Dell | Cloudera Solution User's Guide v.2.4

A Dell User's Guide for Apache™ Hadoop® Deployment
Crowbar v1.6

January 15, 2014

The power to do more

Table of Contents

Figures

Tables

Trademarks

January 15, 2014

Notes, Cautions, and Warnings

A **NOTE** indicates important information that helps you make better use of your computer.

A **CAUTION** indicates potential damage to hardware or loss of data if instructions are not followed.

A **WARNING** indicates a potential for property damage, personal injury, or death.

Abbreviations

Abbreviation	Definition
BMC	Baseboard Management Controller.
DBMS	Database management system.
EDW	Enterprise data warehouse.
EoR	End-of-row switch/router.
HDFS	Hadoop Distributed File System.
IPMI	Intelligent Platform Management Interface.
LAG	Link aggregation group.
LOM	Local Area Network on Motherboard.
NIC	Network interface card.
ToR	Top-of-rack switch/router.

Introduction

This document provides instructions you to use when deploying Cloudera Manager and Apache Hadoop Ecosystem components with Crowbar. This guide is for use with the *Dell Crowbar Software Framework Users Guide*, and is *not* a stand-alone document. It specifically covers Cloudera Manager, Apache Hadoop and the deployment steps from a Crowbar prospective. Please refer to the *Dell Crowbar Software Framework User's Guide* for assistance with installing common Crowbar components and configuring the target systems.

> Concepts beyond the scope of this guide are introduced as needed in notes and references to other documentation.

Overview

Hadoop is an Apache project being built and used by a global community of contributors, written in the Java programming language. Yahoo! has been the largest contributor to the project, and uses Hadoop extensively across its businesses. Other contributors and users include Facebook, LinkedIn, eHarmony, and eBay. Cloudera has created a quality controlled distribution of Hadoop and offers commercial management software, support, and consulting services.

Dell developed a solution for Hadoop that includes optimized hardware, software, and services to streamline deployment and improve the customer experience.

The Dell | Cloudera Solution is based on the Cloudera CDH Enterprise distribution of Hadoop. **Dell's solution** includes:

- Dell Reference architecture (RA) and best practices documentation.

- Optimized hardware and network infrastructure.

- Cloudera CDH software (CDH Community-provided for customer-deployed solutions).

- Cloudera Manager free edition with the ability to upgrade to enterprise level via Cloudera issued license key.

- Cloudera Manager provided Hadoop infrastructure management tools.

- Dell Crowbar software framework.

This solution provides Dell a foundation to offer additional solutions as the Hadoop environment evolves and expands.

Document Scope

The focus of this guide is the use of Crowbar, *not* Apache Hadoop or Cloudera Manager. While Crowbar includes substantial components to assist in the deployment of Apache Hadoop and Cloudera Manager, its operational aspects are completely independent. For more detailed information, please refer to the following links:

Cloudera Manager 4.8 Documentation

- http://www.cloudera.com/content/support/en/documentation/manager/cloudera-manager-v4-latest.html

CDH4 Documentation

- http://www.cloudera.com/content/support/en/documentation/cdh4-documentation/cdh4-documentation-v4-latest.html

Apache Hadoop Documentation

- http://hadoop.apache.org/

cloudera This guide provides this additional information about Cloudera as notes flagged with the Cloudera logo. For detailed operational support for Hadoop, we suggest visiting the Cloudera documentation web site at http://www.cloudera.com.

Opscode Chef Server

Crowbar makes extensive use of Opscode Chef Server**,** http://opscode.com. To explain Crowbar actions, you should understand the underlying Chef implementation. This guide provides this additional Chef information as notes flagged with the Opscode logo.

To use Crowbar, it is not necessary to log into the Chef Server; consequently, use of the Chef UI is not covered in this guide. Supplemental information about Chef is included.

Crowbar is not limited to managing Dell servers and components. Due to driver requirements, some barclamps, for example: BIOS and RAID must be targeted to specific hardware; however, those barclamps are not required for system configuration.

Dell | Cloudera Solution

This section provides detailed information about the basics of Hadoop, and Hadoop components deployment.

Hadoop Basics

The Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programmatic driven processing model. Hadoop is designed to scale up from a minimum of three servers to thousands of machines, each offering local computation and storage.

Rather than rely on hardware to deliver high-availability, the Hadoop library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on a cluster of computers, each of which may be prone to failures.

Hadoop is ideal for organizations with a growing need to store and process massive application datasets. It enables applications to work with thousands of nodes and petabytes of data.

- **Hadoop Core**: The common libraries and utilities that provide the basic Hadoop runtime environment. A set of components and interfaces which implement a distributed filesystem and provide general I/O access for the Hadoop framework (serialization, Java RPC and persistent data storage).
- **Hadoop Distributed File System (HDFS)**: A distributed file system that provides redundant, high-throughput access to application data.
- **MapReduce**: A software framework for distributed processing of large data sets on compute clusters.

Apache Hadoop Component Deployment

Cloudera Manager and Pig employ Crowbar tools to construct a starting proposal, and then edit any parameters to fit the specific needs of your environment. Once the proposal is ready, apply the proposal to deploy each system components.

> The Base Hadoop system (HDFS and Map Reduce), YARN, Zookeeper, HBase, Oozie, Hive, Hue, Flume, Impala, Sqoop, and Solr are deployed using the Cloudera Manager administration console. Crowbar also provides a supplemental Hadoop Ecosystem Barclamp (Pig). You must install the base Hadoop system (HDFS and Map Reduce) using Cloudera Manager before deploying any of these add-ons.

Table 1: Supported Apache Hadoop Components

Component	Deployment Method	Description
HDFS	Cloudera Manager	Apache Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute hosts throughout a cluster to enable reliable, extremely rapid computations.
MapReduce	Cloudera Manager	Apache Hadoop MapReduce supports distributed computing on large data sets across your cluster (requires HDFS).
YARN	Cloudera Manager	Apache Hadoop MapReduce 2.0 (MRv2), or YARN, is a data computation framework that supports MapReduce applications (requires HDFS). The current upstream MRv2 release is not yet considered stable and should not be considered production-ready at this time.
ZooKeeper	Cloudera Manager	Apache ZooKeeper is a centralized service for maintaining and synchronizing configuration data.

Component	Deployment Method	Description
HBase	Cloudera Manager	HBase is an open-source, non-relational, distributed database modeled after Google's BigTable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed Filesystem), providing BigTable-like capabilities for Hadoop. That is, it provides a fault-tolerant way of storing large quantities of sparse data. HBase features compression, in-memory operation, and Bloom filters on a per-column basis as outlined in the original BigTable paper. Tables in HBase can serve as the input and output for MapReduce jobs run in Hadoop, and may be accessed through the Java API but also through REST, Avro or Thrift gateway APIs. HBase is not a direct replacement for a classic SQL Database, although recently its performance has improved, and it is now serving several data-driven websites, including Facebook's Messaging Platform.
Hive	Cloudera Manager	Hive is a data warehouse system that offers a SQL-like language called HiveQL.
Oozie	Cloudera Manager	Oozie is a workflow coordination service to manage data processing jobs on your cluster.
Hue	Cloudera Manager	Hue is a graphical user interface to work with Cloudera's Distribution Including Apache Hadoop (requires HDFS, MapReduce, and Hive).
Flume	Cloudera Manager	Flume collects and aggregates data from almost any source into a persistent store such as HDFS.
Impala	Cloudera Manager	Impala provides a real-time SQL query interface for data stored in HDFS and HBase. Impala requires Hive service and shares Hive Metastore with Hue.
Sqoop	Cloudera Manager	Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases. The version supported by Cloudera Manager is Sqoop 2.
Solr	Cloudera Manager	Solr is a distributed service for indexing and searching data stored in HDFS.
Pig	Crowbar Barclamp	Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data algorithms.

For more information about Hadoop, please visit http://hadoop.apache.org/.

Crowbar User Interface

Crowbar is delivered as a Web application available on the admin node using HTTP on port 3000. By default, you can access it using http://192.168.124.10:3000. Additionally, the default installation contains an implementation of Hadoop specific components (see table below).

> Dell supports running Crowbar on the following browsers: Firefox 3.6, Firefox 11, Google Chrome, Internet Explorer 8, and Internet Explorer 9. HTML5 compatibility and a minimum screen resolution of 1024x768 are recommended.

Table 2: User Interface Service URLs

User Interface Service	Default Location	Port	Example URL
Crowbar	Crowbar Admin Node	3000	http://<crowbar_admin_node>:3000
Cloudera Manager	Hadoop Edge Node	7180	http://<cloudera_manager_server_node>:7180
Hadoop Name Node	Hadoop Name Node	50070	http://<master_name_node>:50070
Hadoop Secondary Name Node	Hadoop Secondary Name Node	50090	http://<secondary_name_node>:50090
Hadoop Data Node	Hadoop Data Node	50075	http://<data_node>: 50075
Hadoop Job Tracker Web	Hadoop Job Tracker Node	50030	http://<job_tracker_node>: 50030
Hadoop Task Tracker Web	Task Tracker Node	50060	http://<task_tracker_node>:50060

> The crowbar admin node IP address (192.168.124.10) is the default address. Replace it with the address assigned to the Crowbar Admin node. Nagios, Ganglia and Chef can be accessed directly from a web browser or via selecting one of the links on the Crowbar Dashboard.

Cloudera Manager Overview

Cloudera Manager deploys and centrally operates a complete Hadoop stack. The application automates the installation process, reducing deployment time from weeks to minutes, gives you a cluster-wide, real time view of the services running and the status of their hosts, provides a single, central place to enact configuration changes across your cluster; and incorporates a full range of reporting and diagnostic tools to help you optimize cluster performance and utilization. Cloudera Manager provides full lifecycle management for Hadoop deployments**.**

Functionality Outline

- Installs the complete Hadoop stack in minutes via a wizard-based interface
- Gives you complete, end-to-end visibility and control over your Hadoop cluster from a single interface
- Enables you to set server roles and configure services across the cluster
- Enables you to gracefully start, stop and restart of services as needed
- Shows information pertaining to hosts in your cluster including status, resident memory, virtual memory and roles

Table 3: Cloudera Manager Standard and Cloudera Enterprise Differences

Feature	Cloudera Standard (Free Edition)	Cloudera Enterprise (60-Day Trial)	Cloudera Enterprise (Licensed Edition)
CDH FEATURES			
Hadoop	✅	✅	✅
Flume	✅	✅	✅
Hive	✅	✅	✅
Mahout	✅	✅	✅
Oozie	✅	✅	✅
Pig	✅	✅	✅
Sqoop	✅	✅	✅
Whirr	✅	✅	✅
Zookeeper	✅	✅	✅
Hue	✅	✅	✅
HBase	✅	✅	✅
Impala	✅	✅	✅
Search	✅	✅	✅
CLOUDERA MANAGER FEATURES			
Deployment & Configuration	✅	✅	✅
Service Management	✅	✅	✅
Service & Host Monitoring	✅	✅	✅
Diagnostics	✅	✅	✅
API	✅	✅	✅
Rolling Updates/Restarts	⚠	✅	✅
SNMP Support	⚠	✅	✅

Feature	Cloudera Standard (Free Edition)	Cloudera Enterprise (60-Day Trial)	Cloudera Enterprise (Licensed Edition)
LDAP Integration	⚠	✅	✅
Configuration History & Rollbacks	⚠	✅	✅
Operational Reports	⚠	✅	✅
Automated Disaster Recovery	⚠	✅	BDR Add-on
CLOUDERA NAVIGATOR FEATURES			
Data Audit – HDFS, Hbase & Hive	⚠	✅	Navigator Add-on
Access Management	⚠	✅	Navigator Add-on
TECHNICAL SUPPORT AND INDEMNITY			
Core Projects	⚠	⚠	✅
Apache HBase	⚠	⚠	RTD Add-on
Cloudera Impala	⚠	⚠	RTQ Add-on
Cloudera Search	⚠	⚠	RTS Add-on
Cloudera Manager	⚠	⚠	✅
Cloudera Navigator	⚠	⚠	Navigator Add-on

Barclamps

> Best practice is to reboot a node whenever a barclamp proposal is applied or updated.

Table 4: Barclamp Descriptions

Barclamp	Description
Hadoop Infrastructure	Provides the basic runtime environment for Hadoop cluster deployment.
Cloudera Manager	Provides end-to-end management for apache Hadoop with the ability to deploy and centrally operate a complete Hadoop stack gives you a cluster wide, real time view of nodes and services running and provides a single central place to enact configuration changes across your cluster. Cloudera Manager incorporates a full range of reporting and diagnostic tools to help you optimize cluster performance and utilization.
Pig	Platform for analyzing large data sets that consists of a high-level language for expressing data algorithms.

Hadoop Infrastructure Barclamp

The Hadoop Infrastructure Barclamp performs all the low level operating system configuration setup for the Hadoop cluster.

> Although Crowbar makes intelligent guesses to preconfigure the node assignments, they may not be optimal for your environment. You can click on the **Remove Node** icon to remove any node from a role.

Installing the Hadoop Infrastructure Barclamp

1. Navigate to the Crowbar interface using a Web browser. Typically, the IP address is http://192.168.124.10:3000.

    a. Username is **crowbar**; password is **crowbar**.

2. Click on the **Barclamps** tab, and then select **Hadoop**.

3. Select the **Hadoop Infrastructure** barclamp, and then click on the **Create** button.

4. In the *Edit Proposal* screen, select **true** from the *Barclamp* > *Log Debug Messages* drop-down.

5. Scroll down to the *Node Deployment* section.

6. Drag and drop nodes from the *Available Nodes* column to their proper roles:

> Ensure that you drag the nodes' **names**, not the link icons.

    a. **Clouderamanager-cb-adminnode** - Preconfigured with the Crowbar Admin Node

> ⚠ This node contains software repositories used by all other nodes. Do not attempt to store repositories elsewhere, as unpredictable results may occur.

    b. **Clouderamanager-server** - Dell recommends that you use the Edge Node

c. **Clouderamanager-namenode** - The primary and secondary Name Nodes

d. **Clouderamanager-datanode** - The Data Nodes

e. **Clouderamanager-edgenode** - The Edge Node

f. **Clouderamanager-ha-journaling node** - The Quorum-based Journaling Node

g. **Clouderamanager-ha-filernode** - The High-availability Filer Node

> You can select only one type of high availability – Quorum-based Journaling or Filer. They are mutually exclusive. Dell recommends that you use Quorum-based Journaling.

7. Click the **Apply** button to commit the barclamp proposal to your nodes.

8. Return to the *Nodes* > *Dashboard* screen.

a. Once all icons are green, the barclamp proposal has been applied.

b. You can view the process of the proposal for each node by viewing their consoles via SSH sessions.

Table 5: Barclamp Parameters

Name	Description	Required	Default
Log Debug Messages	Enable log debug messages (/var/log/chef/client.log).	true	false

Table 6: Operating System Parameters

Name	Description	Required	Default
File System Type	File system type (ext3/ext4).	true	ext4
THP Compaction	Controls the usage of Transparent Huge Pages (THP) Compaction. • **never**: THP Compaction is disabled • **always**: THP Compaction is enabled **Note**: Leave this parameter at the default setting for best performance.	true	never
Map/Reduce File Handles	Maximum number of Map/Reduce open file handles.	true	32768
HDFS File Handles	Maximum number of HDFS open file handles.	true	32768
HBASE File Handles	Maximum number of HBASE open file handles.	true	32768

Table 7: Hadoop High Availability Parameters (HA Filer)

Name	Description	Required	Default
Shared Edits Directory	Specifies the HA shared edits directory.	true	/dfs/ha
Shared Edits Export Options	Specifies the HA shared edits export options.	true	rw,async,no_root_squash,no_subtree_check
Shared Edits Mount Options	Specifies the HA shared edits mount options.	true	rsize=65536,wsize=65536,intr,soft,bg

Cloudera Manager Installation Overview

This section briefly describes the automatic and manual installation processes.

Automatic Installation

An automatic installation will only install Cloudera Manager with the basic options configured. You can then install additional features, based upon your usage requirements, from the Cloudera Manager user interface.

> 🖉 If after an automatic installation you require more Cloudera flexibility and features than are provided by the automatic installation, you must then install them manually via Cloudera Manager after the automatic installation process has completed.

Manual Installation

A manual installation enables you to install any Cloudera Manager features for which you are licensed.

Installing the Cloudera Manager Barclamp

1. Navigate to the Crowbar interface using a Web browser. Typically, the IP address is http://192.168.124.10:3000.

   a. Username is **crowbar**; password is **crowbar**.

2. Click on the **Barclamps** tab, and then select **Hadoop**.

3. Select the **Clouderamanager** barclamp, and then click on the **Create** button.

4. In the *Edit Proposal* screen, select **true** from the *Barclamp* > *Log Debug Messages* drop-down.

5. Select the *Deployment Type* from the dropdown. You can select from:

   - **auto**

   - **manual** (the default)

> 🖉 The Cloudera Manager API parameters in the *Edit Proposal* screen are relevant only if you select **manual** as the Deployment Type. If you select the default **auto** they are ignored, and no further action is required for them.

6. Optionally, you can enter a purchased Cloudera Manager Enterprise license key in the **Cloudera Manager License Key (optional)** field.

> 🖉 You must input the **entire** key from the file that you received from Cloudera. This is not a selection box.

      a.   You can also enter the key later in the Cloudera Manager user interface.

7.   Click the **Apply** button to commit the barclamp proposal to your nodes.

8.   Return to the *Nodes* > *Dashboard* screen.

      a.   Once all icons are green, the barclamp proposal has been applied.

      b.   You can view the process of the proposal for each node by viewing their consoles via SSH sessions.

9.   Optionally, you can install the Pig barclamp. See Pig Barclamp on page 17 for more information.

10.   Reboot the nodes.

📝 It may take some time for all node icons to return to a green "Ready" status.

Once the Clouderamanager barclamp proposal has successfully applied, you can log into the Cloudera Manager user interface.

Local HDFS Access

Data nodes, by default, have local access to the HDFS filesystem. If, after a manual Cloudera Manager installation, you require other nodes in addition to the data nodes to have local access to the HDFS filesystem, you must:

- Add the HDFS Gateway role to those nodes

- Deploy the client configuration from Cloudera Manager

See Appendix A: Post-Installation Gateway Role Assignments for instructions.

Table 8: Cloudera Manager API Parameters

Name	Description	Required	Default
Deployment Type	Specifies the deployment options. • **Auto**: Crowbar preconfigures the initial Hadoop cluster, host, role, and service settings according to the Crowbar-deployed cluster configuration. This will only be applied during the initial cluster setup; any following Hadoop cluster configuration changes must be made from the Cloudera Manager user interface. • **Manual**: You must completely configure the deployed Hadoop cluster manually via the Cloudera Manager user interface.	true	manual

Name	Description	Required	Default
Server Port	Indicates the port upon which the Cloudera Manager server API communicates.	true	7180
User Name	Indicates the Cloudera Manager administrative login username.	true	admin
Password	Indicates the Cloudera Manager administrative login user's password	true	admin
Use TLS (https)	Specifies whether or not the Cloudera Manager server uses TLS cryptography over HTTPS.	true	false
API Version	Indicates the Cloudera Manager API version. This is a read-only field and cannot be changed.	true	2

Table 9: CM Server Parameters

Name	Description	Required	Default
Database Type	Indicates the backing database technology.	true	postgresql

Table 10: Cluster Parameters

Name	Description	Required	Default
Cluster Name	Indicates the name of the cluster.	true	cluster01
CDH Version	Indicates the CDH version in use.	true	CDH4
Cloudera Manager License Key (optional)	If you have a Cloudera Manager License key, you can paste it into this field to activate Cloudera Manager Enterprise level functions upon cluster deployment. You can also use the Cloudera Manager user interface to enter the license key at a later date. This option is located at the Cloudera Manager *Administration* > *License* menu pull-down.	false	N/A

Pig Barclamp

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for **evaluating** these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

Pig's infrastructure layer consists of a compiler that produces sequences of MapReduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:

- **Ease of programming:** It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.

- **Optimization opportunities:** The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.

- **Extensibility:** Users can create their own functions to do special-purpose processing.

Table 11: Pig Barclamp Parameters

Name	Description	Required	Default
java_home	JAVA_HOME environment variable.	true	/usr/java/jdk1.6.0_31/jre
log4jconf	log4jconf log4j configuration file.	true	./conf/log4j.properties
brief	brief logging - no timestamps.	true	false
cluster	Clustername, name of the hadoop jobtracker. If no port is defined port 50020 will be used.	false	
debug_level	Debug level, INFO is default.	true	INFO
file	A file that contains pig script.	false	
jar	Load jarfile, colon separated.	false	
verbose	Verbose print all log messages to screen (default to print only INFO and above to screen).	true	false
exectype	Exectype local or mapreduce - mapreduce is default.	true	mapreduce
ssh_gateway	HOD gateway property.	false	
hod_expect_root	HOD expect root property.	false	
hod_expect_uselatest	HOD use latest root property.	false	
hod_command	HOD command root property.	false	
hod_config_dir	HOD config directory property.	false	
hod_param	HOD param property.	false	
pig_spill_size_threshold	Do not spill temp files smaller than this size (bytes).	true	5000000
pig_spill_gc_activation_size	EXPERIMENT: Activate garbage collection when spilling a file bigger than this size (bytes). This should help reduce the number of files being spilled.	true	40000000
log_file	Log file location.	false	

Cloudera Manager Node Inventory Page

Once the Cloudera barclamp has been deployed, from the Edit Proposal page, there is a link below the Proposal **Attributes section called "Cloudera Manager Nodes." Clicking on this link will display a page titled "Node Inventory." This screen is pictured in** the figure below. You can print this page, as it will be very useful during the Cloudera Manager installation to ensure the correct nodes are selected for their intended Cloudera Manager roles.

Figure 1: Node Inventory Screen
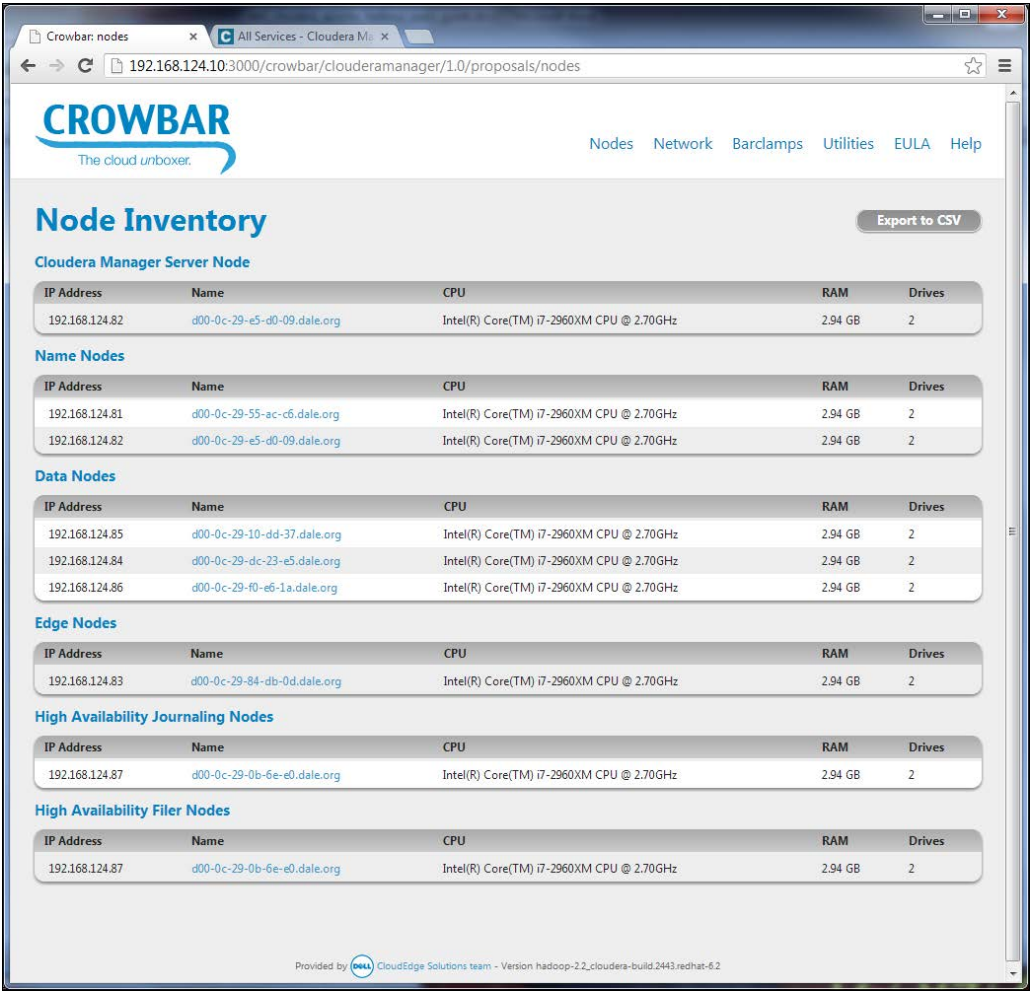
✍️ You can **also export this data to a comma separated value file by selecting the "Export to CSV" button at the top of the page**.

Cloudera Manager Administration Console

> ✎ Dell has tested running the Cloudera Manager Administration console on the following browsers: Firefox 3.6, Firefox 11, Google Chrome, Internet Explorer 8, and Internet Explorer 9.

To start the Cloudera Manager Administration Console:

1. In a web browser, enter the following URL: http(s): // IP_ADDRESS: PORT_NUMBER.

    a. IP_ADDRESS is the name or IP address of the host machine where the Cloudera Manager Web Server is installed. The default machine is the Edge node.

    b. PORT_NUMBER is the default port number (7180).

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

    a. Username: admin

    b. Password: admin

    c. You can also access the Cloudera Manager Administration Console from the Crowbar User Interface using the link located on the crowbar admin node view page (Cloudera Manager).
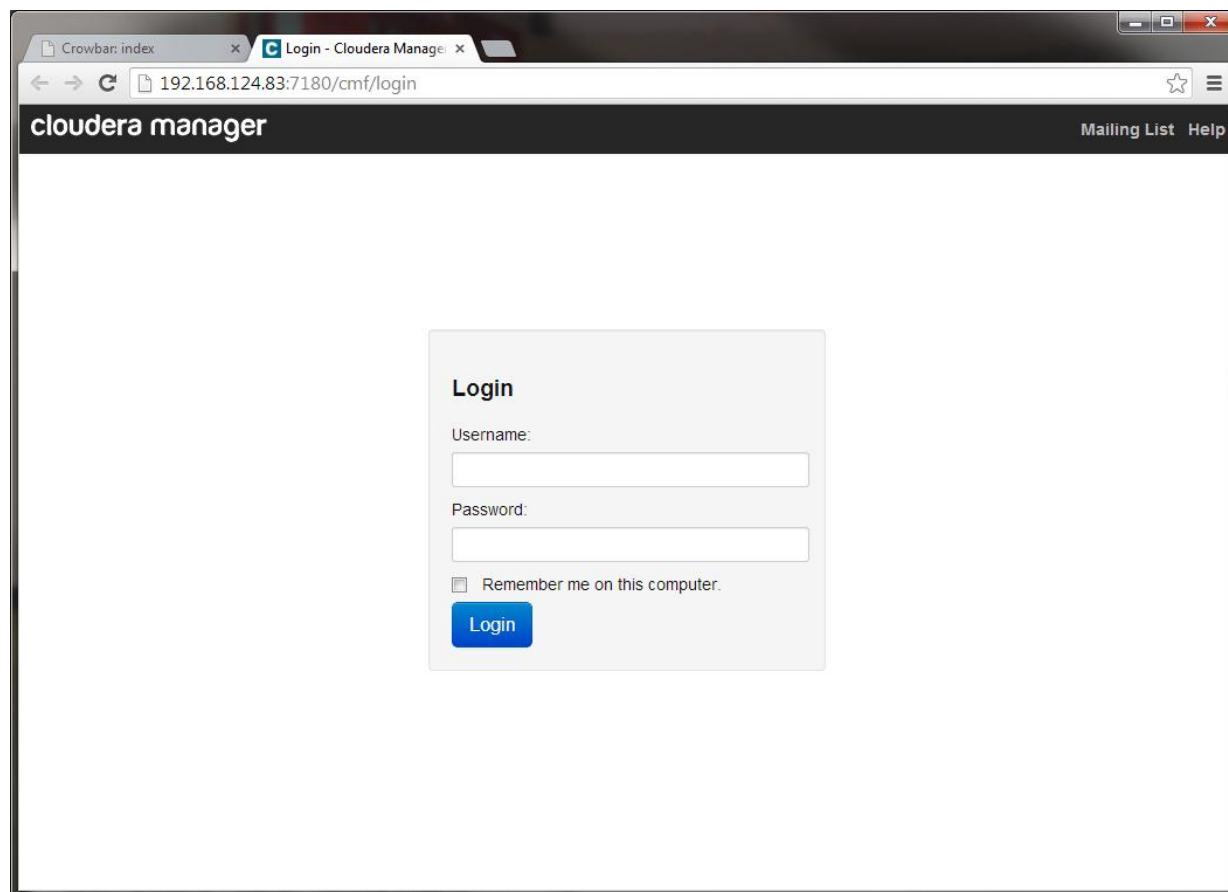
> ✎ For security, you should change the password for the default admin user account as soon as possible. This option is available from the Cloudera Manager application, under the **Administration**->**Password** tab.

Login Screen

1. Enter the user login name and password (default=admin, admin).

2. If you want to save the password, enable the **Remember me on this computer** checkbox.

3. Click the **Login** button to proceed.

Figure 2: Login Screen

Select Edition Screen

This screen enables you to select one of the following Clouder Manager editions:

- **Cloudera Standard - A free edition with limited features.**
- **Cloudera Enterprise Trial - A free, 60-day trial of the full-featured Cloudera Enterprise edition. After 60 days the trial will expire, and the product will continue to function as Cloudera Standard.**
- **Cloudera Enterprise - The full Cloudera Enterprise product. This edition requires a paid, annual license.**

1. Click on the column for the product you wish to install. That column becomes highlighted.

    a. Or, if you wish to use the Cloudera Enterprise Trial Edition, click the **Continue** button to proceed.

2. If you have obtained a Cloudera Manager License key and you wish to upgrade to the Cloudera Manager Enterprise Edition, you can enter the license key.

    a. Click the **Upload License** button.

    b. A file browser window appears, enabling you to select a license key file.

    c. Click the **Upload** button to apply the license key.

    d. Click the **Continue** Button to proceed after the license key has been applied.

Applying the license key is an optional step; you can always enter the license key later on in the process by clicking on the **Administration->License** link in the Cloudera Manager user interface.

Figure 3: Select Edition Screen
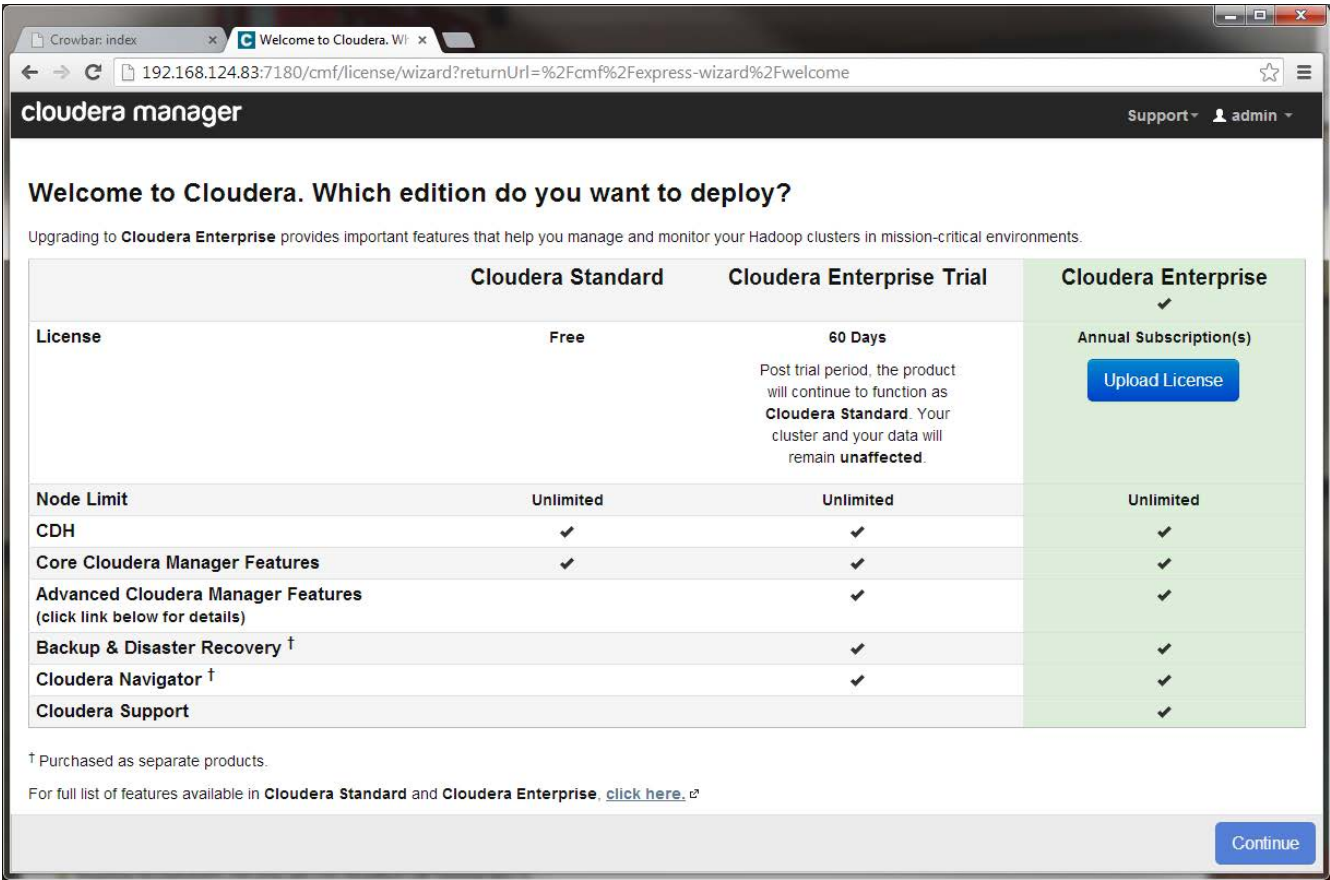
License Key Restart Screen

1. Once the license key has been uploaded, the Cloudera Manager application will ask you to restart the Cloudera Manager server in order for it to take effect. You need to open an SSH console on the node which has the Cloudera Manager (*clouderamanager-server*) role applied to it (login=root/crowbar) and execute the following commands:

```
# service cloudera-scm-server restart
```

2. Once the Cloudera manager server has been restarted, you need to log back into the Cloudera Manager user Interface to proceed.

Figure 4: License Key Restart Screen
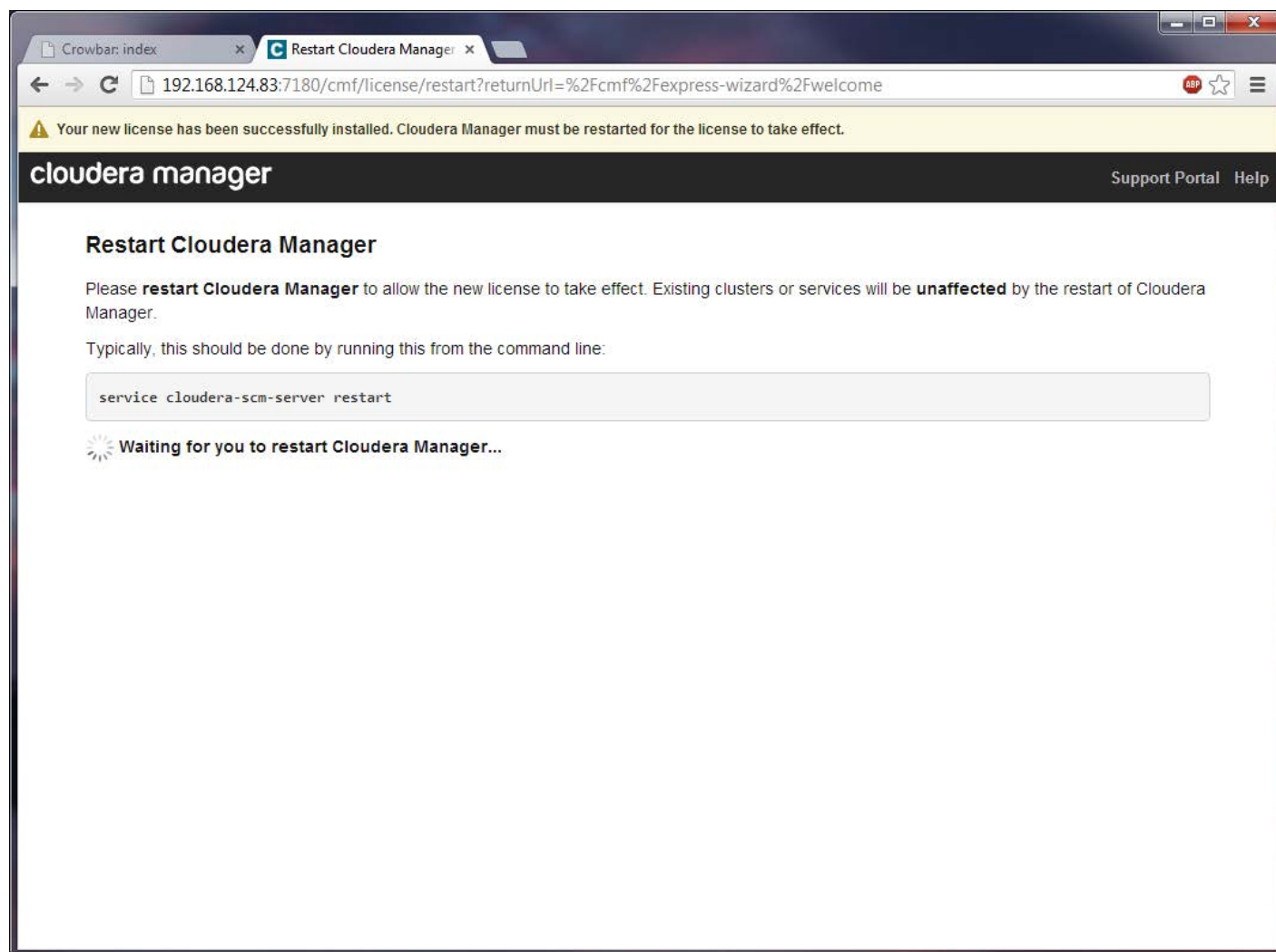
- Upon restarting the service, the screen message transitions from "*Waiting for you to restart Cloudera Manager …*" to "*Restarting …*"

The User interface refreshes to the Login screen.

- Login with username **admin** and password **admin**.

License Key Confirmation Screen

If you have entered the Cloudera Manager License key, you will see this additional screen.

- Click the **Continue** Button to proceed.

Figure 5: License Key Confirmation Screen
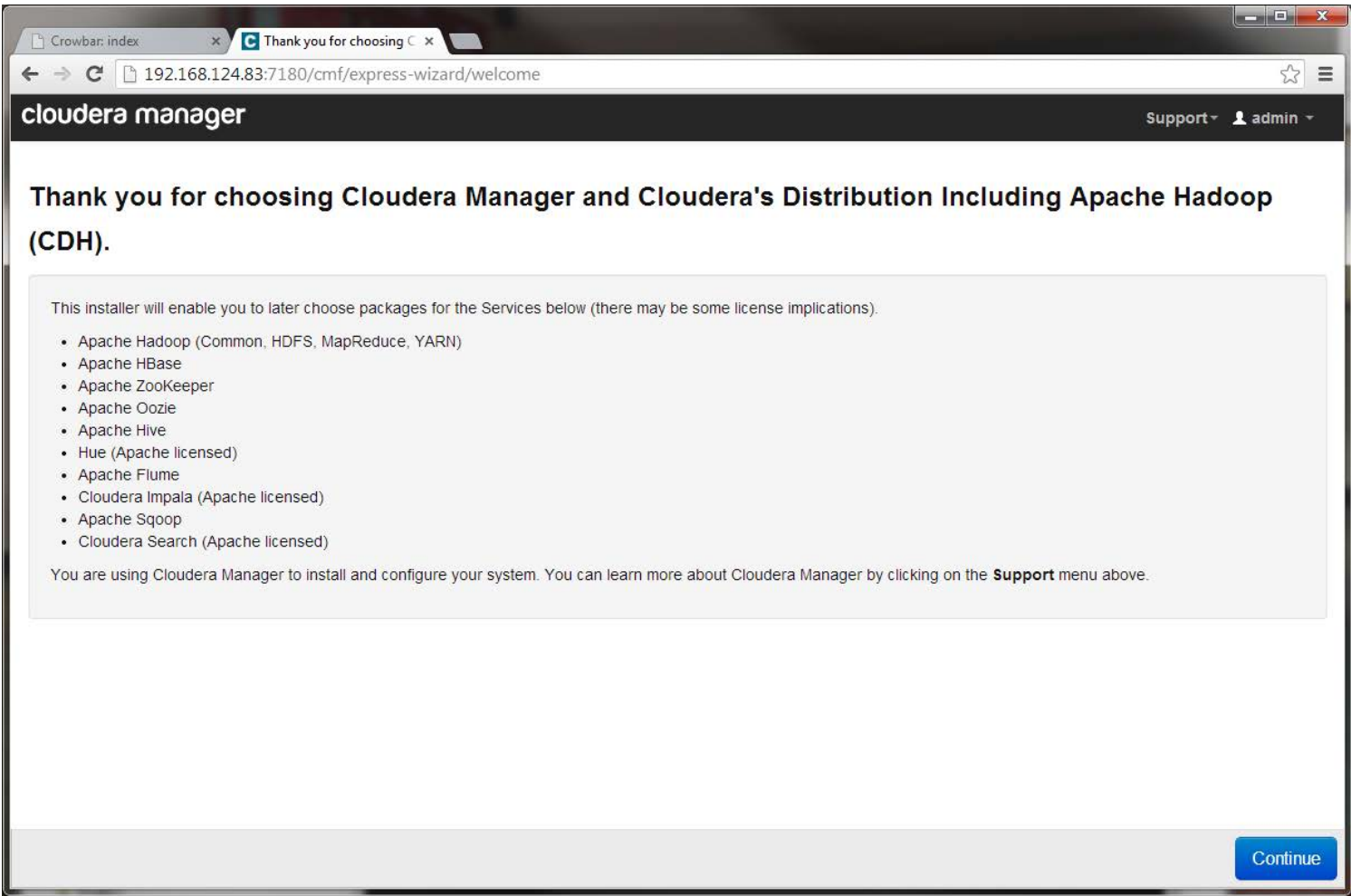
Node Search Screen

1. Enter the IP range or hostname search pattern for all Hadoop cluster nodes. Cloudera Manager will search the cluster using this pattern and will consider any node with a Cloudera Manager agent process running on it as a valid Hadoop node candidate. For example;

   - 192.168.124.[80-90] will attempt to discover all the nodes between 192.168.124.80 and 192.168.124.90

   - 192.168.124.8[1-3] will attempt to discover 192.168.124.81, 192.168.124.82, and 192.168.124.83

   - For additional information on Cloudera Manager search patterns, see the search for hostnames and/or IP addresses using patterns link on the Cloudera Manager user Interface.

2. **Optionally, enter the host's SSH Port**. The default port is 22.

3. Click the **Search** button to proceed.

Figure 6: Cloudera Cluster Node Search Screen

Node Search Results Screen

1. Verify that all your Hadoop nodes have been discovered.

2. Make any cluster configuration adjustments by selecting or deselecting any checkboxes.

3. Click the **Continue** button to proceed.

Figure 7: Node Search Results Screen
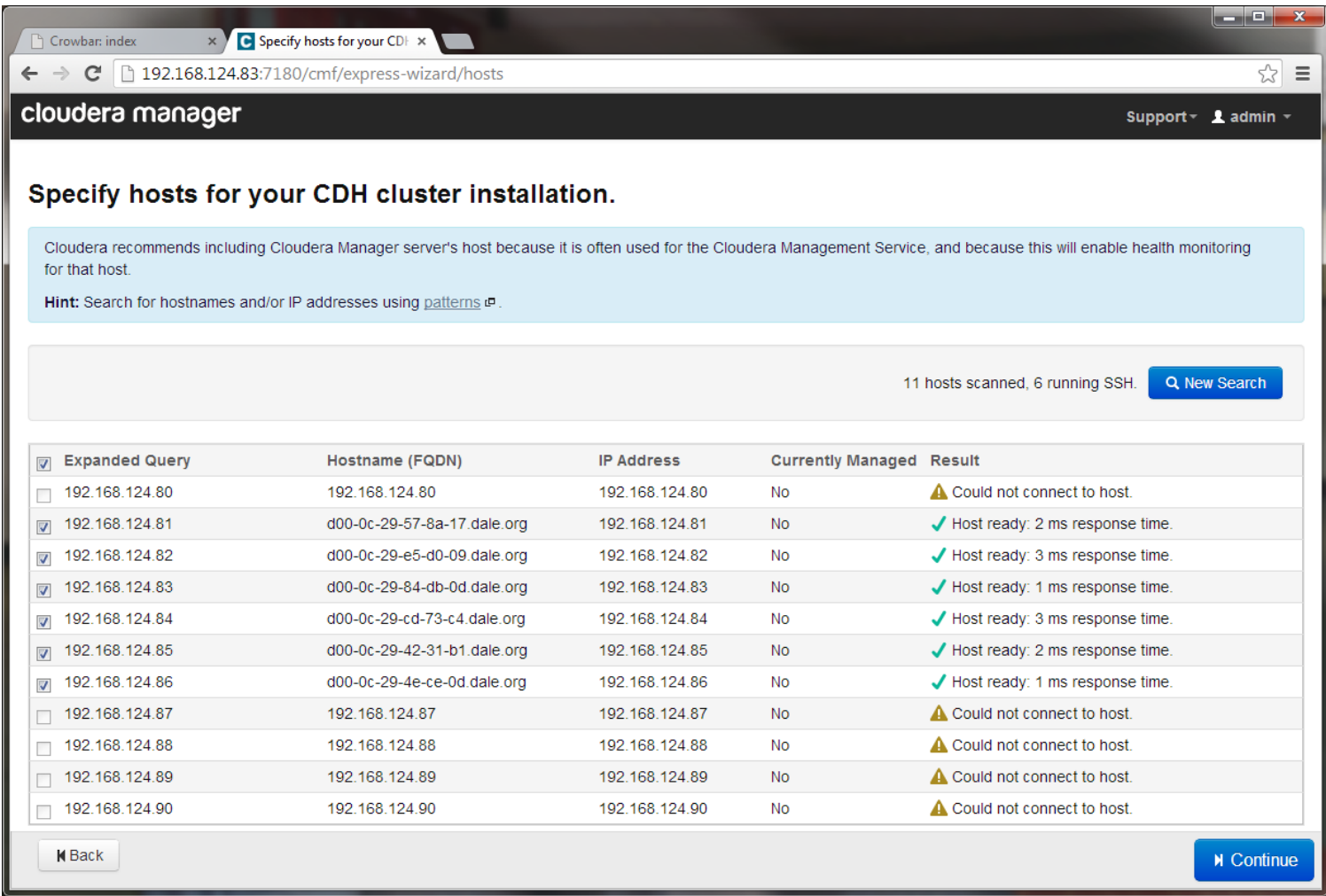
Select Repository Screen

- Select **Use Packages** as the installation method.
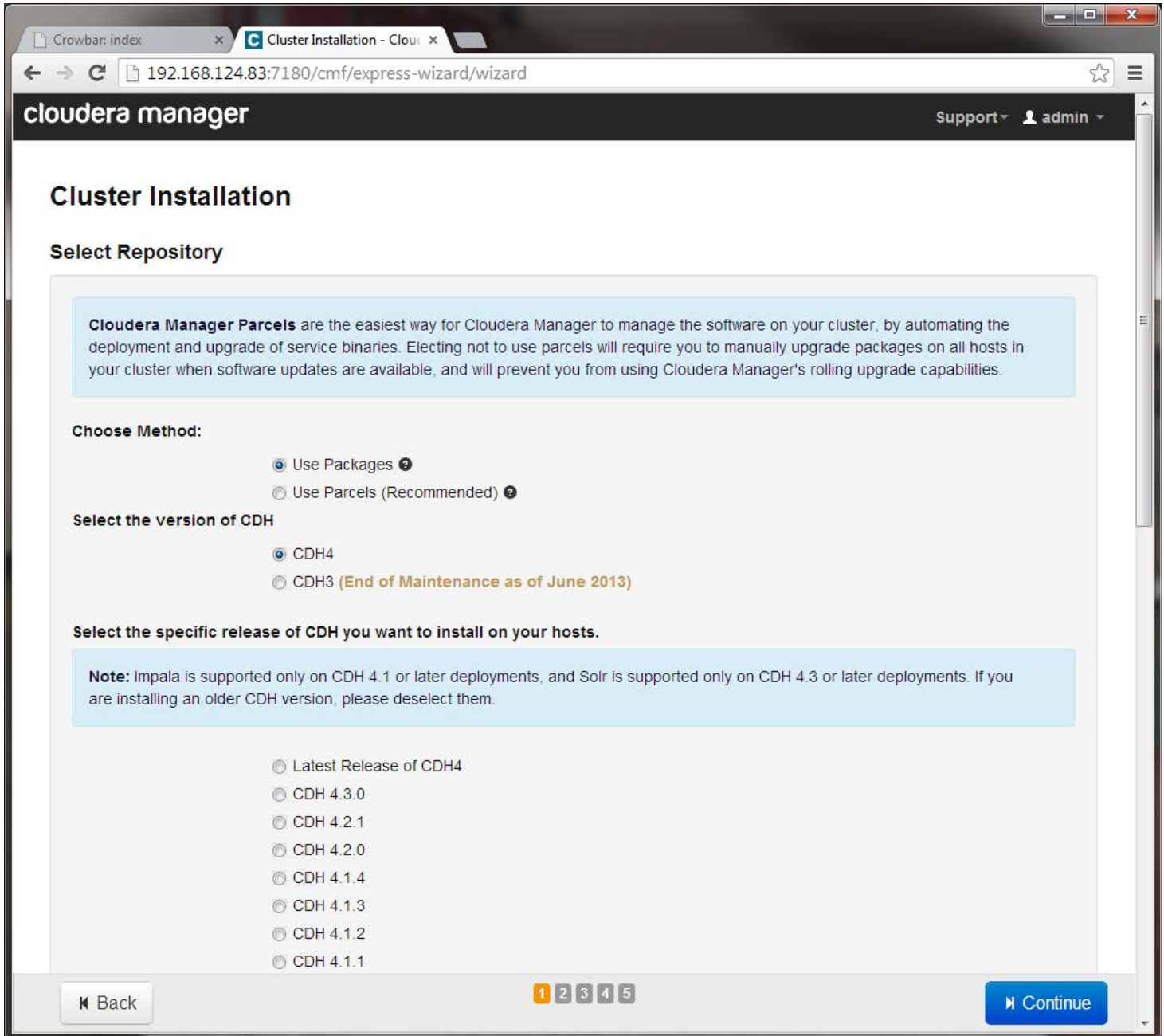
✎ The Dell | Cloudera Solution includes built-in software repositories, accessible via Packages instead of the default Cloudera "parcels". This enables you to install the software without Internet access.

Figure 8: Select Repository Screen

The Select Repository screen expands to display configuration choices.

Repository Configuration Screen

> 📝 RPM based packages are served from the crowbar admin node. By default, the IP address is 192.168.124.10 on port 8091 (http://192.168.124.10:8091). If you configure the crowbar admin node to be on another IP address, you will have to make the appropriate adjustments to the URLs listed above.

1. Select **CDH4** for installation.

2. Select **Custom Repository** for CDH, and then enter the distribution-specific URL.

   - **Red Hat** - http://192.168.124.10:8091/redhat-6.4/crowbar-extra/clouderamanager

   - **CentOS** - http://192.168.124.10:8091/centos-6.4/crowbar-extra/clouderamanager

3. If you wish to install Impala packages:

   a. Select **Custom Repository** for Impala, and then enter the distribution-specific URL.

      - **Red Hat** - http://192.168.124.10:8091/redhat-6.4/crowbar-extra/clouderamanager

      - **CentOS** - http://192.168.124.10:8091/centos-6.4/crowbar-extra/clouderamanager

   See Cloudera's Impala installation documentation for more information.

4. If you wish to install Solr packages:

   a. Select **Custom Repository** for Solr, and then enter the distribution-specific URL.

      - **Red Hat** - http://192.168.124.10:8091/redhat-6.4/crowbar-extra/clouderamanager

      - **CentOS** - http://192.168.124.10:8091/centos-6.4/crowbar-extra/clouderamanager

   See Cloudera's Solr installation documentation for more information.

5. Select **Custom Repository** for Cloudera Manager Agent, and then enter the distribution-specific URL.

   - **Red Hat** - http://192.168.124.10:8091/redhat-6.4/crowbar-extra/clouderamanager.

   - **CentOS** - http://192.168.124.10:8091/centos-6.4/crowbar-extra/clouderamanager

6. Leave the GPG Key URL field empty.

7. Click the **Continue** button to proceed.

About Cloudera Impala

Cloudera Impala enables you to perform fast SQL queries upon HDFS or HBase-stored Apache Hadoop data. It uses the same ODBC driver, SQL (Hive SQL) syntax, storage infrastructure, and user interface as Apache Hive. Impala is not a replacement for MapReduce-based batch processing frameworks.

> 📝 You must point the *Custom Repository for Impala* to Cloudera's corresponding repository in order to download Impala. See Repository Configuration Screen above. Cloudera Manager must be installed and operational upon a node with Internet access in order for Impala to function. Cloudera currently supports Impala running on Red Hat Enterprise Linux (RHEL)/CentOS 6.4 (64-bit) platforms only.

You can find Cloudera's Impala documentation at
http://www.cloudera.com/content/support/en/documentation/cloudera-impala/cloudera-impala-documentation-v1-latest.html.

About Cloudera Search

Cloudera Search, powered by Apache Solr™, enables fast, easy searches within a Hadoop cluster. Users are not required to have deep technical skills in order to use Cloudera Search effectively. Cloudera Search is a replacement for MapReduce-based batch processing frameworks.

> You must point the Custom Repository for Cloudera Search to Cloudera's corresponding repository in order to download Cloudera Search. See Repository Configuration Screen above. Cloudera Manager must be installed and operational upon a node with Internet access in order for Cloudera Search to function. Cloudera currently supports Cloudera Search running on Red Hat Enterprise Linux (RHEL)/CentOS 6.2 (64-bit) platforms only.

You can find Cloudera's Cloudera Search documentation at
http://www.cloudera.com/content/support/en/documentation/cloudera-search/cloudera-search-documentation-v1-latest.html.

Figure 9: Repository Configuration Screen
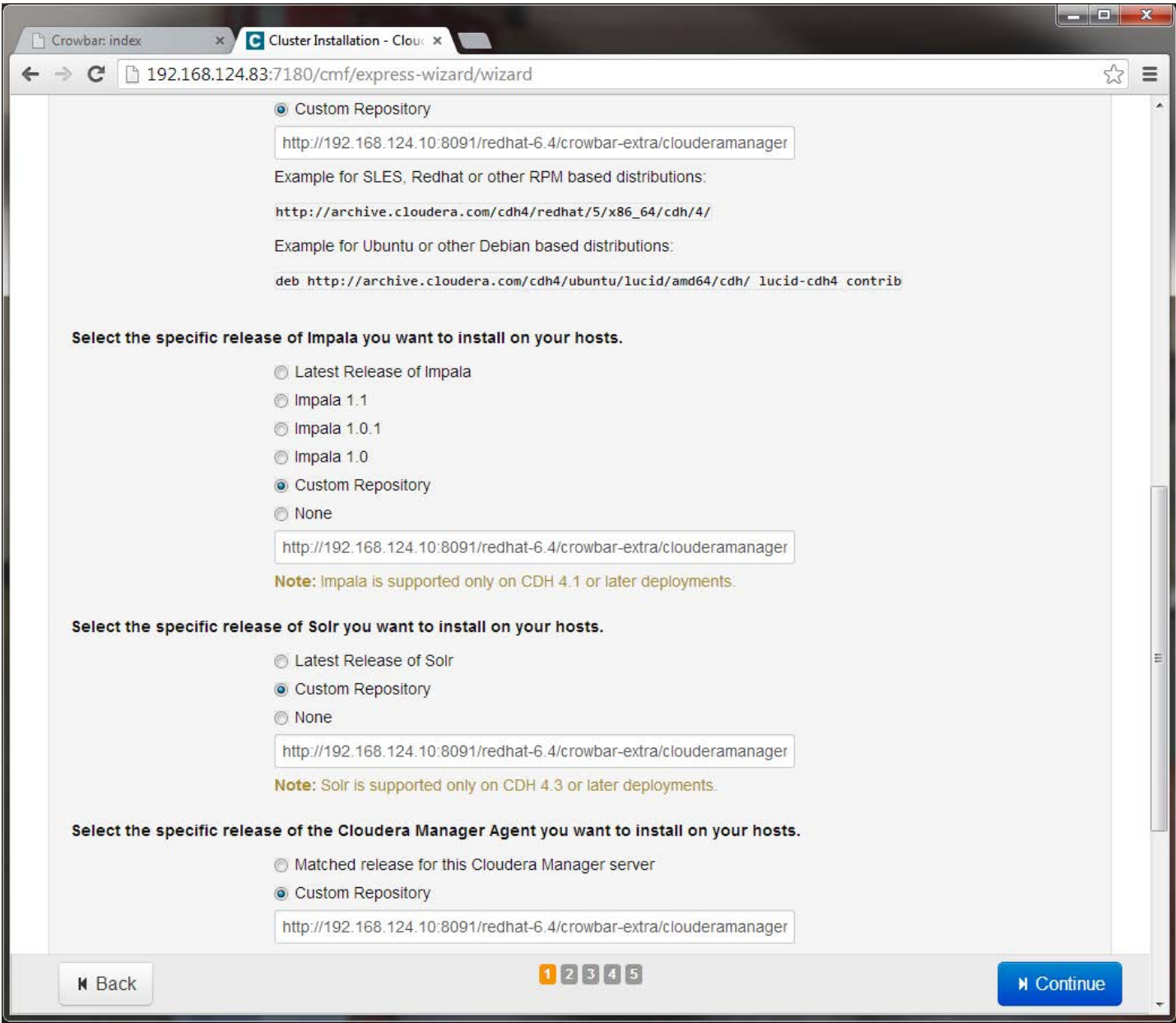
SSH Credentials Screen

1. Select **Login to all hosts as root**.

2. Select **All hosts accept same password**.

3. Enter the **SSH login password** for the cluster (default=crowbar).

4. Confirm the **SSH login password** for the cluster.

5. Accept the default settings for the **SSH port** and **number of simultaneous installations**.

6. Click the **Continue** button to proceed.

Figure 10: SSH Credentials Screen

Package Install Screen

You will see bar graphs next to each node and the name of the package it is installing.

1. Wait for the installation process to complete.

2. Click the **Continue** button to proceed.

Figure 11: Package Install Screen
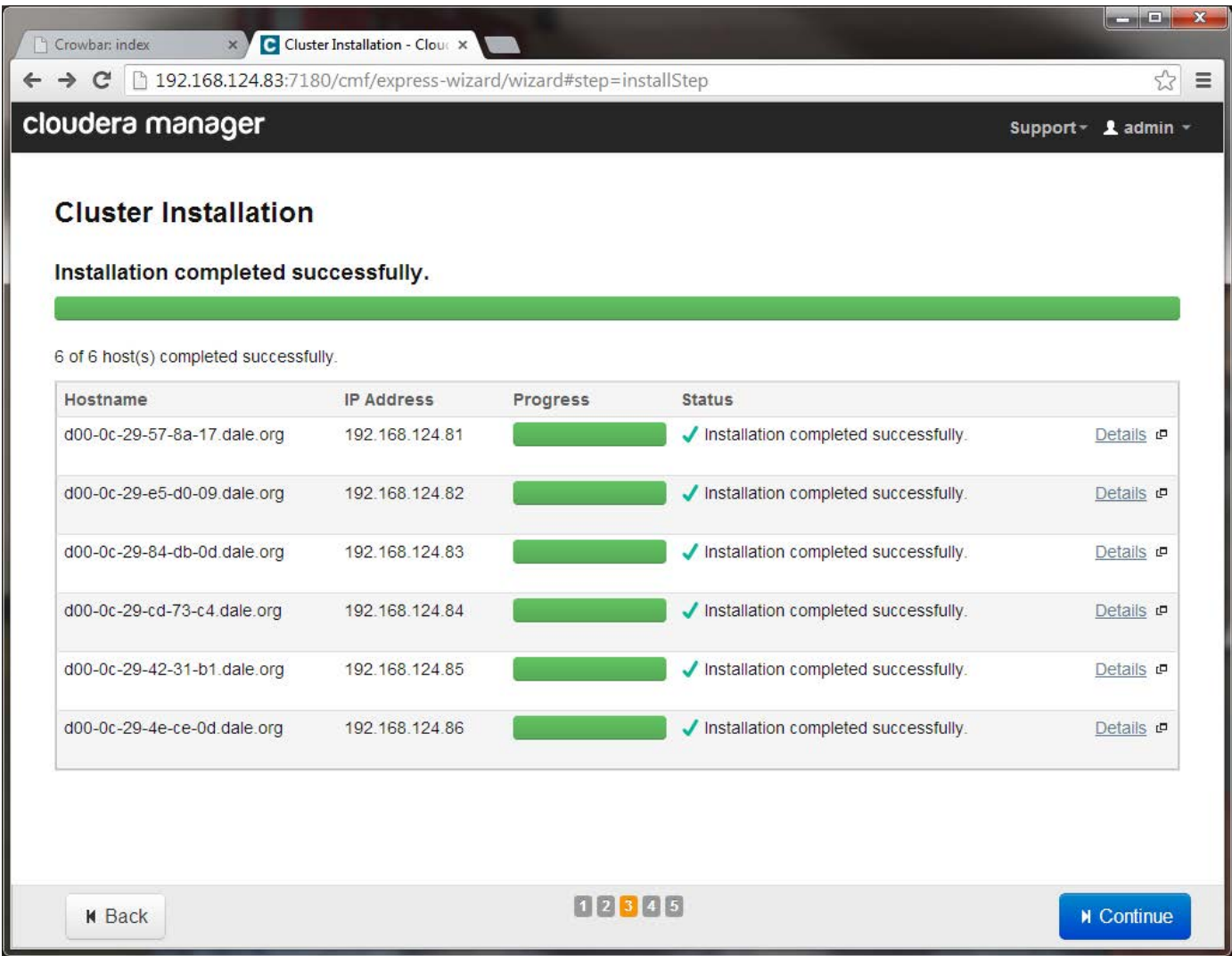
Host Inspector Screen

The Cloudera Manager Host Inspector runs during this part of the installation process in order to validate the proper cluster configuration for the Hadoop installation.

1. Wait for this process to complete.

2. Click the **Run Again** button if you want to run the Host Inspector again.

3. Click the **Continue** button to proceed.

Figure 12: Host Inspector Screen
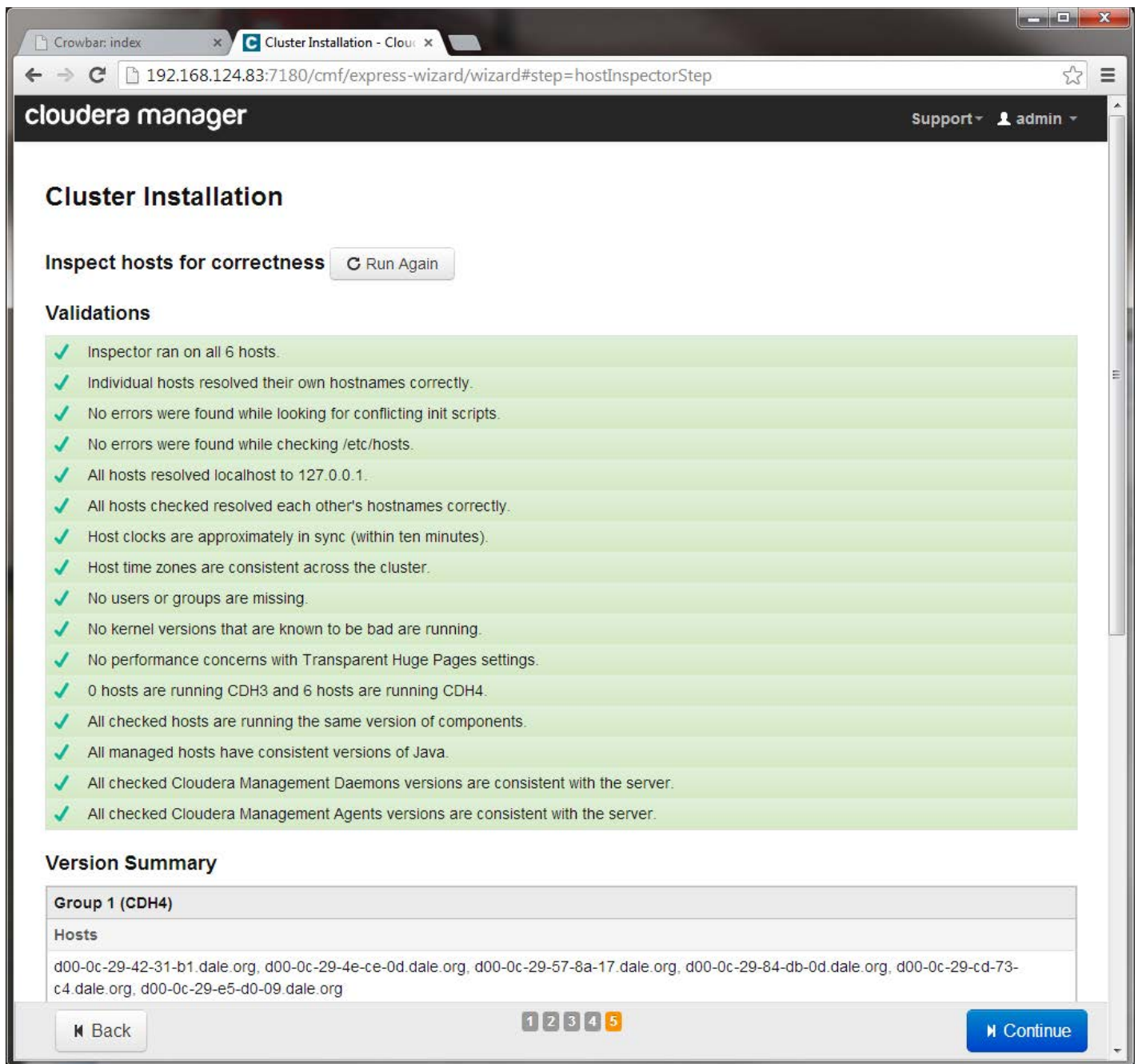
Service Selection Screen

1. Select the services that you want to install.

- **Core Hadoop** – Includes HDFS, MapReduce, Oozie, Hive, and Hue
- **Core with Real-Time Delivery** – Includes HDFS, MapReduce, ZooKeeper, HBase, Oozie, Hive, and Hue
- **Core with Real-Time Query** – Includes HDFS, MapReduce, Impala, Oozie, Hive, and Hue
- **All Services** – Includes HDFS, MapReduce, ZooKeeper, HBase, Impala, Oozie, Hive, and Hue
- **Custom Services** – Select only the services that you want
- **Cloudera Navigator** – A separately-licensed suite of management services

📝 If you select anything other than *All Services*, you can optionally add additional services in the future.

2. If you select **Cloudera Navigator**, first ensure that you have purchased the required licenses. Cloudera Navigator is a separately-licensed feature. Please contact your Dell representative for more information.

3. Click the **Inspect Role Assignments** button to configure the Hadoop cluster services.
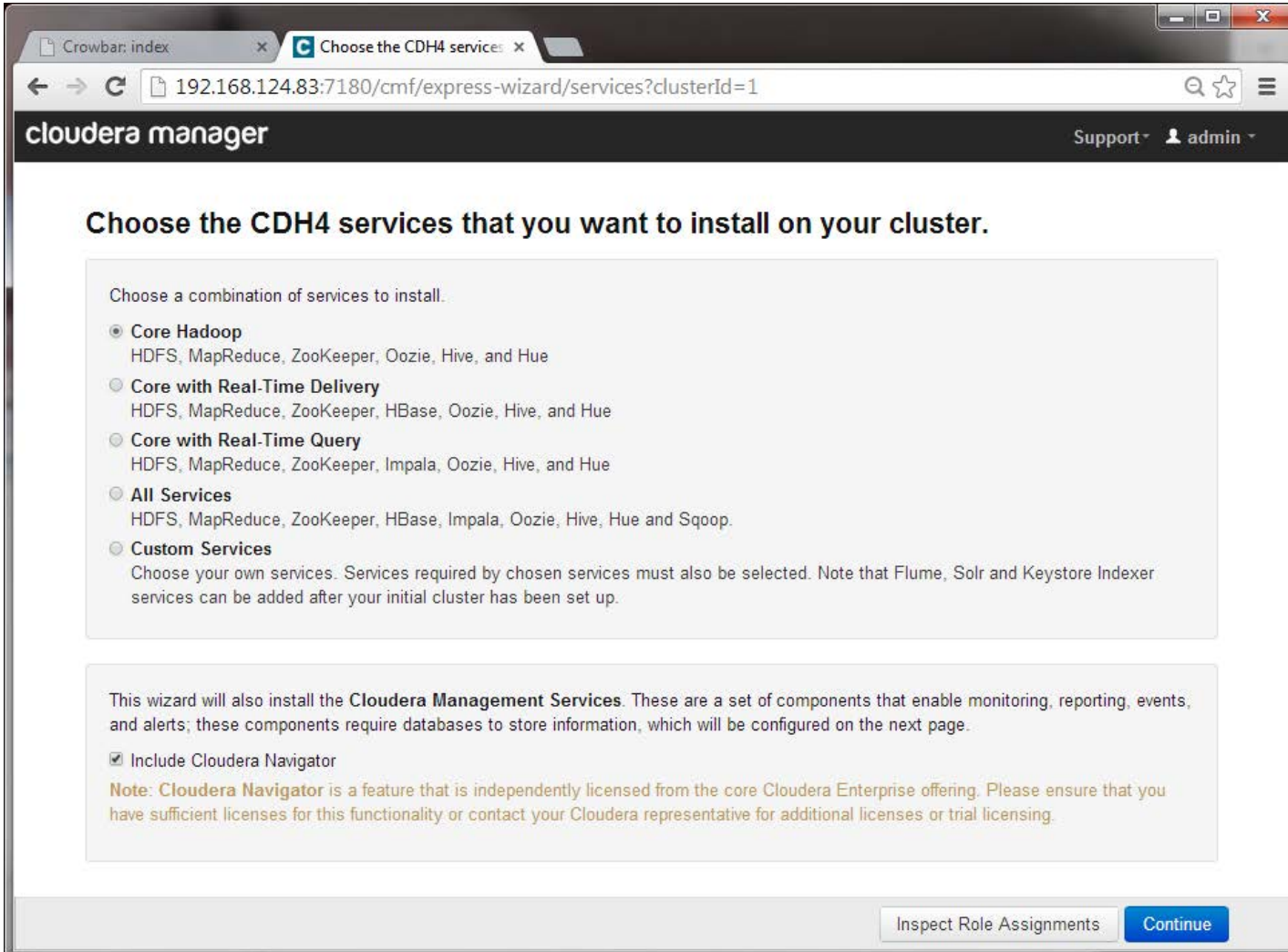
⚠ **Important**: **Do not** select *Continue*, as this will give you the default role assignments, which may not be acceptable to you.

Figure 13: Service Selection Screen

Inspect Role Assignments Screen # 1

1. Select the Cloudera Manager role assignments for Hadoop cluster deployment. Recommended settings for the Dell Reference Architecture:

   - **DataNode** – Crowbar nodes which contain the clouderamanager-datanode role.
   - **NameNode** – 1st Crowbar node which contains the clouderamanager-namenode role.
   - **SecondaryNameNode** – 2nd Crowbar node which contains the clouderamanager-namenode role.
   - **TaskTracker roles** – Crowbar nodes which contains the clouderamanager-datanode role.
   - **JobTracker role** – Crowbar node which contains the clouderamanager-namenode role.
   - **Cloudera Management Service roles** – Crowbar node which contains the clouderamanager-server role. Dell recommends that you assign these roles to the Edge Node.
   - **Zookeeper role** – Crowbar nodes which contains the clouderamanager-namenode role and either the clouderamanager-ha-journaling node role or the clouderamanager-ha-filernode role. At least three (3) nodes should be selected.

2. Please refer to Figure 15: Inspect Role Assignments Screen #2, before clicking the **Continue** button.
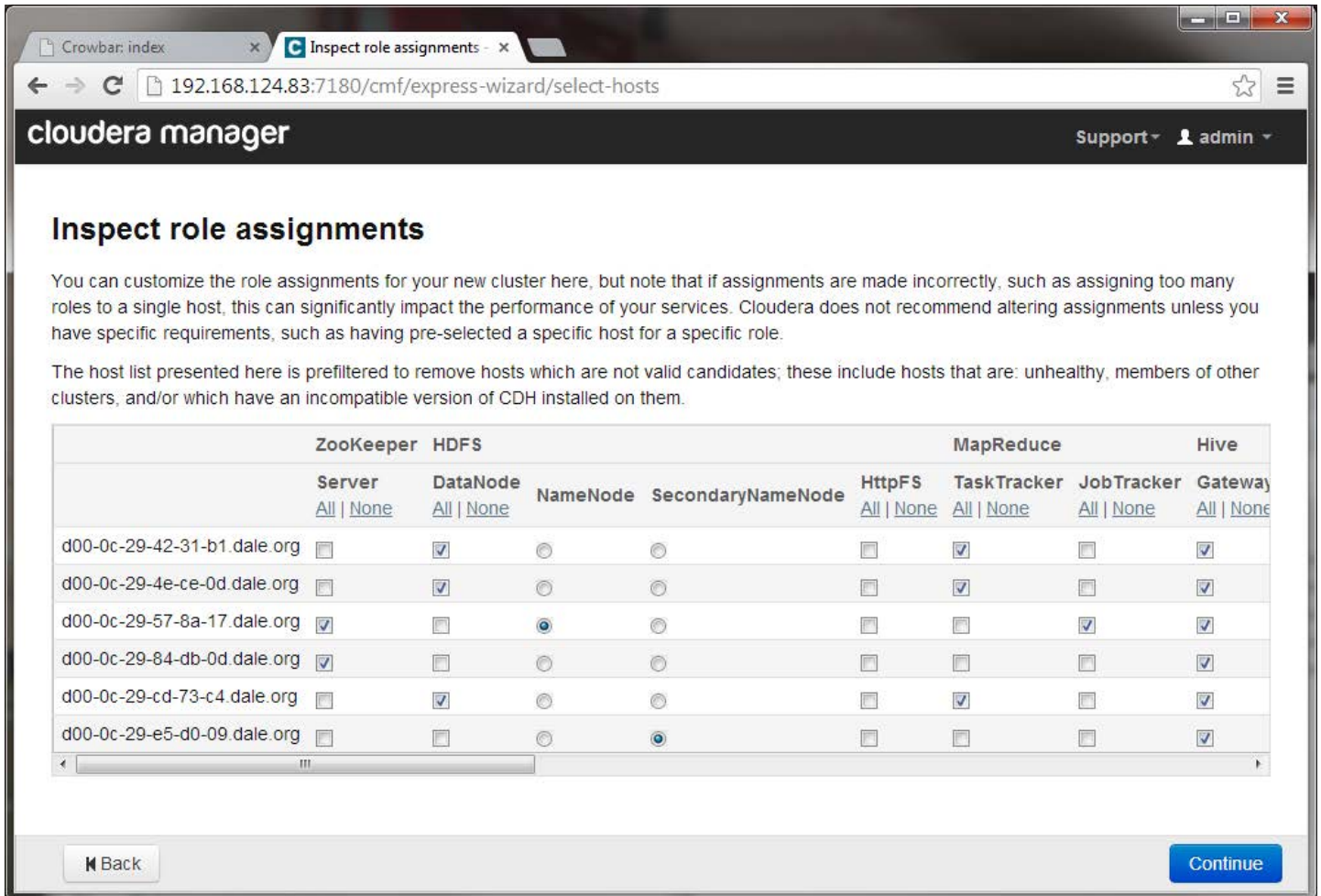
> ✎ The Cloudera Node Inventory page you printed from within the Cloudera Manager barclamp page in Crowbar is very useful for this step to ensure the roles selected in Cloudera Manager are assigned to nodes which have been provisioned (RAID, BIOS, etc.) specifically for that purpose.

Figure 14: Inspect Role Assignments Screen # 1

Inspect Role Assignments Screen # 2

If you entered the Cloudera Manager License key, you will see additional columns in this screen.

1. Select the role assignments for Hadoop add-ons services and monitoring services (Activity Monitor, Service Monitor, Reports Manager, Hbase components, Oozie, Hue, etc.). Dell suggests that you assign these roles to the Cloudera Manager Server role; usually the Edge node.

2. Click the **Continue** button to proceed.

Figure 15: Inspect Role Assignments Screen #2
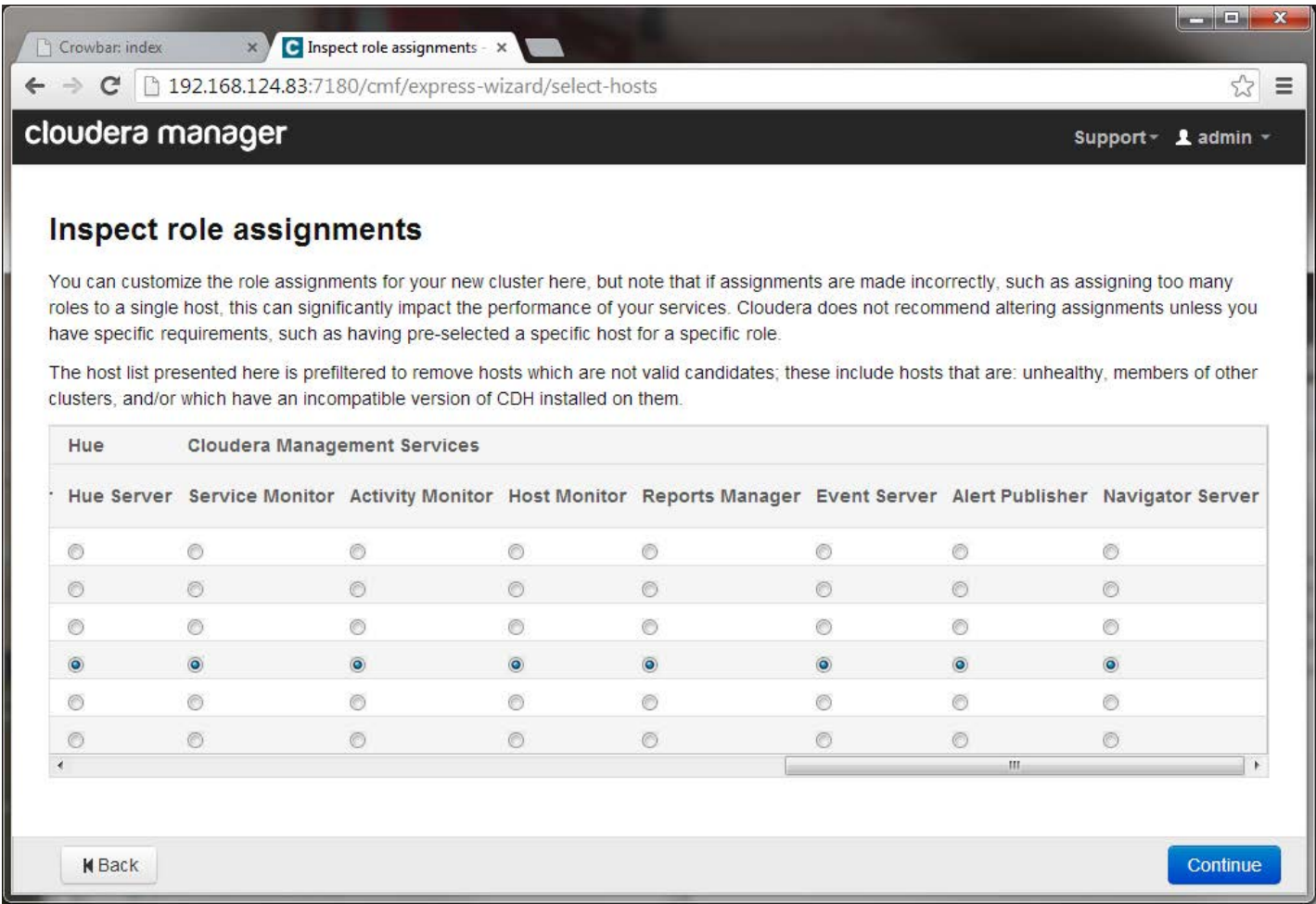
Monitoring Database Setup Screen

If you entered the Cloudera Manager License key, you will see this additional screen.

1. Select **Use Embedded Database**.

2. You can leave the rest of the settings at default values unless you want to change them.

3. Click the **Test Connection** button to make sure you can connect to all the databases (required).

4. Click the **Continue** button to proceed.

Figure 16: Monitoring Database Setup Screen

Review Configuration Changes Screen

If you entered the Cloudera Manager License key, you will see this additional screen.

1. If not set by default, set the Alert Publisher mail server hostname for alerts (*localhost*).

2. If not set by default, set the Alert Publisher mail server message recipients for alerts (*root@localhost*).

3. Click the **Continue** button to proceed.

Figure 17: Review Configuration Changes Screen

Cluster Services Initialization Screen

1. Wait for the Hadoop cluster installation process to complete.

2. Click the **Continue** button to proceed.

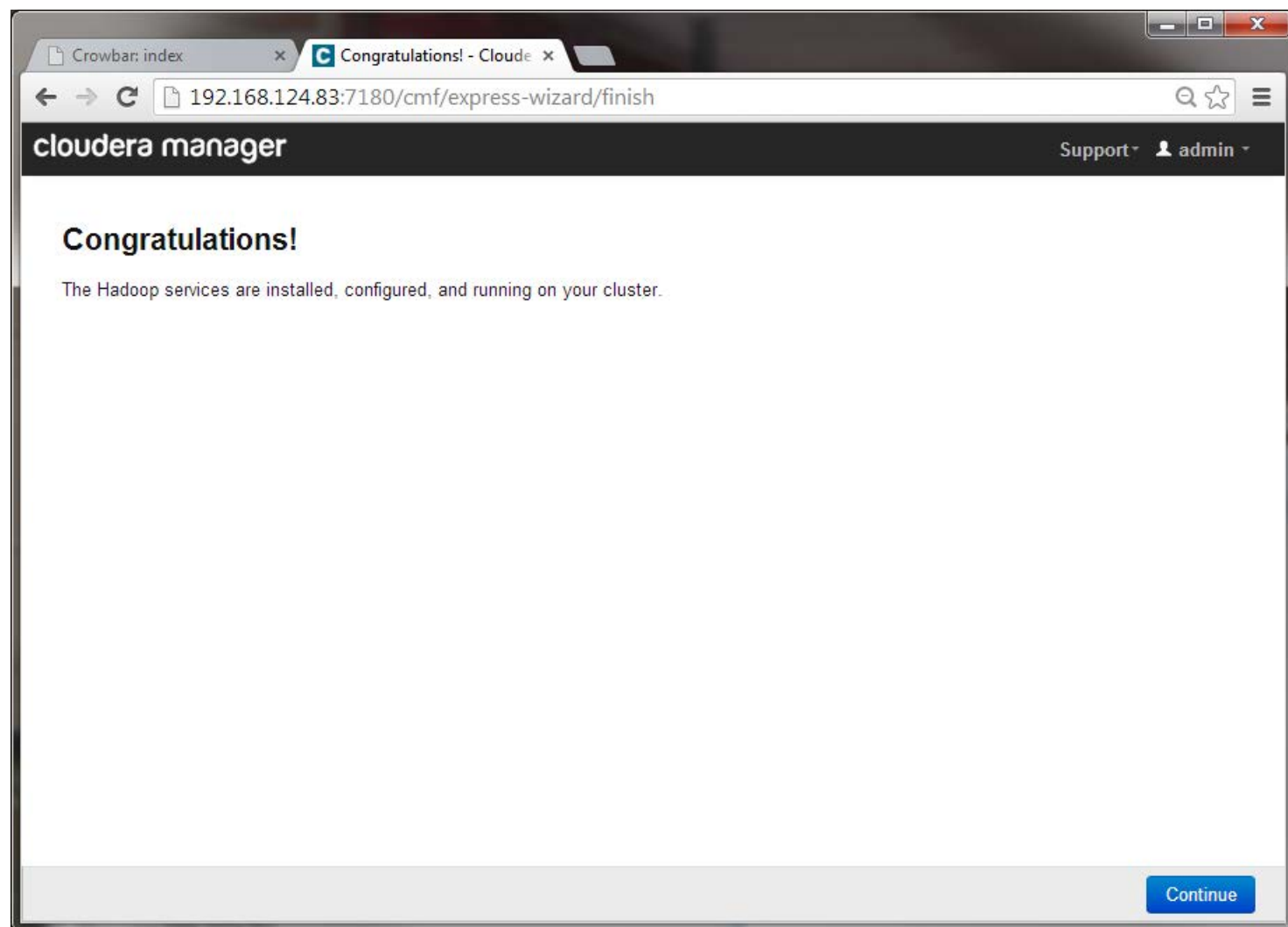**Figure 18: Cluster Services Initialization Screen**
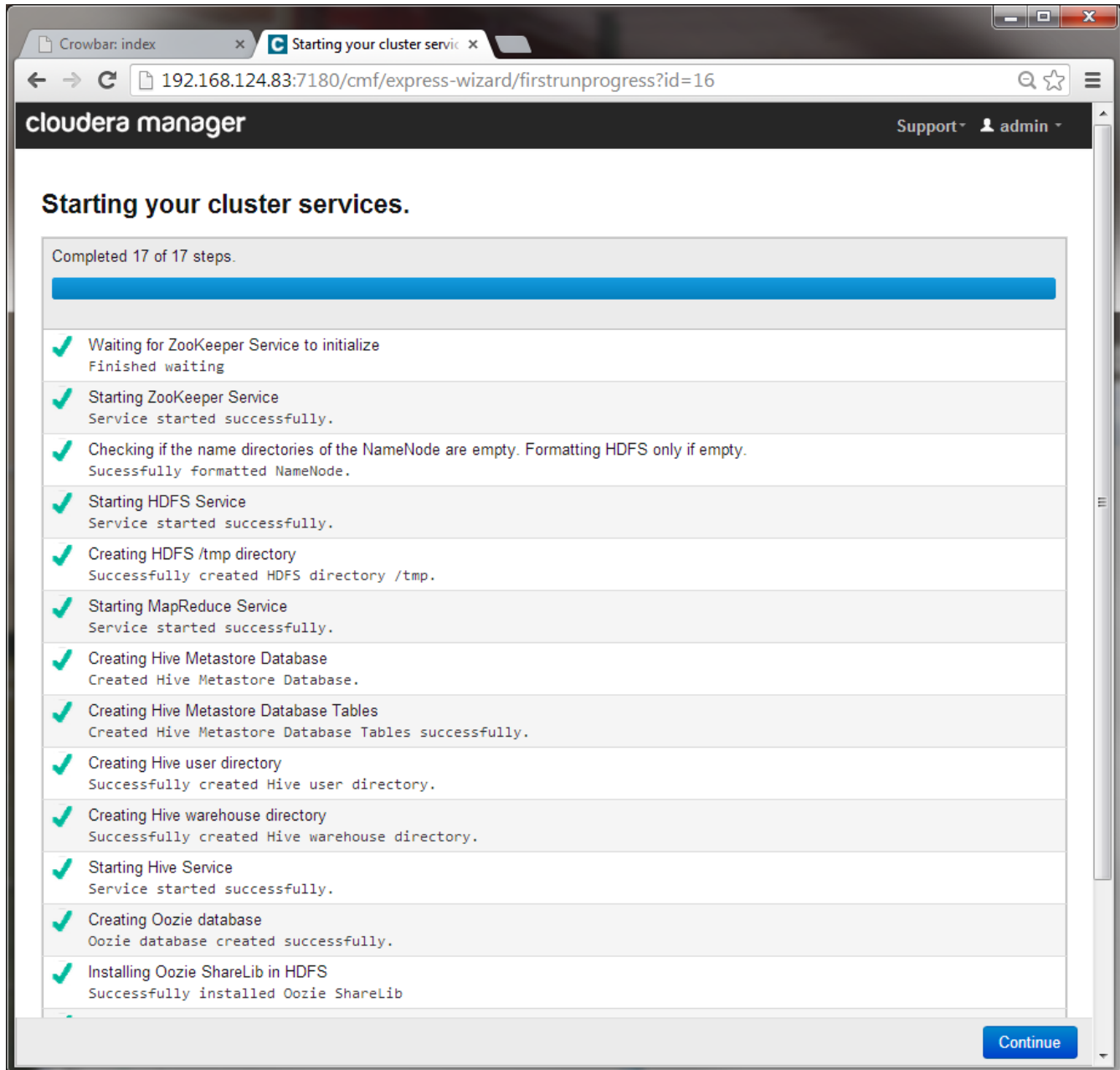
Configuration Completion Screen

If the Hadoop configuration steps complete successfully, you will see the final Cloudera Manager confirmation screen.

- Click the **Continue** button to start using Cloudera Manager.

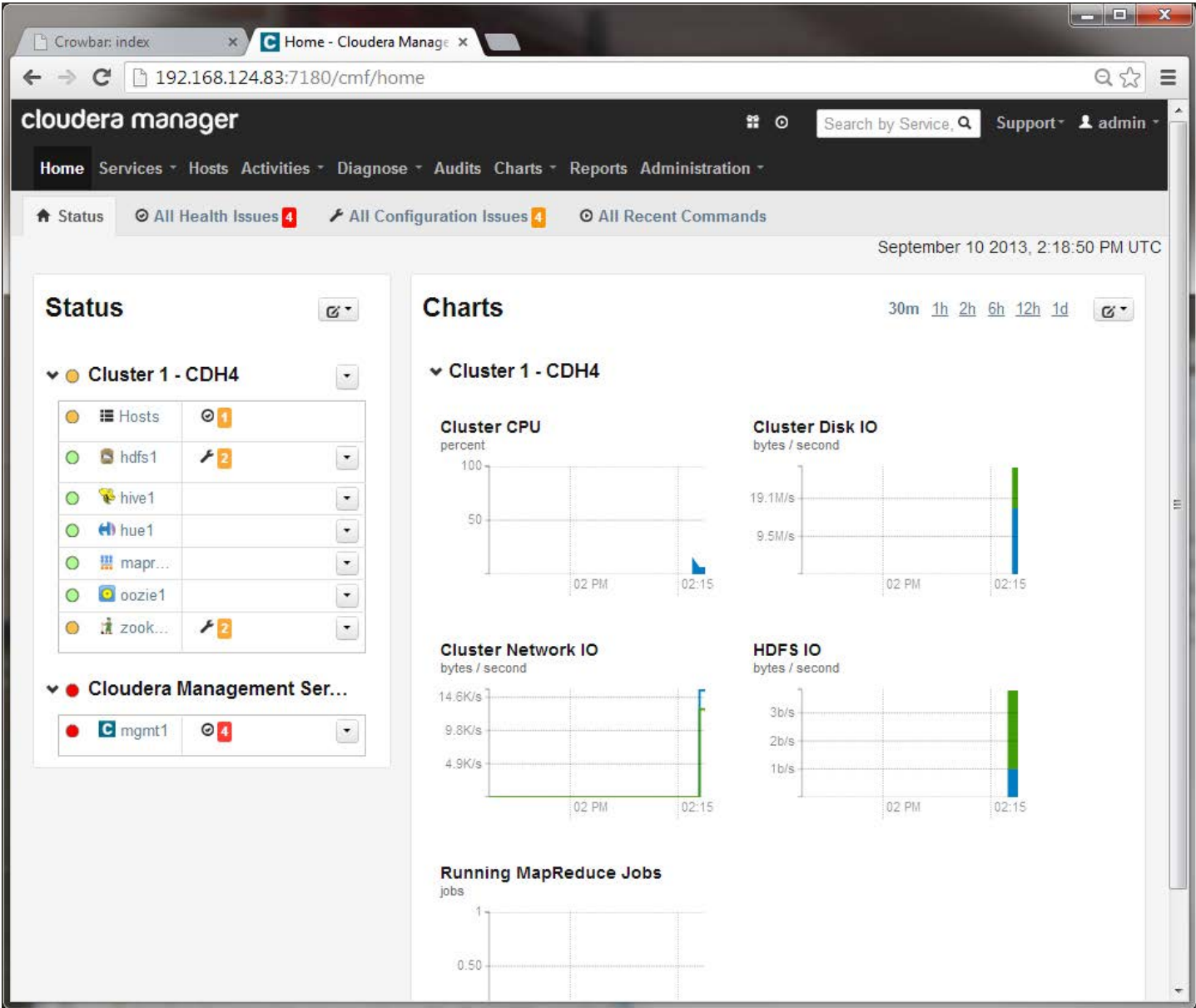**Figure 19: Configuration Completion Screen**

Service Display Screen

This is the normal startup screen after Cloudera Manager has completed the installation steps.

- Please refer to the *Cloudera Manager Users Guide* for additional information on operating Cloudera Manager.

Figure 20: Home Screen

Appendix A: Post-Installation Gateway Role Assignments

Local HDFS Filesystem Access

Data nodes, by default, have local access to the HDFS filesystem. If, after a manual Cloudera Manager installation, you require other nodes in addition to the data nodes to have local access to the HDFS filesystem, you must:

- Add the HDFS Gateway role to those nodes

- Deploy the client configuration from Cloudera Manager

> ✍ Data nodes should already be configured with the HDFS Gateway role.

To Add the HDFS Gateway Role and Deploy the Client Configuration:

1. Log into the Cloudera Manager Web interface.

2. Navigate to **Services** > **hdfs-(*cluster_name*)**. For example, *hdfs-cluster01*.

3. Click on the **Instances** button, and then click on the **Add** button.

4. Click in the **Gateway** column to select any nodes that require local access to the HDFS filesystem.

5. Click on the **Continue** button.

6. Click on the **Actions** drop-down, and then select **Deploy Client Configuration**.

7. Click on the **Deploy Client Configuration** button in the confirmation screen.

8. Wait for the operation to complete.

9. Click on the **Close** button.

All nodes with the Gateway role will now have local access to the HDFS filesystem.

Appendix B: Support

Dell Support

To obtain Dell hardware and software support:

- **Open a request at Dell's** support portal: http://support.dell.com

- See a list of Dell Technical Support call centers near you

Cloudera Support

To obtain support for Hadoop:

- **Open a request at Cloudera's support portal:** http://www.cloudera.com/hadoop-support/

Appendix C: References

- Cloudera: http://www.cloudera.com

- Nagios: http://www.nagios.org

- Ganglia: http://ganglia.sourceforge.net

To Learn More

For more information on the Dell | Cloudera Solution, visit:
www.Dell.com/Hadoop