```
$ sudo apt-get install hue
```

   c. (**Requires CDH 4.2 or later**) Install Impala

      a. Click the entry in the table at Cloudera Impala Version and Download Information and that matches your Ubuntu or Debian system.
      b. Navigate to the list file for your system and save it in the `/etc/apt/sources.list.d/` directory.
      c. Install Impala and the Impala Shell on Impala machines:

```
$ sudo apt-get install impala impala-shell
```

   d. (**Requires CDH 4.3 or later**) Install Search

      a. Click the entry in the table at Cloudera Search Version and Download Information that matches your Ubuntu or Debian system.
      b. Install Solr Server on machines where you want Cloudera Search:

```
$ sudo apt-get install solr-server
```

Start the Cloudera Manager Server

- **Important:** When you start the Cloudera Manager Server and Agents, Cloudera Manager assumes you are not already running HDFS and MapReduce. If these services are running:
  1. Shut down HDFS and MapReduce. See Stopping Services (for CDH 4) or Stopping Services (for CDH 5) for the commands to stop these services.
  2. Configure the init scripts to *not* start on boot, use commands similar to those shown in Configuring init to Start Core Hadoop System Services or Configuring init to Start Core Hadoop System Services but *disable* the start on boot (for example, `$ sudo chkconfig hadoop-hdfs-namenode off`).

  Contact Cloudera Support for help converting your existing Hadoop configurations for use with Cloudera Manager.

1. Run this command on the Cloudera Manager Server host:

```
$ sudo service cloudera-scm-server start
```

If the Cloudera Manager Server does not start, see Troubleshooting Installation and Upgrade Problems on page 131.

(Optional) Start the Cloudera Manager Agents

If you installed the Cloudera Manager Agent packages in Install Cloudera Manager Agent Packages on page 53, run this command on each Agent host:

```
$ sudo service cloudera-scm-agent start
```

When the Agent starts up, it contacts the Cloudera Manager Server. If there is a communication failure between a Cloudera Manager Agent and Cloudera Manager Server, see Troubleshooting Installation and Upgrade Problems on page 131.

When the Agent hosts reboot, `cloudera-scm-agent` starts automatically.

Start the Cloudera Manager Admin Console

The Cloudera Manager Server URL takes the following form `http://Server host:port`, where *Server host* is the fully-qualified domain name or IP address of the host where the Cloudera Manager Server is installed and *port* is the port configured for the Cloudera Manager Server. The default port is 7180.

1. Wait several minutes for the Cloudera Manager Server to complete its startup. To observe the startup process you can perform `tail -f /var/log/cloudera-scm-server/cloudera-scm-server.log` on the Cloudera Manager Server host. If the Cloudera Manager Server does not start, see Troubleshooting Installation and Upgrade Problems on page 131.
2. In a web browser, enter `http://Server host:7180`, where *Server host* is the fully-qualified domain name or IP address of the host where you installed the Cloudera Manager Server. The login screen for Cloudera Manager Admin Console displays.
3. Log into Cloudera Manager Admin Console. The default credentials are: **Username:** `admin` **Password:** `admin`. 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. While 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.

Choose Cloudera Manager Edition and Hosts

The following instructions describe how to use the Cloudera Manager wizard to choose which edition of Cloudera Manager you are using and which hosts will run CDH and managed services.

1. When you start the Cloudera Manager Admin Console, the install wizard starts up. Click **Continue** to get started.
2. Choose which edition to install:
   - Cloudera Express, which does not require a license, but provides a somewhat limited set of features.
   - Cloudera Enterprise Data Hub Edition Trial, which does not require a license, but expires after 60 days and cannot be renewed
   - Cloudera Enterprise with one of the following license types:
     - Basic Edition
     - Flex Edition
     - Data Hub Edition

   If you choose Cloudera Express or Cloudera Enterprise Data Hub Edition Trial, you can elect to upgrade the license at a later time. See Managing Licenses.
3. If you have elected Cloudera Enterprise, install a license:
   a. Click **Upload License**.
   b. Click the document icon to the left of the **Select a License File** text field.
   c. Navigate to the location of your license file, click the file, and click **Open**.
   d. Click **Upload**.

   Click **Continue** to proceed with the installation.
4. Click **Continue** in the next screen. The **Specify Hosts** page displays.
5. Do one of the following:
   - If you installed Cloudera Agent packages in Install Cloudera Manager Agent Packages on page 53, choose from among hosts with the packages installed:
     1. Click the **Currently Managed Hosts** tab.
     2. Choose the hosts to add to the cluster.
   - Search for and choose hosts:
     1. To enable Cloudera Manager to automatically discover hosts on which to install CDH and managed services, enter the cluster hostnames or IP addresses. You can also specify hostname and IP address ranges. For example:

Range Definition	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].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

     You can specify multiple addresses and address ranges by separating them by 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. The scan results will include all addresses scanned, but only scans that reach hosts running SSH will be selected for inclusion in your cluster by default. If you don't know the IP addresses of all of the hosts, you can enter an address range that spans over unused addresses and then deselect the hosts that do not exist (and are not discovered) later in this procedure. However, keep in mind that wider ranges will require more time to scan.

     2. Click **Search**. Cloudera Manager identifies the hosts on your cluster to allow you to configure them for services. 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**. To find additional hosts, click **New Search**, add the host names or IP addresses and click **Search** again. 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. Common causes of loss of connectivity are firewalls and interference from SELinux.

3. Verify that the number of hosts shown matches the number of hosts where you want to install services. Deselect host entries that do not exist and deselect the hosts where you do not want to install services. Click **Continue**. The Select Repository page displays.

6. Click **Continue**. The **Select Repository** page displays.

Choose Software Installation Method and Install Software

The following instructions describe how to use the Cloudera Manager wizard to install Cloudera Manager Agent, CDH, and managed service software.

1. Select how CDH and managed service software is installed: packages or parcels:
   - **Use Packages** - If you *did not* install packages in Install CDH and Managed Service Packages on page 53, click the package versions to install. Otherwise, select the CDH version (CDH 4 or CDH 5) that matches the packages that you installed manually.
   - **Use Parcels**
     1. Choose the parcels to install. The choices you see depend on the repositories you have chosen – a repository may contain multiple parcels. Only the parcels for the latest supported service versions are configured by default.

        You can add additional parcels for previous versions by specifying custom repositories. For example, you can find the locations of the previous CDH 4 parcels at `http://archive.cloudera.com/cdh4/parcels/`. Or, if you are installing CDH 4.3 and want to use Sentry for Policy File-Based Hive Authorization, you can add the Sentry parcel using this mechanism.

        1. To specify the parcel directory, local parcel repository, add a parcel repository, or specify the properties of a proxy server through which parcels are downloaded, click the **More Options** button and do one or more of the following:
           - **Parcel Directory** and **Local Parcel Repository Path** - Specify the location of parcels on cluster hosts and the Cloudera Manager Server host.
           - **Parcel Repository** - In the **Remote Parcel Repository URLs** field, click the + button and enter the URL of the repository. The URL you specify is added to the list of repositories listed in the Configuring Server Parcel Settings on page 84 page and a parcel is added to the list of parcels on the Select Repository page. If you have multiple repositories configured, you will see all the unique parcels contained in all your repositories.
           - **Proxy Server** - Specify the properties of a proxy server.
        2. Click **OK**.

2. If you *did not* install Cloudera Manager Agent packages in Install Cloudera Manager Agent Packages on page 53, do the following:
   a. Select the release of Cloudera Manager Agent to install. You can choose either the version that matches the Cloudera Manager Server you are currently using or specify a version in a custom repository.
   b. If you opted to use custom repositories for installation files, you can provide a GPG key URL that applies for all repositories.
3. Click **Continue**.
   - (Cloudera Manager 5.1.3) Leave **Install Oracle Java SE Development Kit (JDK)** checked to allow Cloudera Manager to install the JDK on each cluster host or uncheck if you plan to install it yourself.
   - If your local laws permit you to deploy unlimited strength encryption and you are running a secure cluster, check the **Install Java Unlimited Strength Encryption Policy Files** checkbox.

Click **Continue**.

4. If your local laws permit you to deploy unlimited strength encryption and you are running a secure cluster, check the **Install Java Unlimited Strength Encryption Policy Files** checkbox.
5. If you chose to have Cloudera Manager install packages, specify host installation properties:
   a. Select **root** or enter the user name for an account that has password-less sudo permission.
   b. Select an authentication method:
      - If you choose to use password authentication, enter and confirm the password.
      - If you choose to use public-key authentication provide a passphrase and path to the required key files.
   c. You can choose to specify an alternate SSH port. The default value is 22.
   d. You can specify the maximum number of host installations to run at once. The default value is 10.
6. Click **Continue**. If you *did not* install packages in (Optional) Install Cloudera Manager Agent, CDH, and Managed Service Software on page 53, Cloudera Manager installs the Oracle JDK, Cloudera Manager Agent,packages and CDH and managed service packages or parcels. During the parcel installation, progress is indicated for the two phases of the parcel installation process (Download and Distribution) in a separate progress bars. If you are installing multiple parcels you will see progress bars for each parcel. When the **Continue** button appears at the bottom of the screen, the installation process is completed. Click **Continue**.
7. Click **Continue**. The Host Inspector runs to validate the installation, and provides a summary of what it finds, including all the versions of the installed components. If the validation is successful, click **Finish**. The Cluster Setup page displays.

Add Services

The following instructions describe how to use the Cloudera Manager wizard to configure and start CDH and managed services.

1. In the first page of the Add Services wizard you choose the combination of services to install and whether to install Cloudera Navigator:
   - Click the radio button next to the combination of services to install:

CDH 4	CDH 5
▪ **Core Hadoop** - HDFS, MapReduce, ZooKeeper, Oozie, Hive, and Hue ▪ **Core with HBase** ▪ **Core with Impala** ▪ **All Services** - HDFS, MapReduce, ZooKeeper, HBase, Impala, Oozie, Hive, Hue, and Sqoop ▪ **Custom Services** - Any combination of services.	▪ **Core Hadoop** - HDFS, YARN (includes MapReduce 2), ZooKeeper, Oozie, Hive, Hue, and Sqoop ▪ **Core with HBase** ▪ **Core with Impala** ▪ **Core with Search** ▪ **Core with Spark** ▪ **All Services** - HDFS, YARN (includes MapReduce 2), ZooKeeper, Oozie, Hive, Hue, Sqoop, HBase, Impala, Solr, Spark, and Key-Value Store Indexer ▪ **Custom Services** - Any combination of services.

   As you select the services, keep the following in mind:

   – Some services depend on other services; for example, HBase requires HDFS and ZooKeeper. Cloudera Manager tracks dependencies and installs the correct combination of services.
   – In a CDH 4 cluster, the MapReduce service is the default MapReduce computation framework. Choose **Custom Services** to install YARN or use the Add Service functionality to add YARN after installation completes.

   > ▪ **Important:** You can create a YARN service in a CDH 4 cluster, but it is not considered production ready.

  - In a CDH 5 cluster, the YARN service is the default MapReduce computation framework. Choose **Custom Services** to install MapReduce or use the Add Service functionality to add MapReduce after installation completes.

    > **Important:** In CDH 5 the MapReduce service has been deprecated. However, the MapReduce service is fully supported for backward compatibility through the CDH 5 life cycle.

  - The Flume service can be added only after your cluster has been set up.

- If you have chosen Data Hub Edition Trial or Cloudera Enterprise, optionally check the **Include Cloudera Navigator** checkbox to enable Cloudera Navigator. See the Cloudera Navigator Documentation.

Click **Continue**. The Customize Role Assignments page displays.

2. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign role instances to hosts of your choosing, if desired.

   Click a field below a role to display a dialog containing a pageable list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

   The following shortcuts for specifying hostname patterns are supported:

   - Range of hostnames (without the domain portion)

Range Definition	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].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

   - IP addresses
   - Rack name

   Click the **View By Host** button for an overview of the role assignment by hostname ranges.

3. When you are satisfied with the assignments, click **Continue**. The Database Setup page displays.
4. On the Database Setup page, configure settings for required databases:

   a. Enter the database host, database type, database name, username, and password for the database that you created when you set up the database.
   b. Click **Test Connection** to confirm that Cloudera Manager can communicate with the database using the information you have supplied. If the test succeeds in all cases, click **Continue**; otherwise check and correct the information you have provided for the database and then try the test again. (For some servers, if you are using the embedded database, you will see a message saying the database will be created at a later step in the installation process.) The Review Changes page displays.

5. Review the configuration changes to be applied. Confirm the settings entered for file system paths. The file paths required vary based on the services to be installed.

   > **Warning:** DataNode data directories should not be placed on NAS devices.

   Click **Continue**. The wizard starts the services.
6. When all of the services are started, click **Continue**. You will see a success message indicating that your cluster has been successfully started.
7. Click **Finish** to proceed to the Home Page.

Change the Default Administrator Password

As soon as possible after running the wizard and beginning to use Cloudera Manager, change the default administrator password:

1. Right-click the logged-in username at the far right of the top navigation bar and select **Change Password**.
2. Enter the current password, and a new password twice and then click **Update**.

Test the Installation

You can test the installation following the instructions in Testing the Installation on page 123.

Deploying Clients

Client configuration files are generated automatically by Cloudera Manager based on the services you install.

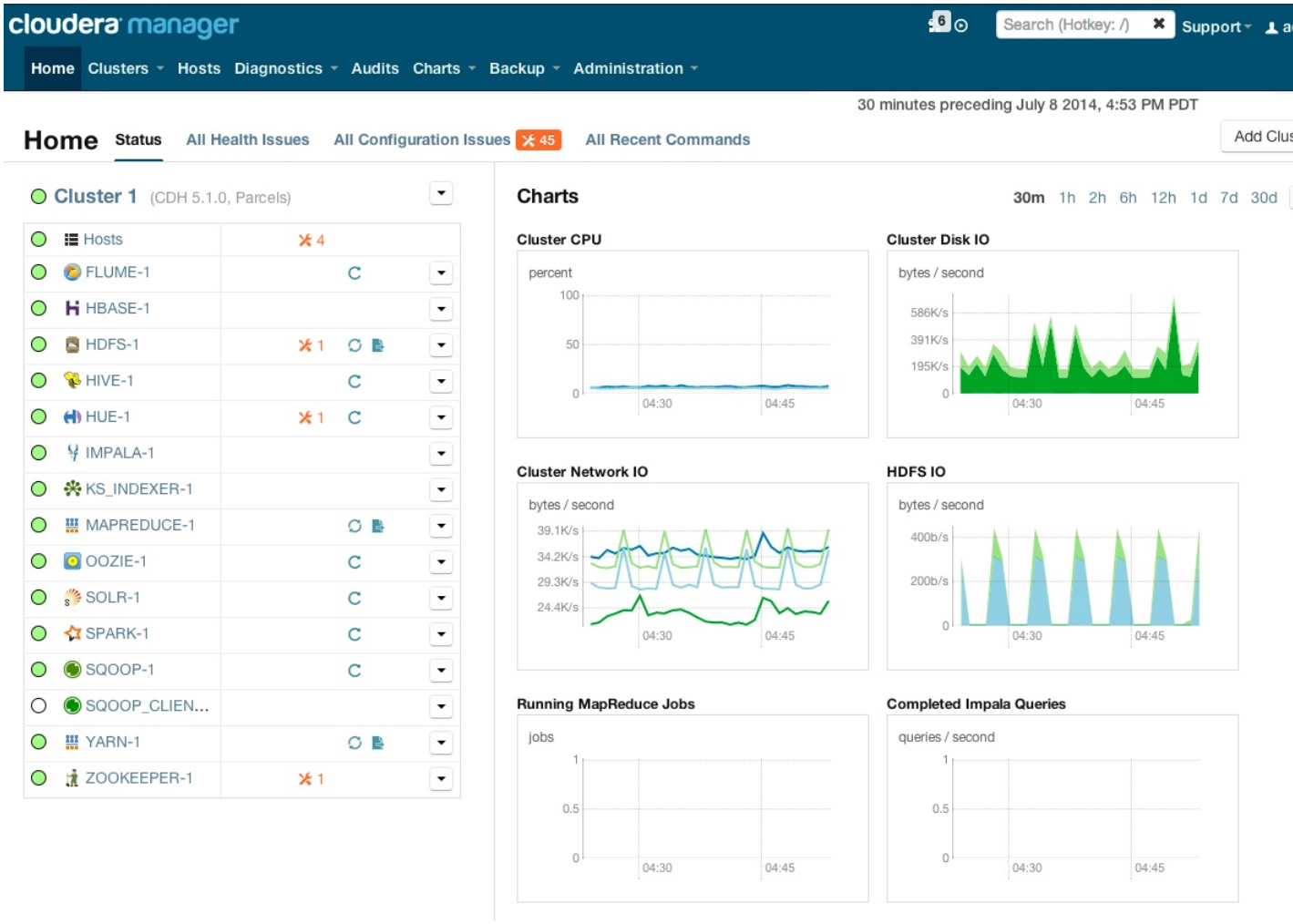
Cloudera Manager deploys these configurations automatically at the end of the installation workflow. You can also download the client configuration files to deploy them manually.

If you modify the configuration of your cluster, you may need to redeploy the client configuration files. If a service's status is "Client configuration redeployment required," you need to redeploy those files.

See Client Configuration Files for information on downloading client configuration files, or redeploying them through Cloudera Manager.

Testing the Installation

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 on 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

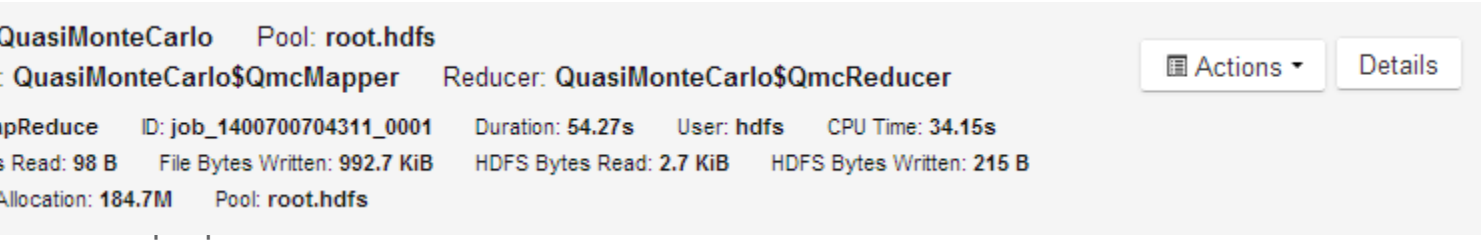
One way to check whether all the Agents are running is to look at the time since their last heartbeat. You can do this by clicking the **Hosts** tab where you can see a list of all the Hosts along with the value of their **Last Heartbeat**. By default, every Agent must heartbeat successfully every 15 seconds. A recent value for the **Last Heartbeat** means that the Server and Agents are communicating successfully.

Running a MapReduce Job

1. Log into a host in the cluster.
2. Run the Hadoop PiEstimator example using one of the following commands:
   - **Parcel** - `sudo -u hdfs hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 10 100`
   - **Package** - `sudo -u hdfs hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 10 100`

   or create and run the WordCount v1.0 application described in Hadoop Tutorial.
3. Depending on whether your cluster is configured to run MapReduce jobs on the YARN or MapReduce service, view the results of running the job by selecting one of the following from the top navigation bar in the Cloudera Manager Admin Console :
   - **Clusters** > ***ClusterName*** > **yarn Applications**
   - **Clusters** > ***ClusterName*** > **mapreduce Activities**

   If you run the PiEstimator job on the YARN service (the default) you will see an entry like the following in **yarn**



Testing with Hue

A good way to test the cluster is by running a job. In addition, you can test the cluster by running one of the Hue web applications. Hue is a graphical user interface that allows you to interact with your clusters by running applications that let you browse HDFS, manage a Hive metastore, and run Hive, Impala, and Search queries, Pig scripts, and Oozie workflows.

1. In the Cloudera Manager Admin Console Home page, click the Hue service.
2. Click the **Hue Web UI** tab, which opens Hue in a new window.
3. Log in with the credentials, **username:** `hdfs`, **password:** `hdfs`.
4. Choose an application in the navigation bar at the top of the browser window.

For more information, see the Hue User Guide.

Uninstalling Cloudera Manager and Managed Software

Use the following instructions to uninstall the Cloudera Manager Server, Agents, managed software, and databases.

Reverting an Incomplete Installation

If you have come to this page because your installation did not complete (for example, if it was interrupted by a virtual machine timeout), and you want to proceed with the installation, do the following before reinstalling:

1. Remove files and directories:

```
$ sudo rm -Rf /usr/share/cmf /var/lib/cloudera* /var/cache/yum/cloudera*
```

Uninstalling Cloudera Manager and Managed Software

Follow the steps in this section to remove software and data.

Record User Data Paths

The user data paths listed Remove User Data on page 129, `/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 /dfs /mapred /yarn`, are the default settings. However, at some point they may 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 CDH and managed services or at some later time, note the location of the paths by checking the configuration in each service.

Stop all Services

1. For each cluster managed by Cloudera Manager:
   a. On the Home page, click  to the right of the cluster name and select **Stop**.
   b. 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.
   c. On the Home page, click  to the right of the Cloudera Management Service entry and select **Stop**. 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.
2. Stop the Cloudera Management Service.

Deactivate and Remove Parcels

If you installed using packages, skip this step and go to Uninstall the Cloudera Manager Server on page 126; you will remove packages in Uninstall Cloudera Manager Agent and Managed Software on page 126. If you installed using parcels remove them as follows:

1. Click the parcel indicator  in the main navigation bar.
2. For each activated parcel, select **Actions** > **Deactivate**. When this action has completed, the parcel button changes to **Activate**.
3. For each activated parcel, select **Actions** > **Remove from Hosts**. When this action has completed, the parcel button changes to **Distribute**.

4. For each activated parcel, select **Actions** > **Delete**. This removes the parcel from the local parcel repository.

There may 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.

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.

- **If you used the cloudera-manager-installer.bin file** - Run the following command on the Cloudera Manager Server host:

```
$ sudo /usr/share/cmf/uninstall-cloudera-manager.sh
```

  - **Note:** If the `uninstall-cloudera-manager.sh` is not installed on your cluster, use the following instructions to uninstall the Cloudera Manager Server.

- **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.

  1. Stop the Cloudera Manager Server and its database:

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

  2. 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.

     Red Hat systems:

```
sudo yum remove cloudera-manager-server
sudo yum remove cloudera-manager-server-db-2
```

     SLES systems:

```
sudo zypper -n rm --force-resolution cloudera-manager-server
sudo zypper -n rm --force-resolution cloudera-manager-server-db-2
```

     Debian/Ubuntu systems:

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

Uninstall Cloudera Manager Agent and Managed Software

Do the following on all Agent hosts:

1. Stop the Cloudera Manager Agent.

   Red Hat/SLES systems:

```
$ sudo service cloudera-scm-agent hard_stop
```

Debian/Ubuntu systems:

```
$ sudo /usr/sbin/service cloudera-scm-agent hard_stop
```

2. Uninstall software:

OS	Parcel Install	Package Install

Red Hat	`$ sudo yum remove 'cloudera-manager-*'`	▪ **CDH 4**
		`$ sudo yum remove 'cloudera-manager-*' bigtop-utils bigtop-jsvc bigtop-tomcat hadoop hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie oozie-client pig zookeeper hue impala impala-shell solr-server`
		▪ **CDH 5**
		`$ sudo yum remove 'cloudera-manager-*' avro-tools crunch flume-ng hadoop-hdfs-fuse hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell kite llama mahout oozie pig pig-udf-datafu search sentry solr-mapreduce spark-python sqoop sqoop2 whirr`

SLES	`$ sudo zypper remove 'cloudera-manager-*'`	▪ **CDH 4**
		`$ sudo zypper remove 'cloudera-manager-*' bigtop-utils bigtop-jsvc bigtop-tomcat hadoop hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie oozie-client pig zookeeper hue impala impala-shell solr-server`
		▪ **CDH 5**
		`$ sudo zypper remove 'cloudera-manager-*' avro-tools crunch flume-ng hadoop-hdfs-fuse hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell kite llama mahout oozie pig pig-udf-datafu search sentry solr-mapreduce spark-python sqoop sqoop2 whirr`

Debian/Ubuntu	`$ sudo apt-get purge 'cloudera-manager-*'`	▪ **CDH 4** `$ sudo apt-get purge 'cloudera-manager-*' bigtop-utils bigtop-jsvc bigtop-tomcat hadoop hadoop-hdfs hadoop-httpfs hadoop-mapreduce hadoop-yarn hadoop-client hadoop-0.20-mapreduce hue-plugins hbase hive oozie oozie-client pig zookeeper hue impala impala-shell solr-server` ▪ **CDH 5**

OS	Parcel Install	Package Install
		$ sudo apt-get purge 'cloudera-manager-*' avro-tools crunch flume-ng hadoop-hdfs-fuse hadoop-hdfs-nfs3 hadoop-httpfs hbase-solr hive-hbase hive-webhcat hue-beeswax hue-hbase hue-impala hue-pig hue-plugins hue-rdbms hue-search hue-spark hue-sqoop hue-zookeeper impala impala-shell kite llama mahout oozie pig pig-udf-datafu search sentry solr-mapreduce spark-python sqoop sqoop2 whirr

3. Run the `clean` command:

Red Hat

```
$ sudo yum clean all
```

SLES

```
$ sudo zypper clean
```

Debian/Ubuntu

```
$ sudo apt-get clean
```

Remove Cloudera Manager and User Data

Kill Cloudera Manager and Managed Processes

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.

Remove Cloudera Manager Data

This step permanently removes 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`. On all Agent hosts, run the following command:

```
$ sudo rm -Rf /usr/share/cmf /var/lib/cloudera* /var/cache/yum/cloudera*
/var/log/cloudera* /var/run/cloudera*
```

Remove the Cloudera Manager Lock File

On all Agent hosts, run this command to remove the Cloudera Manager lock file:

```
$ sudo rm /tmp/.scm_prepare_node.lock
```

Remove User Data

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. 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
```

```
$ sudo rm -Rf /dfs /mapred /yarn
```

- **Note:** For additional information about uninstalling CDH, including clean-up of CDH files, see the entry on Uninstalling CDH Components in the CDH4 Installation Guide or CDH 5 Installation Guide.

Stop and Remove External 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.

Troubleshooting Installation and Upgrade Problems

For information on known issues, see http://www.cloudera.com/content/cloudera-content/cloudera-docs/CM5/latest/Cloudera-Manager-Release-Notes/cm5rn_known_issues.html.

Symptom	Reason	Solution
"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...`	You may have SELinux enabled.	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 won't restart.	You need to do some manual cleanup.	See Uninstalling Cloudera Manager and Managed Software on page 125.
Cloudera Manager Server fails to start and the Server is configured to use a MySQL database to store information about service configuration.	Tables may be configured with the ISAM engine. The Server will not start if its tables are configured with the MyISAM engine, and an error such as the following will appear in the log file: `Tables ... have unsupported engine type ... . InnoDB is required.`	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 MySQL Database on page 30.
Agents fail to connect to server. Error 113 ('No route to host') in `/var/log/cloudera-scm-agent/cloudera-scm-agent.log`	You may have SELinux or `iptables` enabled.	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`.
Some cluster hosts do not appear when you click **Find Hosts** in install or update wizard.	You may have network connectivity problems.	▪ 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 during database configuration for Activity Monitor or Reports Manager.	Hostname mapping or permissions are incorrectly set up.	▪ For hostname configuration, see Configuring Network Names (for CDH 5) or Configuring Network Names (for CDH 4). ▪ 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.

Symptom	Reason	Solution

		For example, if you entered the following for the Activity Monitor database:
		`grant all on activity_monitor.* TO 'amon_user'@'localhost' IDENTIFIED BY 'amon_password';`
		the value you enter here for the database hostname must be `localhost`. On the other hand, 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`. Similarly, 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`.

Activity Monitor, Reports Manager, or Service Monitor databases fail to start.	MySQL `binlog` format problem.	Set `binlog_format=mixed` in `/etc/my.cnf`. For more information, see this MySQL bug report. See also

Symptom	Reason	Solution
		Cloudera Manager and Managed Service Databases on page 21.
You have upgraded the Cloudera Manager Server, but now cannot start services.	You may have mismatched versions of the Cloudera Manager Server and Agents.	Make sure you have upgraded the Cloudera Manager Agents on all hosts. (The previous version of the Agents will heartbeat with the new version of the Server, but you can't start HDFS and MapReduce with this combination.)
Cloudera services fail to start.	Java may not be installed or may be installed at a custom location.	See Configuring a Custom Java Home Location on page 105 for more information on resolving this issue.
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.`	The MySQL thread stack is too small.	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.
The Activity Monitor fails to start. Logs contain the error `read-committed isolation not safe for the statement binlog format`.	The `binlog_format` is not set to mixed.	Modify the `mysql.cnf` file to include the entry for `binlog format` as specified in MySQL Database on page 30.
Attempts to reinstall older versions of CDH or Cloudera Manager using `yum` fails.	It is possible to install, uninstall, and reinstall CDH and Cloudera Manager. In certain cases, this does not complete as expected. If you install Cloudera Manager 5 and CDH 5, then uninstall Cloudera Manager and CDH, and then attempt to install CDH 4 and Cloudera Manager 4, incorrect cached information may result in the installation of an incompatible version of the Oracle JDK.	Clear information in the yum cache: 1. Connect to the CDH host. 2. Execute either of the following commands: `$ yum --enablerepo='*'clean all` or `$ rm -rf /var/cache/yum/cloudera*` 3. After clearing the cache, proceed with installation.
Hive, Impala, or Hue complains about a missing table in the Hive Metastore database.	The Hive Metastore database must be upgraded after a major Hive version change (Hive had a major version change in CDH 4.0, 4.1, 4.2, and 5.0).	Follow the instructions in the Upgrading Hive for upgrading the Hive Metastore database schema. Stop all Hive services before performing the upgrade.
The **Create Hive Metastore Database Tables** command fails due to a problem with an escape string.	PostgreSQL versions 9 and later 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	As the administrator user, use the following command to turn `standard_conforming_strings` off: `ALTER DATABASE <hive_db_name> SET`

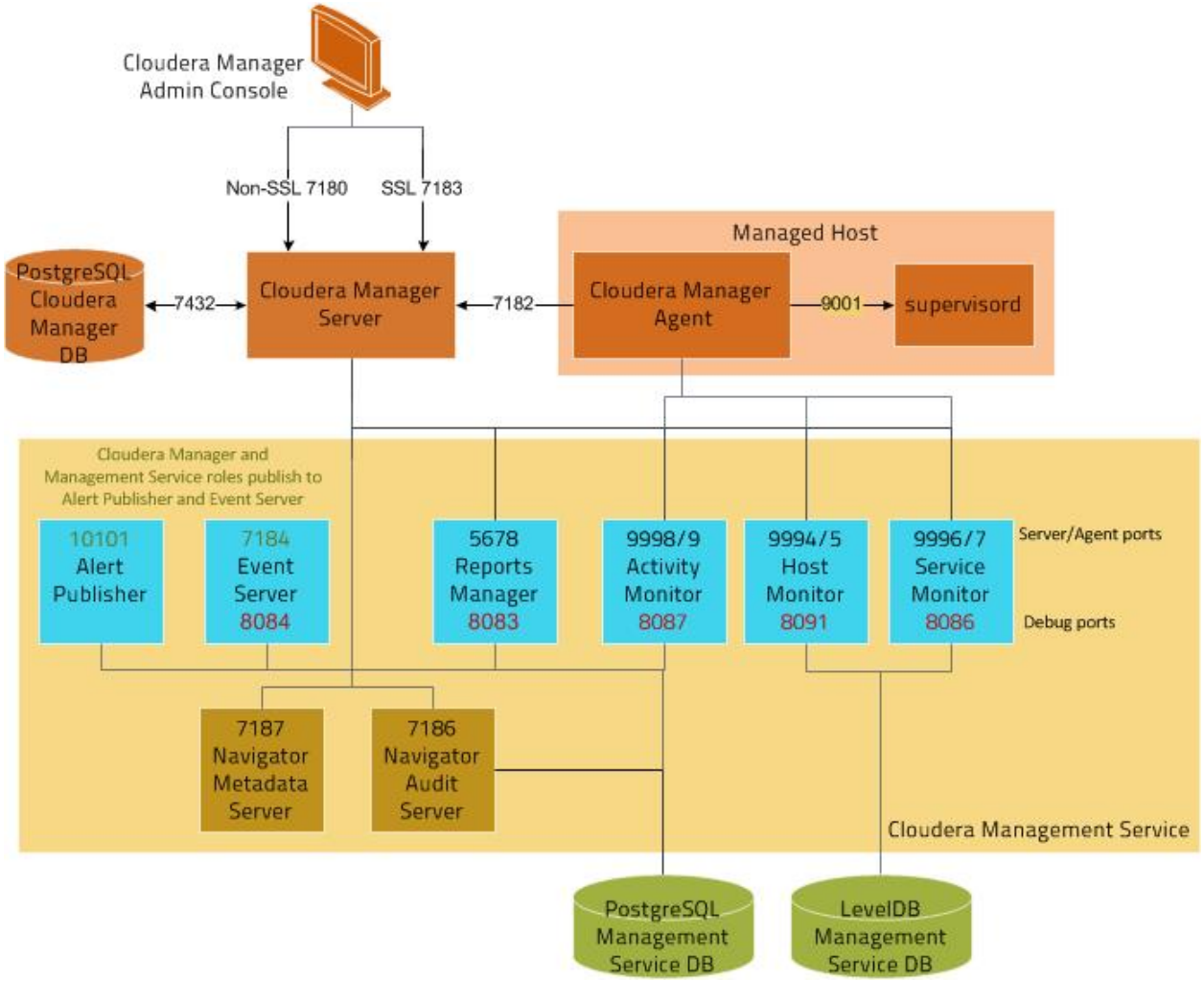
Symptom	Reason	Solution
	9.0 defaulted to `off`, but starting with version 9.0 the default is on.	`standard_conforming_strings = off;`
After upgrading to CDH 5, HDFS DataNodes fail to start with exception: `Exception in securemainjava.lang.RuntimeException: Cannot start datanode because the configured max locked memory size (dfs.datanode.max.locked.memory) of 4294967296 bytes is more than the datanode's available RLIMIT_MEMLOCK ulimit of 65536 bytes.`	HDFS caching, which is enabled by default in CDH 5, requires new memlock functionality from Cloudera Manager 5 Agents.	Do the following: 1. Stop all CDH and managed services. 2. On all hosts with Cloudera Manager Agents, run the command: `$ sudo service cloudera-scm-agent hard_restart` Before performing this step, ensure you understand the semantics of the `hard_restart` command by reading Hard Stopping and Restarting Agents. 3. Start all services.

Configuring Ports for Cloudera Manager

Cloudera Manager, CDH components, managed services, and third-party components use the ports listed in the tables that follow. Before you deploy Cloudera Manager, CDH, and managed services, and third-party components make sure these ports are open on each system. If you are using a firewall, such as iptables, and cannot open all the listed ports, you will need to disable the firewall completely to ensure full functionality.

Ports Used by Cloudera Manager

The following diagram provides an overview of the ports used by Cloudera Manager, Cloudera Navigator, and Cloudera Management Service roles:



For further details, see the following table:

Component	Service	Port	Protocol	Access Requirement	Configuration	Comment
Cloudera Manager Server	HTTP (Web UI)	7180	TCP	External	Administration > Settings > Ports and Addresses	

Component	Service	Port	Protocol	Access Requirement	Configuration	Comment
	HTTPS (Web UI)	7183	TCP	External		Used for HTTPS on master, if enabled. HTTP is the default; only one port is open for either HTTP or HTTPS, not both
	Avro (RPC)	7182	TCP	Internal		Used for Agent to Server heartbeats
	PostgreSQL database managed by cloudera-scm-server-db service	7432	TCP	Internal		The optional embedded PostgreSQL database used for storing configuration information for Cloudera Manager Server.
Cloudera Manager Agent	HTTP (Debug)	9000	TCP	Internal	/etc/cloudera-scm-agent/config.ini	
	Internal supervisord	localhost: 9001	TCP	localhost		supervisord status and control port; used for communication between the Agent and supervisord; only open internally (on localhost)
Event Server	Listens for the publication of events.	7184	TCP	Internal	Cloudera Management Service > Configuration > *ServerName* Default Group > Ports and Addresses	
	Listens for queries for events.	7185	TCP	Internal		
	HTTP (Debug)	8084	TCP	Internal		Allows access to debugging and diagnostic information
Alert Publisher	Internal API	10101	TCP	Internal	Cloudera Management Service > Configuration > *ServerName* Default Group > Ports and Addresses	
Service Monitor	HTTP (Debug)	8086	TCP	Internal	Cloudera Management Service > Configuration > *ServerName* Default Group > Ports and Addresses	
	Listening for Agent messages (private protocol)	9997				

Component	Service	Port	Protocol	Access Requirement	Configuration	Comment
	Internal query API (Avro)	9996				
Activity Monitor	HTTP (Debug)	8087	TCP	Internal	Cloudera Management Service > Configuration > *ServerName* Default Group > Ports and Addresses	
	Listening for Agent messages (private protocol)	9999				
	Internal query API (Avro)	9998				
Host Monitor	HTTP (Debug)	8091	TCP	Internal	Cloudera Management Service > Configuration > *ServerName* Default Group > Ports and Addresses	
	Listening for Agent messages (private protocol)	9995				
	Internal query API (Avro)	9994				
Reports Manager	Queries (Thrift)	5678	TCP	Internal	Cloudera Management Service > Configuration > *ServerName* Default Group > Ports and Addresses	
Cloudera Navigator					Cloudera Management Service > Configuration > *ServerName* Default Group > Ports and Addresses	
Audit Server	HTTP (Debug)	8083	TCP	Internal		
	HTTP (Debug)	7186	TCP	Internal		
Metadata Server	HTTP	7187	TCP	External		

Component	Service	Port	Protocol	Access Requirement	Configuration	Comment
Task Tracker Plug-in (used for activity monitoring)	HTTP (Debug)	localhost: 4867	TCP	localhost		Used only on localhost interface by monitoring agent
Backup and Disaster Recovery	HTTP (Web UI)	7180	TCP	External	Administration > Settings page > Ports and Addresses	Used for communication to peer (source) Cloudera Manager.
	HDFS NameNode	8020	TCP	External	HDFS > Configuration > NameNode Role Group > Ports and Addresses: NameNode Port	HDFS and Hive replication: communication from destination HDFS and MapReduce hosts to source HDFS NameNode(s). Hive Replication: communication from source Hive hosts to destination HDFS NameNode(s).
	HDFS DataNode	50010	TCP	External	HDFS > Configuration > DataNode Role Group(s) > Ports and Addresses: DataNode Transceiver Port	HDFS and Hive replication: communication from destination HDFS and MapReduce hosts to source HDFS DataNode(s). Hive Replication: communication from source Hive hosts to destination HDFS DataNode(s).

Ports Used by Components of CDH 5

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
Hadoop HDFS	DataNode		50010	TCP	External	`dfs.datanode. address`	DataNode HTTP server port
	DataNode	Secure	1004	TCP	External	`dfs.datanode. address`	
	DataNode		50075	TCP	External	`dfs.datanode.http. address`	
	DataNode	Secure	1006	TCP	External	`dfs.datanode.http. address`	
	DataNode		50020	TCP	External	`dfs.datanode.ipc. address`	
	NameNode		8020	TCP	External	`fs.default. name`	`fs.default. name`

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
						or `fs.defaultFS`	is deprecated (but still works)

	NameNode		50070	TCP	External	`dfs.http.address`
						or
						`dfs.namenode.http-address`
							`dfs.http.address`
							is deprecated (but still works)

	NameNode	Secure	50470	TCP	External	`dfs.https.address`
						or
						`dfs.namenode.https-address`
							`dfs.https.address`
							is deprecated (but still works)

	Secondary NameNode		50090	TCP	Internal	`dfs.secondary.http.address`
						or
						`dfs.namenode.secondary.http-address`
							`dfs.secondary.http.address`
							is deprecated (but still works)

	Secondary NameNode	Secure	50495	TCP	Internal	`dfs.secondary.https.address`	
	JournalNode		8485	TCP	Internal	`dfs.namenode.shared.edits.dir`	
	JournalNode		8480	TCP	Internal		
Hadoop MapReduce (MRv1)	JobTracker		8021	TCP	External	`mapred.job.tracker`	
	JobTracker		50030	TCP	External	`mapred.job.tracker.http.address`	
	JobTracker	Thrift Plugin	9290	TCP	Internal	`jobtracker.thrift.address`	Required by Hue and Cloudera Manager Activity Monitor
	TaskTracker		50060	TCP	External	`mapred.task.tracker.http.address`	
	TaskTracker		0	TCP	Localhost	`mapred.task.tracker.report.address`	Communicating with child (umbilical)
Hadoop YARN (MRv2)	ResourceManager		8032	TCP		`yarn.resourcemanager.address`	
	ResourceManager		8030	TCP		`yarn.resourcemanager.scheduler.address`	
	ResourceManager		8031	TCP		`yarn.resourcemanager.`	

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
						`resource-tracker.address`	
	ResourceManager		8033	TCP		`yarn.resourcemanager.admin.address`	
	ResourceManager		8088	TCP		`yarn.resourcemanager.webapp.address`	
	NodeManager		8040	TCP		`yarn.nodemanager.localizer.address`	
	NodeManager		8042	TCP		`yarn.nodemanager.webapp.address`	
	NodeManager		8041	TCP		`yarn.nodemanager.address`	
	MapReduce JobHistory Server		10020	TCP		`mapreduce.jobhistory.address`	
	MapReduce JobHistory Server		19888	TCP		`mapreduce.jobhistory.webapp.address`	
HBase	Master		60000	TCP	External	`hbase.master.port`	IPC
	Master		60010	TCP	External	`hbase.master.info.port`	HTTP
	RegionServer		60020	TCP	External	`hbase.regionserver.port`	IPC
	RegionServer		60030	TCP	External	`hbase.regionserver.info.port`	HTTP
	HQuorumPeer		2181	TCP		`hbase.zookeeper.property.clientPort`	HBase-managed ZK mode
	HQuorumPeer		2888	TCP		`hbase.zookeeper.peerport`	HBase-managed ZK mode
	HQuorumPeer		3888	TCP		`hbase.zookeeper.leaderport`	HBase-managed ZK mode
	REST	REST Service	8080	TCP	External	`hbase.rest.port`	
	REST UI		8085	TCP	External		
	ThriftServer	Thrift Server	9090	TCP	External	Pass `-p <port>` on CLI	

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
	ThriftServer		9095	TCP	External		
		Avro server	9090	TCP	External	`Pass --port <port>` on CLI	
Hive	Metastore		9083	TCP	External		
	HiveServer2		10000	TCP	External	`hive.server2.thrift.port`	
Sqoop	Metastore		16000	TCP	External	`sqoop.metastore.server.port`	
Sqoop 2	Sqoop 2 server		12000	TCP	External		
	Sqoop 2		12001	TCP	External		Admin port
ZooKeeper	Server (with CDH 5 and/or Cloudera Manager 5)		2181	TCP	External	clientPort	Client port
	Server (with CDH 5 only)		2888	TCP	Internal	`X in server.N=host:X:Y`	Peer
	Server (with CDH 5 only)		3888	TCP	Internal	`X in server.N=host:X:Y`	Peer
	Server (with CDH 5 and Cloudera Manager 5)		3181	TCP	Internal	`X in server.N=host:X:Y`	Peer
	Server (with CDH 5 and Cloudera Manager 5)		4181	TCP	Internal	`X in server.N=host:X:Y`	Peer
	ZooKeeper FailoverController (ZKFC)		8019	TCP	Internal		Used for HA
	ZooKeeper JMX port		9010	TCP	Internal		ZooKeeper will also use another randomly selected port for RMI. In order for Cloudera Manager to monitor ZooKeeper, you must open up all ports when the connection originates from the Cloudera Manager server.

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
Hue	Server		8888	TCP	External		
	Beeswax Server		8002		Internal		
	Beeswax Metastore		8003		Internal		
Oozie	Oozie Server		11000	TCP	External	`OOZIE_HTTP_PORT` in `oozie-env.sh`	HTTP
	Oozie Server		11001	TCP	localhost	`OOZIE_ADMIN_PORT` in `oozie-env.sh`	Shutdown port
Spark	Default Master RPC port		7077	TCP	External		
	Default Worker RPC port		7078	TCP			
	Default Master web UI port		18080	TCP	External		
	Default Worker web UI port		18081	TCP			
HttpFS	HttpFS		14000	TCP			
	HttpFS		14001	TCP			

Ports Used by Components of CDH 4

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
Hadoop HDFS	DataNode		50010	TCP	External	`dfs.datanode.address`	DataNode HTTP server port
	DataNode	Secure	1004	TCP	External	`dfs.datanode.address`	
	DataNode		50075	TCP	External	`dfs.datanode.http.address`	
	DataNode	Secure	1006	TCP	External	`dfs.datanode.http.address`	

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
	DataNode		50020	TCP	External	`dfs.datanode.ipc.address`	

	NameNode		8020	TCP	External	`fs.default.name`
						or
						`fs.defaultFS`
							`fs.default.name`
							is deprecated (but still works)

	NameNode		50070	TCP	External	`dfs.http.address`
						or
						`dfs.namenode.http-address`
							`dfs.http.address`
							is deprecated (but still works)

	NameNode	Secure	50470	TCP	External	`dfs.https.address`
						or
						`dfs.namenode.https-address`
							`dfs.https.address`
							is deprecated (but still works)

	Secondary NameNode		50090	TCP	Internal	`dfs.secondary.http.address`
						or
						`dfs.namenode.secondary.http-address`
							`dfs.secondary.http.address`
							is deprecated (but still works)

	Secondary NameNode	Secure	50495	TCP	Internal	`dfs.secondary.https.address`	
	JournalNode		8485	TCP	Internal	`dfs.namenode.shared.edits.dir`	
	JournalNode		8480	TCP	Internal		
Hadoop MRv1	JobTracker		8021	TCP	External	`mapred.job.tracker`	
	JobTracker		50030	TCP	External	`mapred.job.tracker.http.address`	
	JobTracker	Thrift Plugin	9290	TCP	Internal	`jobtracker.thrift.address`	Required by Hue and Cloudera Manager Activity Monitor
	TaskTracker		50060	TCP	External	`mapred.task.tracker.http.address`	
	TaskTracker		0	TCP	Localhost	`mapred.task.tracker.report.address`	Communicating with child (umbilical)
Hadoop YARN	ResourceManager		8032	TCP		`yarn.resourcemanager.address`	
	ResourceManager		8030	TCP		`yarn.resourcemanager.scheduler.address`	

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
	ResourceManager		8031	TCP		yarn.resourcemanager.resource-tracker.address	
	ResourceManager		8033	TCP		yarn.resourcemanager.admin.address	
	ResourceManager		8088	TCP		yarn.resourcemanager.webapp.address	
	NodeManager		8040	TCP		yarn.nodemanager.localizer.address	
	NodeManager		8042	TCP		yarn.nodemanager.webapp.address	
	NodeManager		8041	TCP		yarn.nodemanager.address	
	MapReduce JobHistory Server		10020	TCP		mapreduce.jobhistory.address	
	MapReduce JobHistory Server		19888	TCP		mapreduce.jobhistory.webapp.address	
HBase	Master		60000	TCP	External	hbase.master.port	IPC
	Master		60010	TCP	External	hbase.master.info.port	HTTP
	RegionServer		60020	TCP	External	hbase.regionserver.port	IPC
	RegionServer		60030	TCP	External	hbase.regionserver.info.port	HTTP
	HQuorumPeer		2181	TCP		hbase.zookeeper.property.clientPort	HBase-managed ZK mode
	HQuorumPeer		2888	TCP		hbase.zookeeper.peerport	HBase-managed ZK mode
	HQuorumPeer		3888	TCP		hbase.zookeeper.leaderport	HBase-managed ZK mode
	REST	REST Service	8080	TCP	External	hbase.rest.port	
	REST UI		8085	TCP	External		

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
	ThriftServer	Thrift Server	9090	TCP	External	Pass `-p <port>` on CLI	
	ThriftServer		9095	TCP	External		
		Avro server	9090	TCP	External	Pass `--port <port>` on CLI	
Hive	Metastore		9083	TCP	External		
	HiveServer		10000	TCP	External		
	HiveServer2		10000	TCP	External	`hive.server2.thrift.port`	
Sqoop	Metastore		16000	TCP	External	`sqoop.metastore.server.port`	
Sqoop 2	Sqoop 2 server		12000	TCP	External		
Sqoop 2	Sqoop 2		12001	TCP	External		Admin port
ZooKeeper	Server (with CDH4 and/or Cloudera Manager 4)		2181	TCP	External	clientPort	Client port
	Server (with CDH4 only)		2888	TCP	Internal	`X in server.N=host:X:Y`	Peer
	Server (with CDH4 only)		3888	TCP	Internal	`X in server.N=host:X:Y`	Peer
	Server (with CDH4 and Cloudera Manager 4)		3181	TCP	Internal	`X in server.N=host:X:Y`	Peer
	Server (with CDH4 and Cloudera Manager 4)		4181	TCP	Internal	`X in server.N=host:X:Y`	Peer
	ZooKeeper FailoverController (ZKFC)		8019	TCP	Internal		Used for HA
	ZooKeeper JMX port		9010	TCP	Internal		ZooKeeper will also use another randomly selected port for RMI. In order for Cloudera Manager to monitor ZooKeeper, you must open up all ports when the connection

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
							originates from the Cloudera Manager server.
Hue	Server		8888	TCP	External		
	Beeswax Server		8002		Internal		
	Beeswax Metastore		8003		Internal		
Oozie	Oozie Server		11000	TCP	External	`OOZIE_HTTP_PORT` in `oozie-env.sh`	HTTP
	Oozie Server		11001	TCP	localhost	`OOZIE_ADMIN_PORT` in `oozie-env.sh`	Shutdown port

Ports Used by Cloudera Impala

Impala uses the TCP ports listed in the following table. Before deploying Impala, ensure these ports are open on each system.

Component	Service	Port	Access Requirement	Comment
Impala Daemon	Impala Daemon Frontend Port	21000	External	Used to transmit commands and receive results by `impala-shell`, Beeswax, and version 1.2 of the Cloudera ODBC driver.
Impala Daemon	Impala Daemon Frontend Port	21050	External	Used to transmit commands and receive results by applications, such as Business Intelligence tools, using JDBC and the version 2.0 or higher of the Cloudera ODBC driver.
Impala Daemon	Impala Daemon Backend Port	22000	Internal	Internal use only. Impala daemons use to communicate with each other.
Impala Daemon	StateStoreSubscriber Service Port	23000	Internal	Internal use only. Impala daemons listen on this port for updates from the state store.

Component	Service	Port	Access Requirement	Comment
Impala Daemon	Impala Daemon HTTP Server Port	25000	External	Impala web interface for administrators to monitor and troubleshoot.
Impala StateStore Daemon	StateStore HTTP Server Port	25010	External	StateStore web interface for administrators to monitor and troubleshoot.
Impala Catalog Daemon	Catalog HTTP Server Port	25020	External	Catalog service web interface for administrators to monitor and troubleshoot. New in Impala 1.2 and higher.
Impala StateStore Daemon	StateStore Service Port	24000	Internal	Internal use only. State store listens on this port for registration/unregistration requests.
Impala Catalog Daemon	StateStore Service Port	26000	Internal	Internal use only. The catalog service uses this port to communicate with the Impala daemons. New in Impala 1.2 and higher.
Impala Daemon	Llama Callback Port	28000	Internal	Internal use only. Impala daemons use to communicate with Llama. New in CDH 5.0.0 and higher.
Impala Llama ApplicationMaster	Llama Thrift Admin Port	15002	Internal	Internal use only. New in CDH 5.0.0 and higher.
Impala Llama ApplicationMaster	Llama Thrift Port	15000	Internal	Internal use only. New in CDH 5.0.0 and higher.
Impala Llama ApplicationMaster	Llama HTTP Port	15001	External	Llama service web interface for administrators to monitor and troubleshoot. New in CDH 5.0.0 and higher.

Ports Used by Cloudera Search

Component	Service	Port	Protocol	Access Requirement	Comment
Cloudera Search	Solr search/update	8983	http	External	All Solr-specific actions, update/query.
Cloudera Search	Solr (admin)	8984	http	Internal	Solr administrative use.

Ports Used by Third-Party Components

Component	Service	Qualifier	Port	Protocol	Access Requirement	Configuration	Comment
Ganglia	ganglia-gmond		8649	UDP/TCP	Internal		
	ganglia-web		80	TCP	External	Via Apache httpd	
Kerberos	KRB5 KDC Server	Secure	88	UDP/TCP	External	kdc_ports and kdc_tcp_ports in either the [kdcdefaults] or [realms] sections of kdc.conf	By default only UDP
	KRB5 Admin Server	Secure	749	TCP	Internal	kadmind_port in the [realms] section of kdc.conf	