VMware vSphere Big Data Extensions Administrator's and User's Guide

vSphere Big Data Extensions 2.2

This document supports the version of each product listed and supports all subsequent versions until the document is replaced by a new edition. To check for more recent editions of this document, see http://www.vmware.com/support/pubs.

EN-001701-00

vmware®

You can find the most up-to-date technical documentation on the VMware Web site at:

http://www.vmware.com/support/

The VMware Web site also provides the latest product updates.

If you have comments about this documentation, submit your feedback to:

docfeedback@vmware.com

VMware, Inc.
3401 Hillview Ave.
Palo Alto, CA 94304
www.vmware.com

Contents

About This Book

VMware vSphere Big Data Extensions Administrator's and User's Guide describes how to install VMware vSphere Big Data Extensions™ within your vSphere environment, and how to manage and monitor Hadoop and HBase clusters using the Big Data Extensions plug-in for vSphere Web Client.

VMware vSphere Big Data Extensions Administrator's and User's Guide also describes how to perform Hadoop and HBase operations using the VMware Serengeti™ Command-Line Interface Client, which provides a greater degree of control for certain system management and big data cluster creation tasks.

Intended Audience

This guide is for system administrators and developers who want to use Big Data Extensions to deploy and manage Hadoop clusters. To successfully work with Big Data Extensions, you should be familiar with VMware® vSphere® and Hadoop and HBase deployment and operation.

VMware Technical Publications Glossary

VMware Technical Publications provides a glossary of terms that might be unfamiliar to you. For definitions of terms as they are used in VMware technical documentation, go to http://www.vmware.com/support/pubs.

About VMware vSphere Big Data Extensions 1

VMware vSphere Big Data Extensions lets you deploy and centrally operate big data clusters running on VMware vSphere. Big Data Extensions simplifies the Hadoop and HBase deployment and provisioning process, and gives you a real time view of the running services and the status of their virtual hosts. It provides a central place from which to manage and monitor your big data cluster, and incorporates a full range of tools to help you optimize cluster performance and utilization.

This chapter includes the following topics:

Getting Started with Big Data Extensions

Big Data Extensions lets you deploy big data clusters. The tasks in this section describe how to set up VMware vSphere® for use with Big Data Extensions, deploy the Big Data Extensions vApp, access the VMware vCenter Server® and command-line interface (CLI) administrative consoles, and configure a Hadoop distribution for use with Big Data Extensions.

Prerequisites

- Understand what Project Serengeti® and Big Data Extensions is so that you know how they fit into your big data workflow and vSphere environment.
- Verify that the Big Data Extensions features that you want to use, such as data-compute separated clusters and elastic scaling, are supported by Big Data Extensions for the Hadoop distribution that you want to use.
- Understand which features are supported by your Hadoop distribution.

Procedure

1 Do one of the following.

   - Install Big Data Extensions for the first time. Review the system requirements, install vSphere, and install the Big Data Extensions components: Big Data Extensions vApp, Big Data Extensions plug-in for vCenter Server, and Serengeti CLI Client.
   - Upgrade Big Data Extensions from a previous version. Perform the upgrade steps.

2 (Optional) Install and configure a distribution other than Apache Bigtop for use with Big Data Extensions.

   Apache Bigtop is included in the Serengeti Management Server, but you can use any Hadoop distribution that Big Data Extensions supports.

What to do next

After you have successfully installed and configured your Big Data Extensions environment, you can perform the following additional tasks.

- Stop and start the Serengeti services, create user accounts, manage passwords, and log in to cluster nodes to perform troubleshooting.
- Manage the vSphere resource pools, datastores, and networks that you use to create Hadoop and HBase clusters.
- Create, provision, and manage big data clusters.
- Monitor the status of the clusters that you create, including their datastores, networks, and resource pools, through the vSphere Web Client and the Serengeti Command-Line Interface.
- On your Big Data clusters, run HDFS commands, Hive and Pig scripts , and MapReduce jobs, and access Hive data.
- If you encounter any problems when using Big Data Extensions, see Chapter 13, “Troubleshooting,” on page 139.

Big Data Extensions and Project Serengeti

Big Data Extensions runs on top of Project Serengeti, the open source project initiated by VMware to automate the deployment and management of Hadoop and HBase clusters on virtual environments such as vSphere.

Big Data Extensions and Project Serengeti provide the following components.

Project Serengeti

An open source project initiated by VMware, Project Serengeti lets users deploy and manage big data clusters in a vCenter Server managed environment. The major components are the Serengeti Management Server, which provides cluster provisioning, software configuration, and management services; an elastic scaling framework; and command-line interface. Project Serengeti is made available under the Apache 2.0 license, under which anyone can modify and redistribute Project Serengeti according to the terms of the license.

Serengeti Management Server

Provides the framework and services to run Big Data clusters on vSphere. The Serengeti Management Server performs resource management, policy-based virtual machine placement, cluster provisioning, software configuration management, and environment monitoring.

Serengeti Command-Line Interface Client

The command-line interface (CLI) client provides a comprehensive set of tools and utilities with which to monitor and manage your Big Data deployment. If you are using the open source version of Serengeti without Big Data Extensions, the CLI is the only interface through which you can perform administrative tasks. For more information about the CLI, see the *VMware vSphere Big Data Extensions Command-Line Interface Guide*.

Big Data Extensions

The commercial version of the open source Project Serengeti from VMware, Big Data Extensions, is delivered as a vCenter Server Appliance.
Big Data Extensions includes all the Project Serengeti functions and the following additional features and components.

- Enterprise level support from VMware.
- Bigtop distribution from the Apache community.

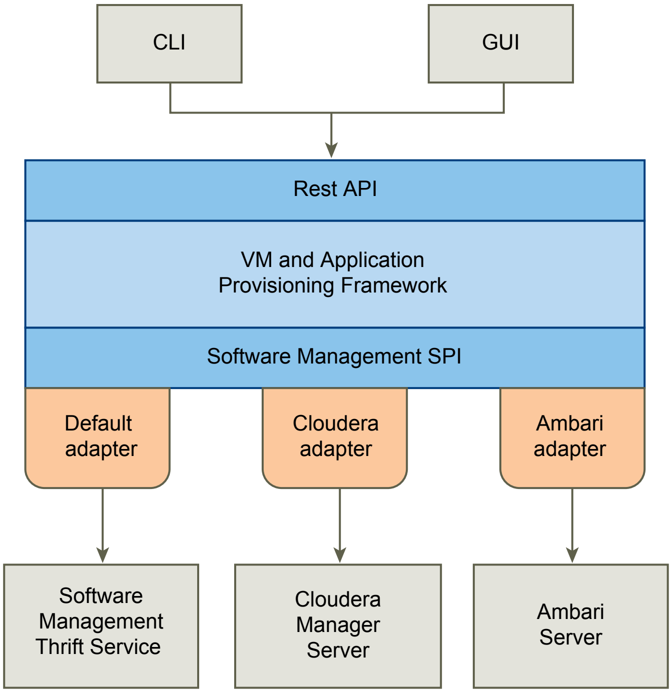
  NOTE VMware provides the Hadoop distribution as a convenience but does not provide enterprise-level support. The Apache Bigtop distribution is supported by the open source community.

- The Big Data Extensions plug-in, a graphical user interface integrated with vSphere Web Client. This plug-in lets you perform common Hadoop infrastructure and cluster management administrative tasks.
- Elastic scaling lets you optimize cluster performance and utilization of physical compute resources in a vSphere environment. Elasticity-enabled clusters start and stop virtual machines, adjusting the number of active compute nodes based on configuration settings that you specify, to optimize resource consumption. Elasticity is ideal in a mixed workload environment to ensure that workloads can efficiently share the underlying physical resources while high-priority jobs are assigned sufficient resources.

About Big Data Extensions Architecture

The Serengeti Management Server and Hadoop Template virtual machine work together to configure and provision big data clusters.

Figure 1-1. Big Data Extensions Architecture



Big Data Extensions performs the following steps to deploy a big data cluster.

1 The Serengeti Management Server searches for ESXi hosts with sufficient resources to operate the cluster based on the configuration settings that you specify, and then selects the ESXi hosts on which to place Hadoop virtual machines.

2 The Serengeti Management Server sends a request to the vCenter Server to clone and configure virtual machines to use with the big data cluster.

3 The Serengeti Management Server configures the operating system and network parameters for the new virtual machines.

4 Each virtual machine downloads the Hadoop software packages and installs them by applying the distribution and installation information from the Serengeti Management Server.

5 The Serengeti Management Server configures the Hadoop parameters for the new virtual machines based on the cluster configuration settings that you specify.

6 The Hadoop services are started on the new virtual machines, at which point you have a running cluster based on your configuration settings.

About Application Managers

You can use Cloudera Manager, Apache Ambari, and the default application manager to provision and manage clusters with VMware vSphere Big Data Extensions.

After you add a new Cloudera Manager or Ambari application manager to Big Data Extensions, you can redirect your software management tasks, including monitoring and managing clusters, to that application manager.

You can use an application manager to perform the following tasks:

- List all available vendor instances, supported distributions, and configurations or roles for a specific application manager and distribution.
- Create clusters.
- Monitor and manage services from the application manager console.

Check the documentation for your application manager for tool-specific requirements.

Restrictions

The following restrictions apply to Cloudera Manager and Ambari application managers:

- To add an application manager with HTTPS, use the FQDN instead of the URL.
- You cannot rename a cluster that was created with a Cloudera Manager or Ambari application manager.
- You cannot change services for a big data cluster from Big Data Extensions if the cluster was created with Ambari or Cloudera Manager application manager.
- To change services, configurations, or both, you must make the changes from the application manager on the nodes.

  If you install new services, Big Data Extensions starts and stops the new services together with old services.
- If you use an application manager to change services and big data cluster configurations, those changes cannot be synced from Big Data Extensions. The nodes that you create with Big Data Extensions do not contain the new services or configurations.

Services and Operations Supported by the Application Managers

If you use Cloudera Manager or Apache Ambari with Big Data Extensions, there are several additional services that are available for your use.

Supported Application Managers and Distributions

Big Data Extensions supports certain application managers and Hadoop distributions.

Table 1-1. Supported application managers and Hadoop distributions

Application Managers	Supported Version	Supported Distributions	Supported EMC Isilon OneFS
Cloudera Manager	5.3, 5.4	CDH 5.3, 5.4	Isilon OneFS 7.1, 7.2
Apache Ambari	1.6, 1.7	HDP 2.1, 2.2, PHD 3.0	Isilon OneFS 7.1, 7.2 Big Data Extensions does not support the provisioning of compute-only clusters with Ambari Manager. However Ambari can provision compute-only clusters when using Isilon OneFS. Refer to the EMC Isilon Hadoop Starter Kit for Hortonworks documentation for information on configuring Ambari and Isilon OneFS.
Default		Apache Bigtop 0.8, CDH 5.2, 5.3, HDP 2.1, PHD 2.0, 2.1 MapR 4.0, 4.1	Isilon OneFS 7.1, 7.2

Supported Features and Operations

The following features and operations are available when you use the Ambari application manager 1.6 and 1.7 (with versions HDP 2.1, 2.2, or PHD 3.0) and the Cloudera Manager application manager 5.3 and 5.4 (with versions CDH 5.3 and 5.4) on Big Data Extensions.

- Create Hadoop Cluster
- Create HBase Cluster
- Scale Cluster In/Out
- Cluster Delete
- Cluster Export (can only be performed with the Serengeti CLI)
- Cluster List
- Cluster Resume
- Cluster Start/Stop
- NameNode High Availability (available only with Ambari and Cloudera Manager)

  To use NameNode High Availability (HA) with Ambari, you must configure NameNode HA for use with your Hadoop deployment. See NameNode High Availability for Hadoop in the Hortonworks documentation.
- Hadoop Topology Awareness (RACK_AS_RACK, HOST_AS_RACK or HVE)
- vSphere Fault Tolerance
- vSphere High Availability

Services supported on Cloudera Manager and Ambari

Table 1-2. Services supported on Cloudera Manager and Ambari

Service Name	Cloudera Manager 5.3, 5.4	Ambari 1.6, 1.7
Falcon		X
Flume	X	X
Ganglia		X
HBase	X	X
HCatalog		X
HDFS	X	X
Hive	X	X
Hue	X	X
Impala	X	
MapReduce	X	X
Nagios		X
Oozie	X	X
Pig		X
Sentry		
Solr	X	
Spark	X	
Sqoop	X	X

Table 1-2. Services supported on Cloudera Manager and Ambari (Continued)

Service Name	Cloudera Manager 5.3, 5.4	Ambari 1.6, 1.7
Storm		X
TEZ		X
WebHCAT		X
YARN	X	X
Zookeeper	X	X

About Service Level vSphere High Availability for Ambari

Ambari supports NameNode HA, however, you must configure NameNode HA for use with your Hadoop deployment. See NameNode High Availability for Hadoop in the Hortonworks documentation.

About Service Level vSphere High Availability for Cloudera

The Cloudera distributions offer the following support for Service Level vSphere HA.

- Cloudera using MapReduce v1 provides service level vSphere HA support for JobTracker.
- Cloudera provides its own service level HA support for NameNode through HDFS2.

For information about how to use an application manager with the CLI, see the *VMware vSphere Big Data Extensions Command-Line Interface Guide*.

Big Data Extensions Support for MapReduce Distribution Features

Big Data Extensions provides different levels of feature support depending on the distribution and version that you configure for use with the default application manager.

Support for Hadoop MapReduce v1 Distribution Features

Table 1-3 lists the supported Hadoop MapReduce v1 distributions and indicates which features are supported when you use the distribution with the default application manager on Big Data Extensions.

Table 1-3. Big Data Extensions Feature Support for Hadoop MapReduce v1 Distributions

	Cloudera	Hortonworks	MapR
Version	5.3, 5.4	1.3	4.0, 4.1
Automatic Deployment	Yes	Yes	Yes
Scale Out	Yes	Yes	Yes
Create Cluster with Multiple Networks	Yes	Yes	No
Data-Compute Separation	Yes	Yes	Yes
Compute-only	Yes	Yes	No
Elastic Scaling of Compute Nodes	Yes when using MapReduce v1	Yes	No
Hadoop Configuration	Yes	Yes	No
Hadoop Topology Configuration	Yes	Yes	No
Hadoop Virtualization Extensions (HVE)	No	Yes	No

Table 1-3. Big Data Extensions Feature Support for Hadoop MapReduce v1 Distributions (Continued)

	Cloudera	Hortonworks	MapR
vSphere HA	Yes	Yes	Yes
Service Level vSphere HA	See "About Service Level vSphere High Availability for Cloudera," on page 15	Yes	No
vSphere FT	Yes	Yes	Yes

Support for Hadoop MapReduce v2 (YARN) Distribution Features

Table 1-4 lists the supported Hadoop MapReduce v2 distributions and indicates which features are supported when you use the distribution with the default application manager on Big Data Extensions.

Table 1-4. Big Data Extensions Feature Support for Hadoop MapReduce v2 (YARN) Distributions

	Apache Bigtop	Cloudera	Hortonworks	MapR	Pivotal
Version	0.8	5.3, 5.4	2.1	4.0, 4.1	2.0, 2.1
Automatic Deployment	Yes	Yes	Yes	Yes	Yes
Scale Out	Yes	Yes	Yes	Yes	Yes
Create Cluster with Multiple Networks	Yes	Yes	Yes	No	Yes
Data-Compute Separation	Yes	Yes	Yes	Yes	Yes
Compute-only	Yes	Yes	Yes	No	Yes
Elastic Scaling of Compute Nodes	No	No when using MapReduce 2	Yes	No	No
Hadoop Configuration	Yes	Yes	Yes	No	Yes
Hadoop Topology Configuration	Yes	Yes	Yes	No	Yes
Hadoop Virtualization Extensions (HVE)	Support only for HDFS	Support only for HDFS	Support only for HDFS.	No	Yes
vSphere HA	No	No	No	Yes	No
Service Level vSphere HA	No	See "About Service Level vSphere High Availability for Cloudera," on page 15	No	No	No
vSphere FT	No	No	No	Yes	No

Feature Support By Hadoop Distribution

Each Hadoop distribution and version provides differing feature support. Learn which Hadoop distributions support which features.

Hadoop Features

The table illustrates which Hadoop distributions support which features when you use the distributions with the default application manager on Big Data Extensions.

Table 1-5. Hadoop Feature Support

	Apache Bigtop	Cloudera	Hortonworks	MapR	Pivotal
Version	0.8	5.3, 5.4	2.1, 2.2	4.0, 4.1	2.0 , 2.1
HDFS1	No	Yes	No	No	No
HDFS2	Yes	Yes	Yes	No	Yes
MapReduce v1	No	Yes	No	Yes	No
MapReduce v2 (YARN)	Yes	Yes	Yes	Yes	Yes
Pig	Yes	Yes	Yes	Yes	Yes
Hive	Yes	Yes	Yes	Yes	Yes
Hive Server	Yes	Yes	Yes	Yes	Yes
HBase	Yes	Yes	Yes	Yes	Yes
ZooKeeper	Yes	Yes	Yes	Yes	Yes

Installing Big Data Extensions 2

To install Big Data Extensions so that you can create and provision big data clusters, you must install the Big Data Extensions components in the order described.

What to do next

If you want to create clusters on any Hadoop distribution other than Apache Bigtop, which is included in theSerengeti Management Server, install and configure the distribution for use with Big Data Extensions.

This chapter includes the following topics:

System Requirements for Big Data Extensions

Before you begin the Big Data Extensions deployment tasks, your system must meet all of the prerequisites for vSphere, clusters, networks, storage, hardware, and licensing.

Big Data Extensions requires that you install and configure vSphere and that your environment meets minimum resource requirements. Make sure that you have licenses for the VMware components of your deployment.

vSphere Requirements

Before you install Big Data Extensions, set up the following VMware products.

- Install vSphere 5.5 (or later) Enterprise or Enterprise Plus.

- When you install Big Data Extensions on vSphere 5.5 or later, use VMware® vCenter™ Single Sign-On to provide user authentication. When logging in to vSphere 5.5 or later you pass authentication to the vCenter Single Sign-On server, which you can configure with multiple identity sources such as Active Directory and OpenLDAP. On successful authentication, your user name and password is exchanged for a security token that is used to access vSphere components such as Big Data Extensions.
- Configure all ESXi hosts to use the same Network Time Protocol (NTP) server.
- On each ESXi host, add the NTP server to the host configuration, and from the host configuration's Startup Policy list, select **Start and stop with host**. The NTP daemon ensures that time-dependent processes occur in sync across hosts.

Cluster Settings

Configure your cluster with the following settings.

- Enable vSphere HA and VMware vSphere® Distributed Resource Scheduler™.
- Enable Host Monitoring.
- Enable admission control and set the policy you want. The default policy is to tolerate one host failure.
- Set the virtual machine restart priority to high.
- Set the virtual machine monitoring to virtual machine and application monitoring.
- Set the monitoring sensitivity to high.
- Enable vMotion and Fault Tolerance logging.
- All hosts in the cluster have Hardware VT enabled in the BIOS.
- The Management Network VMkernel Port has vMotion and Fault Tolerance logging enabled.

Network Settings

Big Data Extensions can deploy clusters on a single network or use multiple networks. The environment determines how port groups that are attached to NICs are configured and which network backs each port group.

You can use either a vSwitch or vSphere Distributed Switch (vDS) to provide the port group backing a Serengeti cluster. vDS acts as a single virtual switch across all attached hosts while a vSwitch is per-host and requires the port group to be configured manually.

When you configure your networks to use with Big Data Extensions, verify that the following ports are open as listening ports.

- Ports 8080 and 8443 are used by the Big Data Extensions plug-in user interface and the Serengeti Command-Line Interface Client.
- Port 5480 is used by vCenter Single Sign-On for monitoring and management.
- Port 22 is used by SSH clients.
- To prevent having to open a network firewall port to access Hadoop services, log into the Hadoop client node, and from that node you can access your cluster.

- To connect to the internet (for example, to create an internal yum repository from which to install Hadoop distributions), you may use a proxy.
- To enable communications, be sure that firewalls and web filters do not block the Serengeti Management Server or other Serengeti nodes.

Direct Attached Storage

Attach and configure direct attached storage on the physical controller to present each disk separately to the operating system. This configuration is commonly described as Just A Bunch Of Disks (JBOD). Create VMFS datastores on direct attached storage using the following disk drive recommendations.

- 8-12 disk drives per host. The more disk drives per host, the better the performance.
- 1-1.5 disk drives per processor core.
- 7,200 RPM disk Serial ATA disk drives.

Do not use Big Data Extensions in conjunction with vSphere Storage DRS

Big Data Extensions places virtual machines on hosts according to available resources, Hadoop best practices, and user defined placement policies prior to creating virtual machines. For this reason, you should not deploy Big Data Extensions on vSphere environments in combination with Storage DRS. Storage DRS continuously balances storage space usage and storage I/O load to meet application service levels in specific environments. If Storage DRS is used with Big Data Extensions, it will disrupt the placement policies of your Big Data cluster virtual machines.

Migrating virtual machines in vCenter Server may disrupt the virtual machine placement policy

Big Data Extensions places virtual machines based on available resources, Hadoop best practices, and user defined placement policies that you specify. For this reason, DRS is disabled on all the virtual machines created within the Big Data Extensions environment. While this prevents virtual machines from being automatically migrated by vSphere, it does not prevent you from inadvertently moving virtual machines using the vCenter Server user interface. This may break the Big Data Extensions defined placement policy. For example, this may disrupt the number of instances per host and group associations.

Resource Requirements for the vSphere Management Server and Templates

- Resource pool with at least 27.5GB RAM.
- 40GB or more (recommended) disk space for the management server and Hadoop template virtual disks.

Resource Requirements for the Hadoop Cluster

- Datastore free space is not less than the total size needed by the Hadoop cluster, plus swap disks for each Hadoop node that is equal to the memory size requested.
- Network configured across all relevant ESXi hosts, and has connectivity with the network in use by the management server.
- vSphere HA is enabled for the master node if vSphere HA protection is needed. To use vSphere HA or vSphere FT to protect the Hadoop master node, you must use shared storage.

Hardware Requirements for the vSphere and Big Data Extensions Environment

Host hardware is listed in the *VMware Compatibility Guide*. To run at optimal performance, install your vSphere and Big Data Extensions environment on the following hardware.

- Dual Quad-core CPUs or greater that have Hyper-Threading enabled. If you can estimate your computing workload, consider using a more powerful CPU.
- Use High Availability (HA) and dual power supplies for the master node's host machine.
- 4-8 GBs of memory for each processor core, with 6% overhead for virtualization.
- Use a 1GB Ethernet interface or greater to provide adequate network bandwidth.

Tested Host and Virtual Machine Support

The maximum host and virtual machine support that has been confirmed to successfully run with Big Data Extensions is 256 physical hosts running a total of 512 virtual machines.

vSphere Licensing

You must use a vSphere Enterprise license or above to use VMware vSphere HA and vSphere DRS.

Unicode UTF-8 and Special Character Support

Big Data Extensions supports internationalization (I18N) level 3. However, there are resources you specify that do not provide UTF-8 support. You can use only ASCII attribute names consisting of alphanumeric characters and underscores (_) for these resources.

Big Data Extensions Supports Unicode UTF-8

vCenter Server resources you specify using both the CLI and vSphere Web Client can be expressed with underscore (_), hyphen (-), blank spaces, and all letters and numbers from any language. For example, you can specify resources such as datastores labeled using non-English characters.

When using a Linux operating system, you should configure the system for use with UTF-8 encoding specific to your locale. For example, to use U.S. English, specify the following locale encoding: en_US.UTF-8. See your vendor's documentation for information on configuring UTF-8 encoding for your Linux environment.

Special Character Support

The following vCenter Server resources can have a period (.) in their name, letting you select them using both the CLI and vSphere Web Client.

- portgroup name
- cluster name
- resource pool name
- datastore name

The use of a period is not allowed in the Serengeti resource name.

Resources Excluded From Unicode UTF-8 Support

The Serengeti cluster specification file, manifest file, and topology racks-hosts mapping file do not provide UTF-8 support. When you create these files to define the nodes and resources for use by the cluster, use only ASCII attribute names consisting of alphanumeric characters and underscores (_).

The following resource names are excluded from UTF-8 support:

- cluster name
- nodeGroup name
- node name
- virtual machine name

The following attributes in the Serengeti cluster specification file are excluded from UTF-8 support:

- distro name
- role
- cluster configuration
- storage type
- haFlag
- instanceType
- groupAssociationsType

The rack name in the topology racks-hosts mapping file, and the placementPolicies field of the Serengeti cluster specification file is also excluded from UTF-8 support.

The Customer Experience Improvement Program

You can configure Big Data Extensions to collect data to help improve your user experience with VMware products. The following section contains important information about the VMware Customer Experience Improvement Program.

The goal of the Customer Experience Improvement Program is to quickly identify and address problems that might be affecting your experience. If you choose to participate in the Customer Experience Improvement Program,Big Data Extensions will regularly send anonymous data to VMware. You can use this data for product development and troubleshooting purposes.

Before collecting the data, VMware makes anonymous all fields that contain information that is specific to your organization. VMware sanitizes fields by generating a hash of the actual value. When a hash value is collected, VMware cannot identify the actual value but can detect changes in the value when you change your environment.

Categories of Information in Collected Data

When you choose to participate in VMware’s Customer Experience Improvement Program (CEIP), VMware will receive the following categories of data:

Configuration Data

Data about how you have configured VMware products and information related to your IT environment. Examples of Configuration Data include: version information for VMware products; details of the hardware and software running in your environment; product configuration settings, and information about your networking environment. Configuration Data may include hashed versions of your device IDs and MAC and Internet Protocol Addresses.

Feature Usage Data

Data about how you use VMware products and services. Examples of Feature Usage Data include: details about which product features are used; metrics of user interface activity; and details about your API calls.

Performance Data

Data about the performance of VMware products and services. Examples of Performance Data include metrics of the performance and scale of VMware products and services; response times for User Interfaces, and details about your API calls.

Enabling and Disabling Data Collection

By default, enrollment in the Customer Experience Improvement Program is enabled during installation. You have the option of disabling this service during installation. You can discontinue participation in the Customer Experience Improvement Program at any time, and stop sending data to VMware. See“Disable the Big Data Extensions Data Collector,” on page 124.

If you have any questions or concerns regarding the Customer Experience Improvement Program for Log Insight, contact bde-info@vmware.com.

Deploy the Big Data Extensions vApp in the vSphere Web Client

Deploying the Big Data Extensions vApp is the first step in getting your cluster up and running with Big Data Extensions.

Prerequisites

- Install and configure vSphere.
  - Configure all ESXi hosts to use the same NTP server.
  - On each ESXi host, add the NTP server to the host configuration, and from the host configuration's Startup Policy list, select **Start and stop with host**. The NTP daemon ensures that time-dependent processes occur in sync across hosts.
  - When installing Big Data Extensions on vSphere 5.1 or later, use vCenter Single Sign-On to provide user authentication.
- Verify that you have one vSphere Enterprise license for each host on which you deploy virtual Hadoop nodes. You manage your vSphere licenses in the vSphere Web Client or in vCenter Server.
- Install the Client Integration plug-in for the vSphere Web Client. This plug-in enables OVF deployment on your local file system.

  NOTE Depending on the security settings of your browser, you might have to approve the plug-in when you use it the first time.

- Download the Big Data Extensions OVA from the VMware download site.
- Verify that you have at least 40GB disk space available for the OVA. You need additional resources for the Hadoop cluster.
- Ensure that you know the vCenter Single Sign-On Look-up Service URL for your vCenter Single Sign-On service.

  If you are installing Big Data Extensions on vSphere 5.1 or later, ensure that your environment includes vCenter Single Sign-On. Use vCenter Single Sign-On to provide user authentication on vSphere 5.1 or later.
- Review the Customer Experience Improvement Program description, and determine if you wish to collect data and send it to VMware help improve your user experience using Big Data Extensions. See "The Customer Experience Improvement Program," on page 23.

Procedure

1. In the vSphere Web Client vCenter Hosts and Clusters view, select **Actions > All vCenter Actions > Deploy OVF Template**.
2. Choose the location where the Big Data Extensions OVA resides and click **Next**.

Option	Description
Deploy from File	Browse your file system for an OVF or OVA template.
Deploy from URL	Type a URL to an OVF or OVA template located on the internet. For example: `http://vmware.com/VMTN/appliance.ovf`.

3. View the OVF Template Details page and click **Next**.
4. Accept the license agreement and click **Next**.
5. Specify a name for the vApp, select a target datacenter for the OVA, and click **Next**.

   The only valid characters for Big Data Extensions vApp names are alphanumeric and underscores. The vApp name must be < 60 characters. When you choose the vApp name, also consider how you will name your clusters. Together the vApp and cluster names must be < 80 characters.
6. Select a vSphere resource pool for the OVA and click **Next**.

   Select a top-level resource pool. Child resource pools are not supported by Big Data Extensions even though you can select a child resource pool. If you select a child resource pool, you will not be able to create clusters from Big Data Extensions.
7. Select shared storage for the OVA and click **Next**.

   If shared storage is not available, local storage is acceptable.
8. For each network specified in the OVF template, select a network in the **Destination Networks** column in your infrastructure to set up the network mapping.

   The first network lets the Management Server communicate with your Hadoop cluster. The second network lets the Management Server communicate with vCenter Server. If your vCenter Server deployment does not use IPv6, you can specify the same IPv4 destination network for use by both source networks.

9 Configure the network settings for your environment, and click **Next**.

   a Enter the network settings that let the Management Server communicate with your Hadoop cluster.

      Use a static IPv4 (IP) network. An IPv4 address is four numbers separated by dots as in aaa.bbb.ccc.ddd, where each number ranges from 0 to 255. You must enter a netmask, such as 255.255.255.0, and a gateway address, such as 192.168.1.253.

      If the vCenter Server or any ESXi host or Hadoop distribution repository is resolved using a fully qualified domain name (FQDN), you must enter a DNS address. Enter the DNS server IP address as **DNS Server 1**. If there is a secondary DNS server, enter its IP address as **DNS Server 2**.

      NOTE You cannot use a shared IP pool with Big Data Extensions.

   b (Optional) If you are using IPv6 between the Management Server and vCenter Server, select the **Enable Ipv6 Connection** checkbox.

      Enter the IPv6 address, or FQDN, of the vCenter Server. The IPv6 address size is 128 bits. The preferred IPv6 address representation is: xxxx:xxxx:xxxx:xxxx:xxxx:xxxx:xxxx:xxxx where each x is a hexadecimal digit representing 4 bits. IPv6 addresses range from 0000:0000:0000:0000:0000:0000:0000:0000 to ffff:ffff:ffff:ffff:ffff:ffff:ffff:ffff. For convenience, an IPv6 address may be abbreviated to shorter notations by application of the following rules.

      - Remove one or more leading zeroes from any groups of hexadecimal digits. This is usually done to either all or none of the leading zeroes. For example, the group 0042 is converted to 42.
      - Replace consecutive sections of zeroes with a double colon (::). You may only use the double colon once in an address, as multiple uses would render the address indeterminate. RFC 5952 recommends that a double colon not be used to denote an omitted single section of zeroes.

      The following example demonstrates applying these rules to the address `2001:0db8:0000:0000:0000:ff00:0042:8329`.

      - Removing all leading zeroes results in the address `2001:db8:0:0:0:ff00:42:8329`.
      - Omitting consecutive sections of zeroes results in the address `2001:db8::ff00:42:8329`.

      See RFC 4291 for more information on IPv6 address notation.

10 Verify that the **Initialize Resources** check box is selected and click **Next**.

   If the check box is unselected, the resource pool, data store, and network connection assigned to the vApp will not be added to Big Data Extensions.

   If you do not add the resource pool, datastore, and network when you deploy the vApp, use the vSphere Web Client or the Serengeti CLI Client to specify the resource pool, datastore, and network information before you create a Hadoop cluster.

11 Run the vCenter Single Sign-On Lookup Service URL to enable vCenter Single Sign-On.

   - If you use vCenter 5.x, use the following URL: `https://`*`FQDN_or_IP_of_SSO_SERVER`*`:7444/lookupservice/sdk`
   - If you use vCenter 6.0, use the following URL: `https://`*`FQDN_of_SSO_SERVER`*`:443/lookupservice/sdk`

   If you don't input the URL, vCenter Single Sign-On is disabled.

12 To disable the Big Data Extensions data collector, uncheck the Customer Experience Improvement Program checkbox.

13 (Optional) To disable the Big Data Extensions Web plug-in from automatically registering, uncheck the enable checkbox.

    By default the checkbox to enable automatic registration of the Big Data Extensions Web plug-in is selected. When you first login to the Big Data Extensions Web client, it automatically connects to the Serengeti management server.

14 Specify a remote syslog server, such as VMware vRealize Log Insight, to which Big Data Extensions can send logging information to across the network.

    Retention, rotation and the splitting of logs received and managed by a syslog server are controlled by that syslog server. Big Data Extensions cannot configure or control log management on a remote syslog server. For more information on log management, see the documentation for the syslog server.

    Regardless of the additional syslog configuration specified with this option, logs continue to be placed in the default locations of the Big Data Extensions environment.

15 Verify the vService bindings and click **Next**.

16 Verify the installation information and click **Finish**.

    vCenter Server deploys the Big Data Extensions vApp. When deployment finishes, two virtual machines are available in the vApp.

    - The Management Server virtual machine, management-server (also called the Serengeti Management Server), which is started as part of the OVA deployment.
    - The Hadoop Template virtual machine, hadoop-template, which is not started. Big Data Extensions clones Hadoop nodes from this template when provisioning a cluster. Do not start or stop this virtual machine without good reason. The template does not include a Hadoop distribution.

IMPORTANT Do not delete any files under the `/opt/serengeti/.chef` directory. If you delete any of these files, such as the `serengeti.pem` file, subsequent upgrades to Big Data Extensions might fail without displaying error notifications.

What to do next

Install the Big Data Extensions plug-in within the vSphere Web Client. See "Install the Big Data Extensions Plug-In," on page 28.

If the **Initialize Resources** check box is not selected, add resources to the Big Data Extensions server before you create a Hadoop cluster.

Install RPMs in the Serengeti Management Server Yum Repository

Install the wsdl4j and mailx Red Hat Package Manager (RPM) packages within the internal Yum repository of the Serengeti Management Server.

The wsdl4j and mailx RPM packages are not embedded within big Data Extension due to licensing agreements. For this reason you must install them within the internal Yum repository of the Serengeti Management Server.

Prerequisites

Deploy the Big Data Extensions vApp.

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as the user `serengeti`.

2 Download and install the wsdl4j and mailx RPM packages.

- If the Serengeti Management Server can connect to the Internet, run the commands as shown in the example below to download the RPMs, copy the files to the required directory, and create a repository.

  ```
  cd /opt/serengeti/www/yum/repos/centos/6/base/RPMS/
  wget http://mirror.centos.org/centos/6/os/x86_64/Packages/mailx-12.4-7.el6.x86_64.rpm
  wget http://mirror.centos.org/centos/6/os/x86_64/Packages/wsdl4j-1.5.2-7.8.el6.noarch.rpm
  createrepo ..
  ```

- If the Serengeti Management Server cannot connect to the Internet, you must run the following tasks manually.

a Download the RPM files as shown in the example below.

```
http://mirror.centos.org/centos/6/os/x86_64/Packages/mailx-12.4-7.el6.x86_64.rpm
http://mirror.centos.org/centos/6/os/x86_64/Packages/wsdl4j-1.5.2-7.8.el6.noarch.rpm
```

b Copy the RPM files to `/opt/serengeti/www/yum/repos/centos/6/base/RPMS/`.

c Run the `createrepo` command to create a repository from the RPMs you downloaded.

```
createrepo /opt/serengeti/www/yum/repos/centos/6/base/
```

Install the Big Data Extensions Plug-In

To enable the Big Data Extensions user interface for use with a vCenter Server Web Client, register the plug-in with the vSphere Web Client. The Big Data Extensions graphical user interface is supported only when you use vSphere Web Client 5.1 and later.

The Big Data Extensions plug-in provides a GUI that integrates with the vSphere Web Client. Using the Big Data Extensions plug-in interface you can perform common Hadoop infrastructure and cluster management tasks.

Note Use only the Big Data Extensions plug-in interface in the vSphere Web Client or the Serengeti CLI Client to monitor and manage your Big Data Extensions environment. Performing management operations in vCenter Server might cause the Big Data Extensions management tools to become unsynchronized and unable to accurately report the operational status of your Big Data Extensions environment.

Prerequisites

- Deploy the Big Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24.
- By default, the Big Data Extensions Web plug-in automatically installs and registers when you deploy the Big Data Extensions vApp. To install the Big Data Extensions Web plug-in after deploying the Big Data Extensions vApp, you must has opted not to enable automatic registration of the Web plug-in during deployment. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24.
- Ensure that you have login credentials with administrator privileges for the vCenter Server system with which you are registering Big Data Extensions.

  Note The user name and password you use to login cannot contain characters whose UTF-8 encoding is greater than 0x8000.

- If you want to use the vCenter Server IP address to access the vSphere Web Client, and your browser uses a proxy, add the vCenter Server IP address to the list of proxy exceptions.

Procedure

1 Open a Web browser and go to the URL of vSphere Web Client 5.1 or later.

```
https://hostname-or-ip-address:port/vsphere-client
```

The *hostname-or-ip-address* can be either the DNS hostname or IP address of vCenter Server. By default the port is 9443, but this might have changed during installation of the vSphere Web Client.

2 Enter the user name and password with administrative privileges that has permissions on vCenter Server, and click **Login**.

3 Using the vSphere Web Client Navigator pane, locate the ZIP file on the Serengeti Management Server that contains the Big Data Extensions plug-in to register to the vCenter Server.

You can find the Serengeti Management Server under the datacenter and resource pool to which you deployed it.

4 From the inventory tree, select **management-server** to display information about the Serengeti Management Server in the center pane.

Click the **Summary** tab in the center pane to access additional information.

5 Note the IP address of the Serengeti Management Server virtual machine.

6 Open a Web browser and go to the URL of the management-server virtual machine.

```
https://management-server-ip-address:8443/register-plugin
```

The *management-server-ip-address* is the IP address you noted in Step 5.

7 Enter the information to register the plug-in.

Option	Action
Register or Unregister	Click **Install** to install the plug-in. Select **Uninstall** to uninstall the plug-in.
vCenter Server host name or IP address	Enter the server host name or IP address of vCenter Server. Do not include `http://` or `https://` when you enter the host name or IP address.
User Name and Password	Enter the user name and password with administrative privileges that you use to connect to vCenter Server. The user name and password cannot contain characters whose UTF-8 encoding is greater than 0x8000.
Big Data Extensions Package URL	Enter the URL with the IP address of the management-server virtual machine where the Big Data Extensions plug-in package is located: **`https://management-server-ip-address/vcplugin/serengeti-plugin.zip`**

8 Click **Submit**.

The Big Data Extensions plug-in registers with vCenter Server and with the vSphere Web Client.

9 Log out of the vSphere Web Client, and log back in using your vCenter Server user name and password.

The Big Data Extensions icon appears in the list of objects in the inventory.

10 Click **Big Data Extensions** in the Inventory pane.

What to do next

Connect the Big Data Extensions plug-in to the Big Data Extensions instance that you want to manage by connecting to the corresponding Serengeti Management Server. See "Connect to a Serengeti Management Server," on page 30.

Configure vCenter Single Sign-On Settings for the Serengeti Management Server

If the Big Data Extensions Single Sign-On (SSO) authentication settings are not configured or if they change after you install the Big Data Extensions plug-in, you can use the Serengeti Management Server Administration Portal to enable SSO, update the certificate, and register the plug-in so that you can connect to the Serengeti Management Server and continue managing clusters.

The SSL certificate for the Big Data Extensions plug-in can change for many reasons. For example, you install a custom certificate or replace an expired certificate.

Prerequisites

- Ensure that you know the IP address of the Serengeti Management Server to which you want to connect.
- Ensure that you have login credentials for the Serengeti Management Server `root` user.

Procedure

1. Open a Web browser and go the URL of the Serengeti Management Server Administration Portal.

   `https://management-server-ip-address:5480`

2. Type **root** for the user name, type the password, and click **Login**.
3. Select the SSO tab.
4. Do one of the following.

Option	Description
Update the certificate	Click **Update Certificate**.
Enable SSO for the first time	Type the **Lookup Service URL**, and click **Enable SSO**.

The Big Data Extensions and vCenter SSO server certificates are synchronized.

What to do next

Reregister the Big Data Extensions plug-in with the Serengeti Management Server. See "Connect to a Serengeti Management Server," on page 30.

Connect to a Serengeti Management Server

To use the Big Data Extensions plug-in to manage and monitor big data clusters and Hadoop distributions, you must connect the Big Data Extensions plug-in to the Serengeti Management Server in your Big Data Extensions deployment.

You can deploy multiple instances of the Serengeti Management Server in your environment. However, you can connect the Big Data Extensions plug-in with only one Serengeti Management Server instance at a time. You can change which Serengeti Management Server instance the plug-in connects to, and use the Big Data Extensions plug-in interface to manage and monitor multiple Hadoop and HBase distributions deployed in your environment.

IMPORTANT The Serengeti Management Server that you connect to is shared by all users of the Big Data Extensions plug-in interface in the vSphere Web Client. If a user connects to a different Serengeti Management Server, all other users are affected by this change.

Prerequisites

- Verify that the Big Data Extensions vApp deployment was successful and that the Serengeti Management Server virtual machine is running.
- Verify that the version of the Serengeti Management Server and the Big Data Extensions plug-in is the same.
- Ensure that vCenter Single Sign-On is enabled and configured for use by Big Data Extensions for vSphere 5.1 and later.
- Install theBig Data Extensions plug-in.

Procedure

1. Use the vSphere Web Client to log in to vCenter Server.
2. Select **Big Data Extensions**.
3. Click the **Summary** tab.
4. In the Connected Server pane, click the **Connect Server** link.
5. Navigate to the Serengeti Management Server virtual machine in the Big Data Extensions vApp to which to connect, select it, and click **OK**.

   The Big Data Extensions plug-in communicates using SSL with the Serengeti Management Server. When you connect to a Serengeti server instance, the plug-in verifies that the SSL certificate in use by the server is installed, valid, and trusted.

The Serengeti server instance appears as the connected server on the **Summary** tab of the Big Data Extensions Home page.

What to do next

You can add resource pool, datastore, and network resources to your Big Data Extensions deployment, and create big data clusters that you can provision for use.

Install the Serengeti Remote Command-Line Interface Client

Although theBig Data Extensions Plug-in for vSphere Web Client supports basic resource and cluster management tasks, you can perform a greater number of the management tasks using the Serengeti CLI Client.

Prerequisites

- Verify that the Big Data Extensions vApp deployment was successful and that the Management Server is running.
- Verify that you have the correct user name and password to log into the Serengeti CLI Client. If you are deploying on vSphere 5.1 or later, the Serengeti CLI Client uses your vCenter Single Sign-On credentials.
- Verify that the Java Runtime Environment (JRE) is installed in your environment, and that its location is in your PATH environment variable.

Procedure

1. Use the vSphere Web Client to log in to vCenter Server.
2. Select **Big Data Extensions**.
3. Click the **Getting Started** tab, and click the **Download Serengeti CLI Console** link.

   A ZIP file containing the Serengeti CLI Client downloads to your computer.

4 Unzip and examine the download, which includes the following components in the `cli` directory.

- The `serengeti-cli-`*`version`* JAR file, which includes the Serengeti CLI Client.
- The `samples` directory, which includes sample cluster configurations.
- Libraries in the `lib` directory.

5 Open a command shell, and navigate to the directory where you unzipped the Serengeti CLI Client download package.

6 Change to the `cli` directory, and run the following command to open the Serengeti CLI Client:

```
java -jar serengeti-cli-version.jar
```

What to do next

To learn more about using the Serengeti CLI Client, see the *VMware vSphere Big Data Extensions Command-line Interface Guide*.

Access the Serengeti CLI By Using the Remote CLI Client

You can access the Serengeti Command-Line Interface (CLI) to perform Serengeti administrative tasks with the Serengeti Remote CLI Client.

Prerequisites

- Use the VMware vSphere Web Client to log in to the VMware vCenter Server® on which you deployed the Serengeti vApp.
- Verify that the Serengeti vApp deployment was successful and that the Management Server is running.
- Verify that you have the correct password to log in to Serengeti CLI. See the *VMware vSphere Big Data Extensions Administrator's and User's Guide*.

  The Serengeti CLI uses its vCenter Server credentials.
- Verify that the Java Runtime Environment (JRE) is installed in your environment and that its location is in your path environment variable.

Procedure

1 Download the Serengeti CLI package from the Serengeti Management Server.

Open a Web browser and navigate to the following URL: `https://`*`server_ip_address`*`/cli/VMware-Serengeti-CLI.zip`

2 Download the ZIP file.

The filename is in the format `VMware-Serengeti-cli-`*`version_number-build_number`*`.ZIP`.

3 Unzip the download.

The download includes the following components.

- The `serengeti-cli-`*`version_number`* JAR file, which includes the Serengeti Remote CLI Client.
- The `samples` directory, which includes sample cluster configurations.
- Libraries in the `lib` directory.

4 Open a command shell, and change to the directory where you unzipped the package.

5 Change to the `cli` directory, and run the following command to enter the Serengeti CLI.

- For any language other than French or German, run the following command.

  ```
  java -jar serengeti-cli-version_number.jar
  ```

- For French or German languages, which use code page 850 (CP 850) language encoding when running the Serengeti CLI from a Windows command console, run the following command.

  ```
  java -Dfile.encoding=cp850 -jar serengeti-cli-version_number.jar
  ```

6 Connect to the Serengeti service.

You must run the `connect host` command every time you begin a CLI session, and again after the 30 minute session timeout. If you do not run this command, you cannot run any other commands.

a Run the `connect` command.

```
connect --host xx.xx.xx.xx:8443
```

b At the prompt, type your user name, which might be different from your login credentials for the Serengeti Management Server.

Note If you do not create a user name and password for the Serengeti Command-Line Interface Client, you can use the default vCenter Server administrator credentials. The Serengeti Command-Line Interface Client uses the vCenter Server login credentials with read permissions on the Serengeti Management Server.

c At the prompt, type your password.

A command shell opens, and the Serengeti CLI prompt appears. You can use the `help` command to get help with Serengeti commands and command syntax.

- To display a list of available commands, type `help`.
- To get help for a specific command, append the name of the command to the `help` command.

  ```
  help cluster create
  ```

- Press Tab to complete a command.

Upgrading Big Data Extensions 3

You can use VMware vSphere® Update Manager™ to upgrade Big Data Extensions from earlier versions.

This chapter includes the following topics:

Prepare to Upgrade Big Data Extensions

As a prerequisite to upgrading Big Data Extensions, you must prepare your system to ensure that you have all necessary software installed and configured properly, and that all components are in the correct state.

Data from nonworking Big Data Extensions deployments is not migrated during the upgrade process. If Big Data Extensions is not working and you cannot recover according to the troubleshooting procedures, do not try to perform the upgrade. Instead, uninstall the previous Big Data Extensions components and install the new version.

IMPORTANT Do not delete any files in the `/opt/serengeti/.chef` directory. If you delete any of these files, such as the `sernegeti.pem` file, subsequent upgrades to Big Data Extensions might fail without displaying error notifications.

Prerequisites

- Install vSphere Update Manager. For more information, see the vSphere Update Manager documentation.
- Verify that your previous Big Data Extensions deployment is working normally.
- Verify that you can create a default Hadoop cluster.

Procedure

1 Install vSphere Update Manager on a Windows Server.

  - Use the same version of vSphere Update Manager as vCenter Server. For example, if you are using vCenter Server 5.5, use vSphere Update Manager 5.5.

- vSphere Update Manager requires network connectivity with vCenter Server. Each vSphere Update Manager instance must be registered with a single vCenter Server instance.

2 Log in to vCenter Server with the vSphere Web Client.

3 Power on the Hadoop Template virtual machine.

4 If the Serengeti Management Server is configured to use a static IP network, make sure that the Hadoop Template virtual machine receives a valid IP address.

You must have a valid IP address and be connected to the network for the Hadoop Template virtual machine to connect to vSphere Update Manager.

5 Open a command shell and log in to the Serengeti Management Server as the user **serengeti**, and run the following series of commands.

```
/usr/sbin/vmware-rpctool 'info-set guestinfo.vmware.vami.version 2010100'
/usr/sbin/vmware-rpctool 'info-set guestinfo.vmware.vami.features SUP'
/usr/sbin/vmware-rpctool 'info-set guestinfo.vmware.vami.https-port 5489'
/usr/sbin/vmware-rpctool 'info-set guestinfo.vmware.vami.http-port 5488'
/usr/sbin/vmware-rpctool 'info-set guestinfo.vmware.vami.https-
clientkeyfile /opt/vmware/etc/sfcb/file.pem'
/usr/sbin/vmware-rpctool 'info-set guestinfo.vmware.vami.https-
clienttruststorefile /opt/vmware/etc/sfcb/client.pem'
```

6 For each cluster that is in AUTO scaling mode, change the scaling mode to MANUAL.

a Open a command shell and log in to the Serengeti Management Server as the user **serengeti**.

b Set the scaling mode of the cluster to MANUAL and the `--targetComputeNodeNum` parameter value to the number of provisioned compute nodes in the cluster.

```
cluster setParam --name cluster-name --elasticityMode manual --targetComputeNodeNum num-
provisioned-compute-nodes
```

7 Verify that all Hadoop clusters are in one of the following states:

- RUNNING
- STOPPED
- CONFIGURE_ERROR

If the status of the cluster is PROVISIONING, wait for the process to finish and for the state of the cluster to change to RUNNING.

8 Make sure that the host name of the Serengeti Management Server matches its fully qualified domain name (FQDN).

9 Access the Admin view of vSphere Update Manager.

a Start vSphere Update Manager.

b On the Home page of the vSphere Web Client, select **Hosts and Clusters**.

c Click **Update Manager**.

d Open the Admin view.

Perform the upgrade tasks in the Admin view.

Upgrade Big Data Extensions Virtual Appliance

You must perform several tasks to complete the upgrade of the Big Data Extensions virtual appliance. Because the versions of the virtual appliance, the Serengeti CLI, and the Big Data Extensions plug-in must all be the same, it is important that you upgrade all components to the new version.

Prerequisites

Complete the preparation steps for upgrading Big Data Extensions.

Procedure

1. Configure Proxy Settings on page 37

   You must have access to the Internet to upgrade your Big Data Extensions virtual appliance. If your site uses a proxy server to access the Internet, you must configure vSphere Update Manager to use the proxy server.

2. Download the Upgrade Source and Accept the License Agreement on page 38

   To start the Big Data Extensions upgrade process, you download the upgrade source and accept the license agreement (EULA).

3. Create an Upgrade Baseline on page 38

   When you upgrade Big Data Extensions virtual appliances, you must create a custom virtual appliance upgrade baseline.

4. Specify Upgrade Compliance Settings on page 39

   Upgrade compliance settings ensure that the upgrade baseline does not conflict with the current state of your Big Data Extensions virtual appliance.

5. Configure the Upgrade Remediation Task and Run the Upgrade Process on page 39

   The upgrade remediation task is the process by which vSphere Update Manager applies patches, extensions, and upgrades to the Big Data Extensions virtual appliance. You configure and run the remediation task to finish the Big Data Extensions virtual appliance upgrade process.

6. Replace the Hadoop Template Virtual Machine Under Big Data Extensions vApp on page 40

   To complete upgrade process, you must manually replace the virtual machine template after you upgrade the Big Data Extensions vApp.

Configure Proxy Settings

You must have access to the Internet to upgrade your Big Data Extensions virtual appliance. If your site uses a proxy server to access the Internet, you must configure vSphere Update Manager to use the proxy server.

If you do not use a proxy server, continue to "Download the Upgrade Source and Accept the License Agreement," on page 38.

Prerequisites

Verify that you have obtained the values for the proxy server URL and port from your network administrator.

Procedure

1. In the Admin view of vSphere Update Manager, click **Configuration** and then select **Download Settings**.
2. In the Proxy Settings section, click **Use proxy**.
3. Enter the values for the proxy URL and port.

4 Click **Test Connection** to ensure that the settings are correct.

5 If the settings are correct, click **Apply**.

vSphere Update Manager can now access the Web using the proxy server for your site.

Download the Upgrade Source and Accept the License Agreement

To start the Big Data Extensions upgrade process, you download the upgrade source and accept the license agreement (EULA).

Prerequisites

Verify that you have the URL from which to download the upgrade source.

For information about the upgrade source and the URL where you can download the upgrade source, see the VMware knowledge base article at `http://kb.vmware.com/` and search on article number 1004543.

Procedure

1 In the Admin view of vSphere Update Manager, click **Configuration** and then select **Download Settings**.

2 On the Download Settings page, click **Add Download Source**.

3 Enter the upgrade source URL in the **Source URL** text box.

4 Click **Validate URL** to verify connectivity to the upgrade URL.

5 Click **OK** to add the download source to vSphere Update Manager.

6 Click **Apply**.

7 Click **Download Now**.

8 On the **VA Upgrades** tab, select the upgrade.

9 Click **EULA** to accept the end user license agreement.

The upgrade source is downloaded.

Create an Upgrade Baseline

When you upgrade Big Data Extensions virtual appliances, you must create a custom virtual appliance upgrade baseline.

Prerequisites

Verify that you are logged in to a vSphere Web Client as an administrator and that the vSphere Web Client is connected to a vCenter Server system with which vSphere Update Manager is registered.

Procedure

1 On the **Baselines and Groups** tab, click **VMs/Vas** to review the existing baselines and groups.

2 Click **Create**.

3 Enter a meaningful name, such as Big Data Extensions VA Upgrade 1.5, and click **Next**.

4 Click **Add Multiple Rules** to create a set of rules that determine the target upgrade version for virtual appliances.

5 Review the baseline settings and click **Finish.**

Specify Upgrade Compliance Settings

Upgrade compliance settings ensure that the upgrade baseline does not conflict with the current state of your Big Data Extensions virtual appliance.

Prerequisites

Verify that you are logged in to a vSphere Web Client as an administrator and that the vSphere Web Client is connected to a vCenter Server system with which vSphere Update Manager is registered.

Procedure

1 In vSphere Web Client, navigate to **VMs and Templates** and click **Upgrade Manager**.

2 Open **Compliance View** and select the virtual appliance to upgrade.

3 Click **Attach**.

4 Select the upgrade baseline.

5 Click **Attach** again.

6 Verify that the virtual appliance needs to be updated.

   a In the inventory list, right-click the baseline.

   b Select **Scan for Updates**.

   vSphere Update Manager scans the baseline against the virtual appliance and determines whether the virtual appliance is up-to-date with the latest Big Data Extensions version. A vSphere Update Manager scan result of 100 percent indicates that your Big Data Extensions version is up-to-date.

What to do next

If the Big Data Extensions virtual appliance is up-to-date, discontinue the upgrade process. If the Big Data Extensions virtual appliance is not up-to-date, continue the upgrade process.

Configure the Upgrade Remediation Task and Run the Upgrade Process

The upgrade remediation task is the process by which vSphere Update Manager applies patches, extensions, and upgrades to the Big Data Extensions virtual appliance. You configure and run the remediation task to finish the Big Data Extensions virtual appliance upgrade process.

Prerequisites

Verify that you logged in to a vSphere Web Client as an administrator and that the vSphere Web Client is connected to a vCenter Server system with which vSphere Update Manager is registered.

NOTE The upgrade can take a few hours to finish.

Procedure

1 In the left pane of the VMs and Templates view, right-click the virtual appliance to upgrade and select **Remediate.**

   All virtual machines and appliances in the container are also remediated.

2 On the Remediation Selection page of the Remediate wizard, select the baseline group and upgrade baselines to apply.

3 Select the virtual machines and appliances that you want to remediate and click **Next.**

4 On the Schedule page, enter a unique name and an optional description for the task.

5 Select **Immediately** to begin the upgrade process immediately after the configuration is finished and click **Next**.

6 Configure the rollback options.

7 Specify the snapshot backup to roll back to and click **Next.**

8 Review the task definition and click **Finish.**

   Big Data Extensions restarts when the upgrade remediation task finishes.

9 Verify that the Big Data Extensions virtual appliance upgrade was successful.

   You can view the version of the Big Data Extensions virtual appliance in the vCenter client.

Some upgrade process errors are written to the Serengeti virtual appliance deployment logs in vCenter Server rather than appearing as error messages.

Replace the Hadoop Template Virtual Machine Under Big Data Extensions vApp

To complete upgrade process, you must manually replace the virtual machine template after you upgrade the Big Data Extensions vApp.

Procedure

1 Download the Big Data Extensions template OVA file from the Big Data Extensions download page.

2 Login to the vSphere Client.

3 Delete the original hadoop-template virtual machine under the Big Data Extensions vApp or move it out of the Big Data Extensions vApp which has been upgraded.

4 Select **File > Deploy OVA Template**.

5 Enter the downloaded template OVA path in the **Deploy OVA Template** dialog.

6 Complete the steps in the OVA Deployment wizard.

   On the step to configure the name and location, enter **`hadoop-template`** or any other valid name. On the step to configure the resource pool, select the Big Data Extensions vApp which has been upgraded.

7 Login to the Big Data Extensions server via SSH.

8 Restart Big Data Extensions Web services: `sudo service tomcat restart`

Upgrade the Big Data Extensions Plug-in

You must use the same version of the Serengeti Management Server and the Big Data Extensions plug-in.

By default, the Big Data Extensions Web plug-in automatically installs and registers with the Serengeti Management Server when you deploy the Big Data Extensions vApp. If you chose not to install and register the Big Data Extensions Web plug-in when installing the Big Data Extensions vApp, you must perform this task to upgrade the plug-in.

Procedure

1 Open a Web browser and go to the URL of the Serengeti Management Server plug-in manager service.

```
https://management-server-ip-address:8443/register-plugin
```

2 Select **Uninstall** and click **Submit.**

3 Select **Install**.

4 Enter the information to register the new plug-in, and click **Submit**.

Upgrade the Serengeti CLI

The Serengeti CLI must be the same version as your Big Data Extensions deployment. If you run the CLI remotely to connect to the management server, you must upgrade the Serengeti CLI.

Procedure

1. Log in to the vSphere Web Client.
2. Select Big Data Extensions from the navigation panel.
3. Click the **Summary** tab.
4. In the Connected Server panel, click **Connect Server**.
5. Select the Serengeti Management Server virtual machine in the Big Data Extensions vApp to which you want to connect and click **OK**.
6. Click the **Getting Started** tab, and click **Download Serengeti CLI Console**.

   A ZIP file containing the Serengeti CLI Client downloads to your computer.
7. Unzip and examine the ZIP file, which includes the following components in the CLI directory:
   - The `serengeti-cli-version` JAR file, which includes the Serengeti CLI Client.
   - The samples directory, which includes sample cluster configurations.
   - Libraries in the `lib` directory.
8. Open a command shell and navigate to the directory where you unzipped the Serengeti CLI Client download package.
9. Change to the CLI directory, and run the following command to open the Serengeti CLI Client:

   ```
   java -jar serengeti-cli-version.jar
   ```

What to do next

1. If your clusters are deployed with a Hadoop Template virtual machine that has a customized version of the CentOS 6.x operating system that includes VMware Tools, you must customize a new CentOS 6.x template to use after you upgrade Big Data Extensions.
2. To enable the Serengeti Management Server to manage clusters that you created in a previous version of Big Data Extensions, you must upgrade each cluster.

Upgrade Big Data Extensions Virtual Machine Components by Using the Serengeti Command-Line Interface

To enable the Serengeti Management Server to manage clusters created in a previous version of Big Data Extensions, you must upgrade the components in the virtual machines of each cluster. The Serengeti Management Server uses these components to control the cluster nodes.

When you upgrade from an earlier version of Big Data Extensions, clusters that you need to upgrade are shown with an alert icon next to the cluster name. When you click the alert icon the error message "Upgrade the cluster to the latest version" displays as a tool tip. See "View Provisioned Clusters in the vSphere Web Client," on page 125.

You can also identify clusters you need to upgrade using the `cluster list` command. When you run the `cluster list` command, the message "Earlier" displays where the cluster version normally appears.

Prerequisites

- You must be upgrading a cluster that was created with a previous version of Big Data Extensions.

- Ensure that you have applied the security patch `BDE-2.0.0-Patch1-bash.tar`.

  For more information on the security patch `BDE-2.0.0-Patch1-bash.tar`, see the VMware knowledge base article at http://kb.vmware.com/kb/2091050.

Procedure

1 For each cluster that you created in a previous version of Big Data Extensions, make sure that all of the nodes of a cluster are in the following states: RUNNING, STOPPED, ERROR, CONFIGURE_ERROR and UPGRADE_ERROR.

  If a node does not have a valid IP address, it cannot be upgraded to the new version of Big Data Extensions virtual machine tools.

  a Log into the vSphere Web Client that is connected to vCenter Server and navigate to **Hosts and Clusters**.

  b Select the resource pool of the cluster, select the **Virtual Machines** tab, and power on the cluster's virtual machines.

  IMPORTANT It might take up to five minutes for vCenter Server to assign valid IP addresses to the Big Data cluster nodes. Do not perform the remaining upgrade steps until the nodes have received their IP addresses.

2 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

3 Run the `cluster upgrade` command for each cluster that was created with a previous version of Big Data Extensions.

4 If the upgrade fails for a node, make sure that the failed node has a valid IP address and then rerun the `cluster upgrade` command.

  You can rerun the command as many times as you need to upgrade all the nodes.

What to do next

Stop and restart your big data clusters.

Add a Remote Syslog Server

If you wish to use a remote syslog server after upgrading from earlier versions of Big Data Extensions, you must manually specify the remote syslog server you wish to use.

The retention, rotation and splitting of logs received and managed by a syslog server are controlled by that syslog server. Big Data Extensions cannot configure or control log management on a remote syslog server. For more information on log management, see the documentation for your syslog server.

Prerequisites

- Successfully upgrade to the current release of Big Data Extensions.
- Have a remote syslog server within your environment that Big Data Extensions can send logging information to.

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as the user **serengeti.**

2 Open the file `/etc/rsyslog.d/20-base.conf` in a text editor.

3 Edit the file to include the remote syslog service information.

```
*.* @syslog_ip_address:port_number
```

4 Restart the syslog service.

```
service rsyslog restart
```

Your upgraded Big Data Extensions deployment will send logging information to the remote syslog service you specify.

NOTE Regardless of the additional syslog configuration specified with this procedure, logs continue to be placed in the default locations of the Big Data Extensions environment. See "Log Files for Troubleshooting," on page 140.

Managing Application Managers 4

A key to managing your Hadoop clusters is understanding how to manage the different application managers that you use in your Big Data Extensions environment.

This chapter includes the following topics:

Add an Application Manager by Using the vSphere Web Client

To use either Cloudera Manager or Ambari application managers to manage clusters, you must add the application manager and add server information to Big Data Extensions.

Application manager names can include only alphanumeric characters ([0-9, a-z, A-Z]) and the following special characters; underscores, hyphens, and blank spaces.

Procedure

1 On the Big Data Extensions navigation pane, click **Application Managers**.

2 Click the **Add Application Manager** icon (+) at the top of the page to open the New Application Manager wizard.

3 Follow the prompts to complete the installation of the application manager.

You can use either http or https.

Option	Action
Use http	Enter the server URL with http. The SSL certification text box is disabled.
Use https	Enter the FQDN instead of the URL. The SSL certification text box is enabled.

The vSphere Web UI refreshes the Application Manager list and displays it in the List view.

Modify an Application Manager by Using the Web Client

You can modify the information for an application manager, for example, you can change the manager server IP address if it is not a static IP, or you can upgrade the administrator account.

Prerequisites

Verify that you have at least one external application manager installed on your Big Data Extensions environment.

Procedure

1. In the vSphere Web Client, click **Application Managers** in the navigation menu.
2. From the Application Managers list, right-click the application manager to modify and select **edit settings**.
3. In the Edit Application Manager dialog box, make the changes to the application manager and click **OK.**

Delete an Application Manager by Using the vSphere Web Client

You can delete an application manager with the vSphere Web Client when you no longer need it.

The process fails if the application manager you want to delete contains clusters.

Prerequisites

Verify that you have at least one external application manager installed in your Big Data Extensions environment.

Procedure

1. In the vSphere Web Client, click **Application Managers** in the navigation pane.
2. Right-click the application manager to delete and select **Delete**.

The application manager is removed from the Application Managers list panel.

View Application Managers and Distributions by Using the Web Client

You can view a list of the application managers and distributions that are currently being used in your Big Data Extensions environment.

Procedure

- From Big Data Extensions, click **Application Managers** from the **Inventory Lists**.

  A list opens that contains the distributions, descriptions, application managers, and how many clusters are managed by your Big Data Extensions environment.

View Roles for Application Manager and Distribution by Using the Web Client

You can use the Application Managers pane to view a list and the details of the Hadoop roles for a specific application manager and distribution.

Procedure

1. From Big Data Extensions, click **Inventory Lists > Application Managers**.

2 Select the application manager for which you want to view details.

The details pane opens that contains a list of supported distributions with the name, vendor, version and roles of the distribution.

Managing Hadoop Distributions 5

The Serengeti Management Server includes the Apache Bigtop distribution, but you can add any supported Hadoop distribution to your Big Data Extensions environment.

Procedure

1. Hadoop Distribution Deployment Types on page 50

   You can choose which Hadoop distribution to use when you deploy a cluster. The type of distribution you choose determines how you configure it for use with Big Data Extensions. When you deploy the Big Data Extensions vApp, the Bigtop 0.8.0 distribution is included in the OVA that you download and deploy.

2. Configure a Tarball-Deployed Hadoop Distribution by Using the Serengeti Command-Line Interface on page 50

   You can add and configure Hadoop distributions other than those included with the Big Data Extensions vApp using the command line. You can configure multiple Hadoop distributions from different vendors.

3. Configuring Yum and Yum Repositories on page 52

   You can deploy Cloudera CDH4 and CDH5, Apache Bigtop, MapR, and Pivotal PHD Hadoop distributions using Yellowdog Updater, Modified (yum). Yum enables automatic updates and package management of RPM-based software distributions. To deploy a Hadoop distribution using yum, you must create and configure a yum repository.

4. Create a Hadoop Template Virtual Machine using RHEL Server 6.x and VMware Tools on page 69

   You can create a Hadoop Template virtual machine that has a customized version of the Red Hat Enterprise Linux (RHEL) Server 6.x operating system that includes VMware Tools. Although only a few Hadoop distributions require a custom version of RHEL Server 6.x, you can customize RHEL Server 6.x for any Hadoop distribution.

5. Maintain a Customized Hadoop Template Virtual Machine on page 73

   You can modify or update the Hadoop Template virtual machine operating system. When you make updates, you must remove the snapshot that is created by the virtual machine.

Hadoop Distribution Deployment Types

You can choose which Hadoop distribution to use when you deploy a cluster. The type of distribution you choose determines how you configure it for use with Big Data Extensions. When you deploy the Big Data Extensions vApp, the Bigtop 0.8.0 distribution is included in the OVA that you download and deploy.

Depending on which Hadoop distribution you want to configure to use with Big Data Extensions, use either a tarball or yum repository to install your distribution. The table lists the supported Hadoop distributions, the distribution name, vendor abbreviation, and version number to use as input parameters when you configure the distribution for use with Big Data Extensions.

Table 5-1. Hadoop Deployment Types

Hadoop Distribution	Version Number	Vendor Abbreviation	Deployment Type	HVE Support?
Bigtop	0.8	BIGTOP	Yum	No
Pivotal HD	3.0	PHD	Yum	Yes
Hortonworks Data Platform	2.1.1 - 2.2	HDP	Yum	No
Cloudera	5.3, 5.4	CDH	Yum	No
MapR	4.0, 4.1	MAPR	Yum	No

About Hadoop Virtualization Extensions

Hadoop Virtualization Extensions (HVE), developed by VMware, improves Hadoop performance in virtual environments by enhancing Hadoop's topology awareness mechanism to account for the virtualization layer.

Configure Hadoop 2.x and Later Distributions with DNS Name Resolution

When you create clusters using Hadoop distributions based on Hadoop 2.0 and later, the DNS server in your network must provide forward and reverse FQDN/IP resolution. Without valid DNS and FQDN settings, the cluster creation process might fail, or the cluster is created but does not function. Hadoop distributions based on Hadoop 2.x and later include Apache Bigtop, Cloudera CDH4 and CDH5, Hortonworks HDP 2.x, and Pivotal PHD 1.1 and later releases.

Configure a Tarball-Deployed Hadoop Distribution by Using the Serengeti Command-Line Interface

You can add and configure Hadoop distributions other than those included with the Big Data Extensions vApp using the command line. You can configure multiple Hadoop distributions from different vendors.

Refer to your Hadoop distribution vendor's Web site to obtain the download URLs to use for the components that you want to install. If you are behind a firewall, you might need to modify your proxy settings to allow the download. Before you install and configure tarball-based deployments, ensure that you have the vendor's URLs from which to download the different Hadoop components. Use these URLs as input parameters to the `config-distro.rb` configuration utility.

If you have a local Hadoop distribution and your server does not have access to the Internet, you can manually upload the distribution.

Prerequisites

- Deploy the Big Data Extensions vApp.
- Review the different Hadoop distributions so you know which distribution name abbreviation, vendor name abbreviation, and version number to use as an input parameter, and whether the distribution supports Hadoop Virtualization Extension (HVE).

- (Optional) Set the password for the Serengeti Management Server.

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

2 Run the `/opt/serengeti/sbin/config-distro.rb` Ruby script.

```
config-distro.rb --name distro_name --vendor vendor_name --version version_number
--hadoop hadoop_package_url --pig pig_package_url --hive hive_package_url
--hbase hbase_package_url --zookeeper zookeeper_package_URL --hve {true | false} --yes
```

Option	Description
--name	Name to identify the Hadoop distribution that you are downloading. For example, **hdp** for Hortonworks. This name can include alphanumeric characters ([a-z], [A-Z], [0-9]) and underscores ("_").
--vendor	Vendor name whose Hadoop distribution you want to use. For example, **HDP** for Hortonworks.
--version	Version of the Hadoop distribution that you want to use. For example, **1.3**.
--hadoop	URL from which to download the Hadoop distribution tarball package from the Hadoop vendor's Web site.
--pig	URL from which to download the Pig distribution tarball package from the Hadoop vendor's Web site.
--hive	URL from which to download the Hive distribution tarball package from the Hadoop vendor's Web site.
--hbase	(Optional) URL from which to download the HBase distribution tarball package from the Hadoop vendor's Web site.
--zookeeper	(Optional) URL from which to download the ZooKeeper distribution tarball package from the Hadoop vendor's Web site.
--hve {true | false}	(Optional) Specifies whether the Hadoop distribution supports HVE
--yes	(Optional) Specifies that all confirmation prompts from the `config-distro.rb` script are answered with a "yes" response.

The example downloads the tarball version of Hortonworks Data Platform (HDP), which consists of Hortonworks Hadoop, Hive, HBase, Pig, and ZooKeeper distributions. Note that you must provide the download URL for each of the software components you wish to configure for use with Big Data Extensions.

```
config-distro.rb --name hdp --vendor HDP --version 1.3.2
--hadoop http://public-repo-1.hortonworks.com/HDP/centos6/1.x/updates/1.3.2.0/tars/
hadoop-1.2.0.1.3.2.0-111.tar.gz
--pig http://public-repo-1.hortonworks.com/HDP/centos6/1.x/updates/1.3.2.0/tars/
pig-0.11.1.1.3.2.0-111.tar.gz
--hive http://public-repo-1.hortonworks.com/HDP/centos6/1.x/updates/1.3.2.0/tars/
hive-0.11.0.1.3.2.0-111.tar.gz
--hbase http://public-repo-1.hortonworks.com/HDP/centos6/1.x/updates/1.3.2.0/tars/
hbase-0.94.6.1.3.2.0-111-security.tar.gz
--zookeeper http://public-repo-1.hortonworks.com/HDP/centos6/1.x/updates/1.3.2.0/tars/
zookeeper-3.4.5.1.3.2.0-111.tar.gz
--hve true
```

The script downloads the files.

3 When the download finishes, explore the /opt/serengeti/www/distros directory, which includes the following directories and files.

Item	Description
name	Directory that is named after the distribution. For example, apache.
manifest	The manifest file that is generated by config-distro.rb that is used to download the Hadoop distribution.
manifest.example	Example manifest file. This file is available before you perform the download. The manifest file is a JSON file with three sections: name, version, and packages.

4 To enable Big Data Extensions to use the added distribution, restart the tomcat service.

```
sudo /sbin/service tomcat restart
```

The Serengeti Management Server reads the revised manifest file and adds the distribution to those from which you can create a cluster.

5 Return to the Big Data Extensions Plug-in for vSphere Web Client, and click **Hadoop Distributions** to verify that the Hadoop distribution is available to use to create a cluster.

The distribution and the corresponding role appear.

The distribution is added to the Serengeti Management Server, but is not installed in the Hadoop Template virtual machine. The agent is preinstalled on each virtual machine that copies the distribution components that you specify from the Serengeti Management Server to the nodes during the Hadoop cluster creation process.

What to do next

You can add datastore and network resources for the Hadoop clusters that you create.

You can create and deploy big data clusters using your chosen Hadoop distribution.

Configuring Yum and Yum Repositories

You can deploy Cloudera CDH4 and CDH5, Apache Bigtop, MapR, and Pivotal PHD Hadoop distributions using Yellowdog Updater, Modified (yum). Yum enables automatic updates and package management of RPM-based software distributions. To deploy a Hadoop distribution using yum, you must create and configure a yum repository.

- Yum Repository Configuration Values on page 53

  To create a local yum repository, you create a configuration file that identifies the file and package names of a distribution to download and deploy. When you create the configuration file, you replace a set of placeholder values with values that correspond to your Hadoop distribution. The yum repositories are used to install or update Hadoop software on CentOS and other operating systems that use Red Hat Package Manager (RPM).

- Setup a Local Yum Repository for Apache Bigtop, Cloudera, Hortonworks, and MapR Hadoop Distributions on page 56

  Although publicly available yum repositories exist for Ambari, Apache Bigtop, Cloudera, Hortonworks, and MapReduce distributions, creating your own yum repository can result in faster download times and greater control over the repository.

- Setup a Local Yum Repository for the Pivotal Hadoop Distribution on page 58

  Pivotal does not provide a publicly accessible yum repository from which you can deploy and upgrade the Pivotal Hadoop software distribution. Therefore, you might want to download the Pivotal software tarballs and create your own yum repository for Pivotal which provides you with better access and control over installing and updating your Pivotal HD distribution software.

- Configure a Yum-Deployed Hadoop Distribution on page 60

  You can install Hadoop distributions that use yum repositories (as opposed to tarballs) for use with Big Data Extensions. When you create a cluster for a yum-deployed Hadoop distribution, the Hadoop nodes download and install Red Hat Package Manager (RPM) packages from the official yum repositories for a particular distribution or your local yum repositories.

- Set Up a Local Yum Repository for Cloudera Manager Application Manager on page 61

  When you create a new cluster with an external application manager, you must install agents and distribution packages on each cluster node. If the installation downloads the agents and packages from the Internet, the process might be slow. If you do not have an Internet connection, the cluster creation process is not possible. To avoid these problems, you can create a local yum repository.

- Set Up a Local Yum Repository for Ambari Application Manager on page 64

  When you create a new cluster with an external application manager, you must install agents and distribution packages on each cluster node. If the installation downloads the agents and packages from the Internet, the process might be slow. If you do not have an Internet connection, the cluster creation process is impossible. To avoid these problems, you can create a local yum repository.

Yum Repository Configuration Values

To create a local yum repository, you create a configuration file that identifies the file and package names of a distribution to download and deploy. When you create the configuration file, you replace a set of placeholder values with values that correspond to your Hadoop distribution. The yum repositories are used to install or update Hadoop software on CentOS and other operating systems that use Red Hat Package Manager (RPM).

The following tables list the values to use for the Ambari, Apache Bigtop, Cloudera, Hortonworks, MapR, and Pivotal distributions.

Note If you copy-and-paste values from the table, be sure to include all required information. Some values appear on two lines in the table, for example, "maptech maprecosystem", and they must be combined into a single line when you use them.

Apache Bigtop Yum Repository Configuration Values

Table 5-2. Apache Bigtop Yum Repository Placeholder Values

Placeholder	Value
repo_file_name	bigtop.repo

package_info	[bigtop]
	name=Bigtop
	enabled=1
	gpgcheck=1
	type=NONE
	baseurl=http://bigtop.s3.amazonaws.com/releases/0.7.0/redhat/6/x86_64
	gpgkey=http://archive.apache.org/dist/bigtop/KEYS
	Note If you use a version other than 0.7.0, use the exact version number of your Apache Bigtop distribution in the pathname.

mirror_cmds	reposync -r bigtop

Table 5-2. Apache Bigtop Yum Repository Placeholder Values (Continued)

Placeholder	Value
default_rpm_dir	bigtop
target_rpm_dir	bigtop

local_repo_info	[bigtop]
	name=Apache Bigtop
	baseurl=http://*ip_of_yum_repo_webserver*/bigtop/
	enabled=1
	gpgcheck=0

Cloudera Yum Repository Configuration Values

Table 5-3. Cloudera Yum Repository Placeholder Values

Placeholder	Value
repo_file_name	cloudera-cdh.repo

package_info	If you use CDH4, use the values below.
	[cloudera-cdh]
	name=Cloudera's Distribution for Hadoop
	http://archive.cloudera.com/cdh4/redhat/6/x86_64/cdh/4/
	gpkey=http://archive.cloudera.com/cdh4/redhat/6/x86_64/cdh/RPM-GPG-KEY-cloudera
	gpgcheck=1
	If you use CDH5, use the values below.
	[cloudera-cdh]
	name=Cloudera's Distribution for Hadoop
	baseurl=http://archive.cloudera.com/cdh5/redhat/6/x86_64/cdh/5/
	gpgkey=http://archive.cloudera.com/cdh5/redhat/6/x86_64/cdh/RPM-GPG-KEY-cloudera
	gpgcheck=1

mirror_cmds	reposync -r cloudera-cdh4
default_rpm_dir	cloudera-cdh/RPMS
target_rpm_dir	cdh/*version_number*

local_repo_info	[cloudera-cdh]
	name=Cloudera's Distribution for Hadoop
	baseurl=http://*ip_of_yum_repo_webserver*/cdh/*version_number*/
	enabled=1
	gpgcheck=0

Hortonworks Yum Repository Configuration Values

Table 5-4. Hortonworks Yum Repository Placeholder Values

Placeholder	Value
repo_file_name	hdp.repo

package_info	[hdp]
	name=Hortonworks Data Platform Version - HDP-2.1.1.0
	baseurl=http://public-repo-1.hortonworks.com/HDP/centos6/2.x/GA/2.1.1.0
	gpgcheck=1
	gpgkey=http://public-repo-1.hortonworks.com/HDP/centos6/2.x/GA/2.1.1.0/RPM-GPG-KEY/RPM-GPG-KEY-Jenkins
	enabled=1
	priority=1
	NOTE If you use a version other than HDP 2.1.1.0, use the exact version number of your Hortonworks distribution in the pathname.

mirror_cmds	reposync -r hdp
default_rpm_dir	hdp
target_rpm_dir	hdp/2

local_repo_info	[hdp]
	name=Hortonworks Data Platform Version -HDP-2.1.1.0
	baseurl=http://ip_of_yum_repo_webserver/hdp/2/
	enabled=1
	gpgcheck=0

MapR Yum Repository Configuration Values

Table 5-5. MapR Yum Repository Placeholder Values

Placeholder	Value
repo_file_name	mapr.repo

package_info	[maprtech]
	name=MapR Technologies
	baseurl=http://package.mapr.com/releases/3.1.0/redhat/
	enabled=1
	gpgcheck=0
	protect=1
	[maprecosystem]
	name=MapR Technologies
	baseurl=http://package.mapr.com/releases/ecosystem/redhat
	enabled=1
	gpgcheck=0
	protect=1
	NOTE If you use a version other than 3.1.0, use the exact version number of your MapR distribution in the pathname.

mirror_cmds	reposync -r maprtech reposync -r maprecosystem
default_rpm_dir	maprtech maprecosystem

Table 5-5. MapR Yum Repository Placeholder Values (Continued)

Placeholder	Value
target_rpm_dir	mapr/3

local_repo_info	[mapr]
	name=MapR Version 3
	baseurl=http://*ip_of_yum_repo_webserver*/mapr/3/
	enabled=1
	gpgcheck=0
	protect=1

Pivotal Yum Repository Configuration Values

Table 5-6. Pivotal Yum Repository Placeholder Values

Placeholder	Value
repo_file_name	phd.repo
package_info	Not Applicable
mirror_cmds	Not Applicable
default_rpm_dir	pivotal
target_rpm_dir	phd/1

local_repo_info	[pivotalhd]
	name=PHD Version 1.0
	baseurl=http://*ip_of_yum_repo_webserver*/phd/1/
	enabled=1
	gpgcheck=0

Setup a Local Yum Repository for Apache Bigtop, Cloudera, Hortonworks, and MapR Hadoop Distributions

Although publicly available yum repositories exist for Ambari, Apache Bigtop, Cloudera, Hortonworks, and MapReduce distributions, creating your own yum repository can result in faster download times and greater control over the repository.

Prerequisites

- High-speed Internet access.
- CentOS 6.x 64-bit or Red Hat Enterprise Linux (RHEL) 6.x 64-bit.

  The hadoop-template virtual machine in the Serengeti vApp contains CentOS 6.5 64-bit. You can clone the hadoop-template virtual machine to a new virtual machine and create the yum repository on it.
- An HTTP server with which to create the yum repository. For example, Apache HTTP server.
- If there is a firewall on your system, ensure that the firewall does not block the network port number used by your HTTP server proxy. Typically, this is port 80.
- Refer to the yum repository placeholder values to populate the variables required in the steps. See "Yum Repository Configuration Values," on page 53.

Procedure

1 If your yum repository server requires an HTTP proxy server, open a command shell, such as Bash or PuTTY, log in to the yum repository server, and run the following commands to export the `http_proxy` environment variable.

```
# switch to root user
sudo su
export http_proxy=http://host:port
```

Option	Description
host	The hostname or the IP address of the proxy server.
port	The network port number to use with the proxy server.

2 Install the HTTP server that you want to use as a yum server.

This example installs the Apache HTTP Server and enables the `httpd` server to start whenever the machine is restarted.

```
yum install -y httpd
/sbin/service httpd start
/sbin/chkconfig httpd on
```

3 Install the `yum-utils` and `createrepo` packages.

The `yum-utils` package contains the `reposync` command.

```
yum install -y yum-utils createrepo
```

4 Synchronize the yum server with the official yum repository of your preferred Hadoop vendor.

a Using a text editor, create the file `/etc/yum.repos.d/$repo_file_name`.

b Add the *`package_info`* content to the new file.

c Mirror the remote yum repository to the local machine by running the *`mirror_cmds`* for your distribution packages.

It might take several minutes to download the RPMs from the remote repository. The RPMs are placed in the `$default_rpm_dir` directories.

5 Create the local yum repository.

a Move the RPMs to a new directory under the Apache HTTP Server document root.

The default document root is `/var/www/html/`.

```
doc_root=/var/www/html
mkdir -p $doc_root/$target_rpm_dir
mv $default_rpm_dir $doc_root/$target_rpm_dir/
```

For example, the `mv` command for the MapR Hadoop distribution is the following:

```
mv maprtech maprecosystem $doc_root/mapr/3/
```

b Create a yum repository for the RPMs.

```
cd $doc_root/$target_rpm_dir
createrepo .
```

c Create a new file, `$doc_root/$target_rpm_dir/$repo_file_name`, and include the *`local_repo_info`*.

d From a different machine, ensure that you can download the repository file from `http://`*`ip_of_webserver target_rpm_dir//repo_file_name`*.

6 (Optional) Configure HTTP proxy.

If the virtual machines created by the Serengeti Management Server do not need an HTTP proxy to connect to the local yum repository, skip this step.

On the Serengeti Management Server, edit the `/opt/serengeti/conf/serengeti.properties` file and add the following content anywhere in the file or replace existing items:

```
# set http proxy server
serengeti.http_proxy = http://<proxy_server:port>

# set the FQDNs (or IPs if no FQDN) of the Serengeti Management Server and the
local yum repository servers for 'serengeti.no_proxy'.
The wildcard for matching multi IPs doesn't work.
serengeti.no_proxy = serengeti_server_fqdn_or_ip.
yourdomain.com, yum_server_fqdn_or_ip.
yourdomain.com
```

What to do next

Configure your Apache Bigtop, Cloudera, Hortonworks, or MapR deployment for use with Big Data Extensions. See "Configure a Yum-Deployed Hadoop Distribution," on page 60.

Setup a Local Yum Repository for the Pivotal Hadoop Distribution

Pivotal does not provide a publicly accessible yum repository from which you can deploy and upgrade the Pivotal Hadoop software distribution. Therefore, you might want to download the Pivotal software tarballs and create your own yum repository for Pivotal which provides you with better access and control over installing and updating your Pivotal HD distribution software.

Pivotal does not provide a publicly accessible yum repository from which you can deploy and upgrade the Pivotal Hadoop software distribution. You might want to download the Pivotal software tarballs, and create your own yum repository from which to deploy and configure the Pivotal Hadoop software.

Prerequisites

- High-speed Internet access.
- CentOS 6.x 64-bit or Red Hat Enterprise Linux (RHEL) 6.x 64-bit.

  The hadoop-template virtual machine in the Big Data Extensions vApp contains CentOS 6.5 64-bit. You can clone the hadoop-template virtual machine to a new virtual machine and create the yum repository on it.

  NOTE Because the Pivotal Hadoop distribution requires CentOS 6.2 64-bit version or 6.4 64-bit version (x86_64), the yum server that you create to deploy the distribution must also use a CentOS 6.x 64-bit operating system.

- An HTTP server with which to create the yum repository. For example, Apache HTTP server.
- If there is a firewall on your system, ensure that the firewall does not block the network port number used by your HTTP server proxy. Typically, this is port 80.

Procedure

1 If your yum repository server requires an HTTP proxy server, open a command shell, such as Bash or PuTTY, log in to the yum repository server, and run the following commands to export the `http_proxy` environment variable.

```
# switch to root user
sudo su
export http_proxy=http://host:port
```

Option	Description
host	The hostname or the IP address of the proxy server.
port	The network port number to use with the proxy server.

2 Install the HTTP server that you want to use with a yum server.

This example installs the Apache HTTP Server and enables the httpd server to start whenever the machine is restarted.

```
yum install -y httpd
/sbin/service httpd start
/sbin/chkconfig httpd on
```

3 Install the `yum-utils` and `createrepo` packages.

The `yum-utils` package includes the `reposync` command.

```
yum install -y yum-utils createrepo
```

4 Download the Pivotal HD 1.0 or 2.0 tarball from the Pivotal Web site.

5 Extract the tarball that you downloaded.

The tarball name might vary if you download a different version of Pivotal HD.

```
tar -xf phd_1.0.1.0-19_community.tar
```

6 Extract PHD_1.0.1_CE/PHD-1.0.1.0-19.tar to the *`default_rpm_dir`* directory.

For Pivotal Hadoop the *`default_rpm_dir`* directory is `pivotal`.

The version numbers of the tar that you extract might be different from those used in the example if an update has occurred.

```
tar -xf PHD_1.0.1_CE/PHD-1.0.1.0-19.tar -C pivotal
```

7 Create and configure the local yum repository.

a Move the RPMs to a new directory under the Apache HTTP Server document root.

The default document root is `/var/www/html/`.

```
doc_root=/var/www/html
mkdir -p $doc_root/$target_rpm_dir
mv $default_rpm_dir $doc_root/$target_rpm_dir/
```

This example moves the RPMs for the Pivotal Hadoop distribution.

```
mv pivotal $doc_root/phd/1/
```

b Create a yum repository for the RPMs.

```
cd $doc_root/$target_rpm_dir
createrepo .
```

c Create a file, `$doc_root/$target_rpm_dir/$repo_file_name`, and include the *local_repo_info*.

d From a different machine, ensure that you can download the repository file from http://*ip_of_webserver*/$target_rpm_dir/$repo_file_name.

8 (Optional) Configure an HTTP proxy.

If the virtual machines created by the Serengeti Management Server do not need an HTTP proxy to connect to the local yum repository, skip this step.

On the Serengeti Management Server, edit the file`/opt/serengeti/conf/serengeti.properties`, and add the following content anywhere in the file or replace existing items:

```
# set http proxy server
serengeti.http_proxy = http://<proxy_server:port>

# set the FQDNs (or IPs if no FQDN) of the Serengeti Management Server and the
local yum repository servers for 'serengeti.no_proxy'.
The wildcard for matching multi IPs doesn't work.
serengeti.no_proxy = serengeti_server_fqdn_or_ip.
yourdomain.com, yum_server_fqdn_or_ip.yourdomain.com
```

Configure a Yum-Deployed Hadoop Distribution

You can install Hadoop distributions that use yum repositories (as opposed to tarballs) for use with Big Data Extensions. When you create a cluster for a yum-deployed Hadoop distribution, the Hadoop nodes download and install Red Hat Package Manager (RPM) packages from the official yum repositories for a particular distribution or your local yum repositories.

Prerequisites

- Review the different Hadoop distributions so that you know which distribution name, vendor abbreviation, and version number to use as an input parameter, and whether the distribution supports Hadoop Virtualization Extensions.
- Create a local yum repository for your Hadoop distribution. Creating your own repository can result in better access and more control over the repository.

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

2 Run the `/opt/serengeti/sbin/config-distro.rb` Ruby script.

```
config-distro.rb --name distro_name --vendor vendor_abbreviation --version ver_number
--repos http://url_to_yum_repo/name.repo
```

Option	Description
--name	Name to identify the Hadoop distribution that you are downloading. For example, **chd4** for Cloudera CDH4. This name can include alphanumeric characters ([a-z], [A-Z], [0-9]) and underscores ("_").
--vendor	Abbreviation of vendor name whose Hadoop distribution you want to use. For example, **CDH**.
--version	Version of the Hadoop distribution that you want to use. For example, **4.6.0**.
--repos	URL from which to download the Hadoop distribution yum package. This URL can be a local yum repository that you create or a publicly accessible yum repository hosted by the software vendor.

This example adds the Apache Bigtop Hadoop Distribution to Big Data Extensions.

```
config-distro.rb --name bigtop --vendor BIGTOP --version 0.8.0
--repos http://url_to_yum_repo/bigtop.repo
```

The example adds the Cloudera CDH4 Hadoop distribution to Big Data Extensions.

```
config-distro.rb --name cdh4 --vendor CDH --version 4.6.0 --repos
http://url_to_yum_repo/cloudera-cdh4.repo
```

NOTE The `config-distro.rb` script downloads files only for tarball-deployed distributions. No files are downloaded for yum-deployed distributions.

This example adds the Hortonworks Hadoop Distribution to Big Data Extensions.

```
config-distro.rb --name hdp --vendor HDP --version 2.1.1
--repos http://url_to_yum_repo/hdp.repo
```

The example adds the MapR Hadoop distribution to Big Data Extensions.

```
config-distro.rb --name mapr --vendor MAPR --version 3.1.0 --repos
http://url_to_yum_repo/mapr.repo
```

This example adds the Pivotal Hadoop Distribution to Big Data Extensions.

```
config-distro.rb --name phd --vendor PHD --version 2.0
--repos http://url_to_yum_repo/phd.repo
```

3 To enable Big Data Extensions to use the new distribution, restart the Tomcat service.

```
sudo /sbin/service tomcat restart
```

The Serengeti Management Server reads the revised manifest file and adds the distribution to those from which you can create a cluster.

4 Return to the Big Data Extensions Plug-in for vSphere Web Client, and click **Hadoop Distributions** to verify that the Hadoop distribution is available.

What to do next

You can create Hadoop and HBase clusters.

Set Up a Local Yum Repository for Cloudera Manager Application Manager

When you create a new cluster with an external application manager, you must install agents and distribution packages on each cluster node. If the installation downloads the agents and packages from the Internet, the process might be slow. If you do not have an Internet connection, the cluster creation process is not possible. To avoid these problems, you can create a local yum repository.

Prepare the Software Environment for the Local Repository for Cloudera Manager

The first step to create a local yum repository for Cloudera Manager is to prepare the software environment by setting up necessary servers and directories.

Prerequisites

Verify that you have the following conditions in place.

- High-speed Internet access.
- CentOS 6.x 64-bit or Red Hat Enterprise Linux (RHEL) 6.x 64-bit.

  The hadoop-template virtual machine in the Serengeti vApp contains CentOS 6.5 64-bit. You can clone the hadoop-template virtual machine to a new virtual machine and create the yum repository on it.

- An HTTP server with which to create the yum repository. For example, Apache HTTP server.
- If your system has a firewall, ensure that the firewall does not block the network port number that your HTTP server proxy uses. Typically, this is port 80.
- For more information about the yum repository placeholder values, see "Yum Repository Configuration Values," on page 53.

Procedure

1 If your yum repository server requires an HTTP proxy server, perform the steps:

   a Open a command shell, such as Bash or PuTTY.

   b Log in to the yum repository server.

   c Export the `http_proxy` environment variable.

```
# switch to root user
sudo su
export http_proxy=http://host:port
```

Option	Description
host	The hostname or the IP address of the proxy server.
port	The network port number to use with the proxy server.

2 Install the HTTP server to use as a yum server.

This example installs the Apache HTTP Server and enables the `httpd` server to start whenever the machine restarts.

```
yum install -y httpd
/sbin/service httpd start
/sbin/chkconfig httpd on
```

3 Make the CentOS directory.

```
mkdir -p /var/www/html/yum/centos6
```

4 Make the Cloudera Manager directory.

```
mkdir -p /var/www/html/yum/cm
```

5 Install the createrepo RPM.

```
yum install -y createrepo
```

Set Up the Local CentOS Yum Repository

You must copy all the RPM packages from the CentOS 6 DVD ISO images to set up the local CentOS yum repository.

Prerequisites

Verify that you prepared the software environment for the CentOS yum repository creation, including the directories for CentOS and the application manager. Refer to your CentOS documentation.

Procedure

1 Download the `CentOS-6.5-x86_64-bin-DVD1.iso` and `CentOS-6.5-x86_64-bin-DVD2.iso` CentOS 6 DVD ISO images from the CentOS official website.

2 Download the ISO images to the virtual machine servers.

3 Copy all of the CentOS RPM packages to `/var/www/html/yum/centos6`.

```
mkdir /mnt/centos6-1
 mount -o loop CentOS-6.5-x86_64-bin-DVD1.iso /mnt/centos6-1
 cp /mnt/centos6-1/Packages/* /var/www/html/yum/centos6

    mkdir /mnt/centos6-2
 mount -o loop CentOS-6.5-x86_64-bin-DVD2.iso /mnt/centos6-2
 cp /mnt/centos6-2/Packages/* /var/www/html/yum/centos6
```

4 Create the CentOS 6 yum repository.

```
createrepo /var/www/html/yum/centos6
```

Download Packages for Cloudera Manager

After you set up the local CentOS yum repository, you must download the packages for Cloudera Manager.

Procedure

1 Download the `cm5.0.1-centos6.tar.gz` file.

```
wget http://archive-primary.cloudera.com/cm5/repo-as-tarball/5.0.1
/cm5.0.1-centos6.tar.gz
```

For other versions of Cloudera Manager, the URLs used in the example might vary.

2 Extract the tarball.

```
tar xzf cm5.0.1-centos6.tar.gz
```

For other versions of Cloudera Manager, the URLs used in the example might vary.

3 Copy all of the files in `cm/5.0.1/RPMS/x86_64/` to `/var/www/html/yum/cm/`

```
cp cm/5.0.1/RPMS/x86_64/* /var/www/html/yum/cm/
```

Configure the Yum Repository Server and the Local Parcel Repository

You must configure the yum repository server and the local parcel repository before you can distribute the parcels file.

Procedure

1 Create the yum repository.

The `repodata` directory is created in `/var/www/html/yum/`.

```
createrepo /var/www/html/yum
```

2 Ensure that you can access the URL http://*yum_repo_server_ip*/yum from a browser.

3 Create the `Parcels` directory.

```
mkdir -p /var/www/html/parcels
```

4 Change to the `Parcels` directory.

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

5 Download the `Parcels` file.

```
wget http://archive-primary.cloudera.com/cdh5/parcels/5.0.1/
CDH-5.0.1-1.cdh5.0.1.p0.47-el6.parcel
```

6 Download the `manifest.json` file.

```
wget http://archive-primary.cloudera.com/cdh5/parcels/5.0.1/manifest.json
```

7 In the `manifest.json` file, remove all items except for `CDH–5.0.1–1.cdh5.0.1.p0.47–el6.parcel`

8 Open a browser, go to http://*your_cloudera_manager_server*:7180/cmf/parcel/status and click **Edit Settings**.

9 Select one minute in the **Parcel Update Frequency** text box.

10 Remove the remote parcel repository URL that was replaced by the target parcel URL.

11 Add the URL http://*yum_repo_server_ip*/parcels.

You can now create clusters for the Cloudera Manager by using the local yum repository.

Set Up a Local Yum Repository for Ambari Application Manager

When you create a new cluster with an external application manager, you must install agents and distribution packages on each cluster node. If the installation downloads the agents and packages from the Internet, the process might be slow. If you do not have an Internet connection, the cluster creation process is impossible. To avoid these problems, you can create a local yum repository.

Prepare the Software Environment for the Local Repository for Ambari

The first step to create a local yum repository for Ambari is to prepare the software environment.

Prerequisites

Verify that you have the following conditions in place.

- High-speed Internet access.
- CentOS 6.x 64-bit or Red Hat Enterprise Linux (RHEL) 6.x 64-bit.

  The hadoop-template virtual machine in the Serengeti vApp contains CentOS 6.5 64-bit. You can clone the hadoop-template virtual machine to a new virtual machine and create the yum repository on it.
- An HTTP server with which to create the yum repository. For example, Apache HTTP server.
- If your system has a firewall, ensure that the firewall does not block the network port number that your HTTP server proxy uses. Typically, this is port 80.
- For more information about the yum repository placeholder values, see "Yum Repository Configuration Values," on page 53.

Procedure

1 If your yum repository server requires an HTTP proxy server, open a command shell, such as Bash or PuTTY, log in to the yum repository server, and export the `http_proxy` environment variable.

```
# switch to root user
sudo su
export http_proxy=http://host:port
```

Option	Description
host	The hostname or the IP address of the proxy server.
port	The network port number to use with the proxy server.

2 Install the HTTP server to use as a yum server.

This example installs the Apache HTTP Server and enables the `httpd` server to start whenever the machine restarts.

```
yum install -y httpd
/sbin/service httpd start
/sbin/chkconfig httpd on
```

3 Make the CentOS directory.

```
mkdir -p /var/www/html/yum/centos6
```

4 Make the Ambari directory.

```
mkdir -p /var/www/html/yum/ambari
```

5 Install the createrepo RPM.

```
yum install -y createrepo
```

Set Up the Local CentOS Yum Repository

You must copy all the RPM packages from the CentOS 6 DVD ISO images to set up the local CentOS yum repository.

Prerequisites

Verify that you prepared the software environment for the CentOS yum repository creation, including the directories for CentOS and the application manager. Refer to your CentOS documentation.

Procedure

1 Download the `CentOS-6.5-x86_64-bin-DVD1.iso` and `CentOS-6.5-x86_64-bin-DVD2.iso` CentOS 6 DVD ISO images from the CentOS official website.

2 Download the ISO images to the virtual machine servers.

3 Copy all of the CentOS RPM packages to `/var/www/html/yum/centos6`.

```
mkdir /mnt/centos6-1
 mount -o loop CentOS-6.5-x86_64-bin-DVD1.iso /mnt/centos6-1
 cp /mnt/centos6-1/Packages/* /var/www/html/yum/centos6

    mkdir /mnt/centos6-2
 mount -o loop CentOS-6.5-x86_64-bin-DVD2.iso /mnt/centos6-2
 cp /mnt/centos6-2/Packages/* /var/www/html/yum/centos6
```

4 Create the CentOS 6 yum repository.

```
createrepo /var/www/html/yum/centos6
```

Download Packages for Ambari

After you set up the local CentOS yum repository, you must download the packages for the Ambari application manager.

Procedure

1 Go to `/var/www/html/yum/ambari`.

2 Download the Ambari agent.

```
wget http://s3.amazonaws.com/public-repo-1.hortonworks.com/ambari/centos6/
ambari-1.6.0-centos6.tar.gz
```

If you use other versions of Ambari, for example Ambari 1.6.1, the URL that you use might vary.

3 Download the HDP packages.

```
wget http://s3.amazonaws.com/public-repo-1.hortonworks.com/HDP/
centos6/HDP-2.1.2.0-centos6-rpm.tar.gz
```

If you use other versions of HDP, for example HDP 1.3.2 or HDP 2.0, the URL that you use might vary.

4 Download the HDP-UTILS packages.

```
wget http://s3.amazonaws.com/public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.16/
repos/centos6/HDP-UTILS-1.1.0.16-centos6.tar.gz
wget http://s3.amazonaws.com/public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.17/
repos/centos6/HDP-UTILS-1.1.0.17-centos6.tar.gz
```

5 Extract all of the tarball files and copy them to `/var/www/html/yum/ambari`.

Configure the Ambari Repository File on the Ambari Server

To set up the local yum repository, you must configure the Ambari repository file.

Procedure

1 Stop the Ambari server..

```
ambari-server stop
```

2 Download the `ambari.repo` file.

```
wget http://s3.amazonaws.com/public-repo-1.hortonworks.com/ambari/centos6/1.x/updates/
1.6.0/ambari.repo
```

3 Move the `ambari.repo` file to the `/etc/yum.repos.d/` directory.

4 Edit the `ambari.repo` file.

a Replace the URLs with the yum repository server address.

b Remove the group check.

c Add a new section for CentOS.

Example: Configuring the Ambari Repository File on the Ambari Server

```
[centos]
   name=centos6
   baseurl=http://<yum_repo_server_ip>/yum/centos6/
   gpgcheck=0
   enabled=1

[ambari-1.x]
   name=Ambari 1.x
   baseurl=http://<yum_repo_server_ip>/yum/ambari/ambari/centos6/1.x/updates/1.6.0/
   gpgcheck=0
   enabled=1
   priority=1

[HDP-UTILS-1.1.0.16]
   name=Hortonworks Data Platform Utils Version - HDP-UTILS-1.1.0.16
   baseurl=http://<yum_repo_server_ip>/yum/ambari/HDP-UTILS-1.1.0.16/repos/centos6/
   gpgcheck=0
   enabled=1
   priority=1

[HDP-UTILS-1.1.0.17]
   name=Hortonworks Data Platform Utils Version - HDP-UTILS-1.1.0.17
   baseurl=http://<yum_repo_server_ip>/yum/ambari/HDP-UTILS-1.1.0.17/repos/centos6/
   gpgcheck=0
   enabled=1
   priority=1
```

Configure the HDP Repository URL on the Local Yum Server

After you configure the Ambari repository on the Ambari server, you must configure the HDP repository URL on the local yum server.

Prerequisites

Verify that you configured the Ambari repository on the Ambari server.

Procedure

1. Get the `hdp_urlinfo.json` file from `http://public-repo-1.hortonworks.com/HDP/hdp_urlinfo.json`.
2. Move the `hdp_urlinfo.json` file to `/var/www/html/yum/ambari/HDP/`
3. Edit the `hdp_urlinfo.json` file by replacing the URL for *HDP-2.1* and *centos6* with your local HDP repository URL.

   Set the URL in the correct section based on your HDP and Linux OS version and type.

   ```
    {
         "HDP-2.1": {
           "latest": {
              "centos5": "http://public-
   repo-1.hortonworks.com/HDP/centos5/2.x/updates/2.1.3.0/hdp.repo",
              "centos6":
   "http://<yum_repo_server_ip>/yum/ambari/HDP/centos6/2.x/updates/2.1.2.0/hdp.repo",
              "suse11": "http://public-
   repo-1.hortonworks.com/HDP/suse11/2.x/updates/2.1.3.0/hdp.repo"
           }
        },
        "HDP-2.0": {
   ```

```
        "latest": {
           "centos5": "http://public-
repo-1.hortonworks.com/HDP/centos5/2.x/updates/2.0.12.0/hdp.repo",
           "centos6": "http://public-
repo-1.hortonworks.com/HDP/centos6/2.x/updates/2.0.12.0/hdp.repo",
           "suse11": "http://public-
repo-1.hortonworks.com/HDP/suse11/2.x/updates/2.0.12.0/hdp.repo"
        }
     },
     "HDP-1.3": {
        "latest": {
           "centos5": "http://public-
repo-1.hortonworks.com/HDP/centos5/1.x/updates/1.3.8.0/hdp.repo",
           "centos6": "http://public-
repo-1.hortonworks.com/HDP/centos6/1.x/updates/1.3.8.0/hdp.repo",
           "suse11": "http://public-
repo-1.hortonworks.com/HDP/suse11/1.x/updates/1.3.8.0/hdp.repo"
        }
     }
   }
```

Configure the HDP Repository on the Ambari Server

After you configure the Ambari repository on the Ambari server, you must configure the HDP repository on the Ambari server.

Prerequisites

Verify that you have configured the ambari.repository on the Ambari server.

Procedure

1 Edit the following file:

```
/var/lib/ambari-server/resources/stacks/HDP/2.1/repos/repoinfo.xml
```

a Replace the version number *2.1* with your version number.

b Replace the URL of the *latest* element to the location of our `hdp_urlinfo.json`:

```
http://<yum_repo_server_ip>/yum/ambari/HDP/hdp_urlinfo.json
```

c Replace the *baseurl* in the `os type="redhat6"` with your local HDP repository URL, as shown in the following example:

```
<?xml version="1.0"?>
<!--
  License section(not displayed here).
-->
<reposinfo>
 <latest>http://<yum_repo_server_ip>/yum/ambari/HDP/hdp_urlinfo.json</latest>
 <os type="redhat6">
  <repo>
   <baseurl>http://<yum_repo_server_ip>/yum/ambari/HDP/centos6/2.x/updates/2.1.2.0/</baseurl>
   <repoid>HDP-2.1</repoid>
   <reponame>HDP</reponame>
  </repo>
 </os>
 <os type="redhat5">
  <repo>
```

```
   <baseurl>http://public-repo-1.hortonworks.com/HDP/centos5/2.x/updates/2.1.2.0/</baseurl>
   <repoid>HDP-2.1</repoid>
   <reponame>HDP<?reponame>
  </repo>
 </os>
 <os type="suse11">
  <repo>
   <baseurl>http://public-repo-1.hortonworks.come/HDP/suse11/2.x/updates/2.1.2.0/<baseurl>
   <repoid>HDP-2.1</repoid>
   <reponame>HDP</reponame>
  </repo>
 </os>
 <os type="debian 12">
  <repo>
   <baseurl>http://public-repo-1.hortonworks.com/HDP/ubuntu 12/2.x/updates/2.1.2.0/</baseurl>
   <repoid>HDP-2.1</repoid>
   <reponame>HDP</reponame>
  </repo>
 </os>
</reposinfo>
```

2 Start the Ambari server.

```
ambari-server start
```

You are ready to create clusters for the Ambari server by using the local yum repository.

Create a Hadoop Template Virtual Machine using RHEL Server 6.x and VMware Tools

You can create a Hadoop Template virtual machine that has a customized version of the Red Hat Enterprise Linux (RHEL) Server 6.x operating system that includes VMware Tools. Although only a few Hadoop distributions require a custom version of RHEL Server 6.x, you can customize RHEL Server 6.x for any Hadoop distribution.

Before You Create a Hadoop Template Virtual Machine using RHEL Server 6.x and VMware Tools

Before you create a Hadoop template virtual machine using the RHEL server 6.x and VMware tools, you must perform some prerequisite tasks and be familiar with some important information on the RHEL Server 6.1, hostnames, disk partitioning, and creating Hadoop Template virtual machines with multiple cores per socket.

You can create a Hadoop Template virtual machine that uses RHEL Server 6.1 or later as the guest operating system into which you can install VMware Tools for RHEL 6.x in combination with a supported Hadoop distribution. This allows you to create a Hadoop Template virtual machine that uses your organization's operating system configuration. When you provision Big Data clusters using the customized Hadoop template, the VMware Tools for RHEL 6.x will be in the virtual machines that are created from the Hadoop Template virtual machine.

If you create Hadoop Template virtual machines with multiple cores per socket, when you specify the CPU settings for the virtual machine you must specify a multiple of cores per socket. For example, if the virtual machine uses two cores per socket, the vCPU settings must be an even number. For example: 4, 8, or 12. If you specify an odd number, the cluster provisioning or CPU resizing will fail.

IMPORTANT

- You must use **`localhost.localdomain`** as the hostname when you install the RHEL template otherwise the FQDN of the virtual machine cloned from the template may not be set correctly.
- If you are performing disk partitioning, do not use the Linux Volume Manager (LVM).

Prerequisites

- Deploy theBig Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24.
- Obtain the IP address of the Serengeti Management Server.
- Locate the VMware Tools version that corresponds to the ESXi version in your data center.

Create a Virtual Machine Template with a 20GB Thin Provisioned Disk and Install RHEL 6.x

For more information on this procedure, see the *Red Hat Enterprise Linux Installation Guide*, available on the Red Hat website.

Procedure

1. Download the RHEL Server 6.x installation ISO from **`www.redhat.com`** to a datastore.
2. In vSphere Client, create a new virtual machine with a 20GB thin provision disk and select Red Hat Enterprise Linux 6 (64-bit) as the Guest OS.
3. Right-click the virtual machine and click **Edit Settings**.
4. Select **CD/DVD Device 0**, and select the datastore ISO file for the RHEL ISO file.
5. Select **SCSI controller 0** > **Change Type** > **LSI Logic Parallel** and click OK.
6. Under **Device Status**, select **`connected`** and **`connect at power on`**, and click **OK**.
7. From the console window of the virtual machine, install the RHEL Server 6.x operating system using the default settings for all settings except the following items:
   - You can select the language and time zone you want the operating system to use
   - You can specify that the swap partition use a smaller size to save disk space (for example, 500MB)
   - You can reduce the size of the swap partition because it is not used by Big Data Extensions.
   - Select **Minimal** in the Package Installation Defaults screen.

Ensure the Virtual Machine has a Valid IP and Internet Connectivity

The Hadoop template virtual machine requires a valid IP address and an Internet connection.

Prerequisites

-

Procedure

- Run the `ifconfig` command to ensure that the virtual machine has a valid IP and Internet connectivity.

  This task assumes the use of Dynamic Host Configuration Protocol (DHCP).

  - If IP address information appears in the output of the `ifconfig` command , see "Configure the Network for the Hadoop Template Virtual Machine to use DHCP," on page 71.
  - If no IP address information appears, see "Configure the Network for the Hadoop Template Virtual Machine to use DHCP," on page 71.

Configure the Network for the Hadoop Template Virtual Machine to use DHCP

Procedure

1. Using a text editor open the `/etc/sysconfig/network-scripts/ifcfg-eth0` file.
2. Locate the following parameters and specify the following configuration.

   ```
   ONBOOT=yes
   DEVICE=eth0
   BOOTPROTO=dhcp
   ```

3. Save your changes and close the file.
4. Restart the network service.

   ```
   sudo /sbin/service network restart
   ```

5. Run the `ifconfig` command to ensure that the virtual machine has a valid IP and Internet connectivity.

Install the JDK 7 RPM

Procedure

1. From the Oracle® Java SE 7 Downloads page, download the latest JDK 7 Linux x64 RPM and copy it to the root folder of the virtual machine template.
2. Install the RPM.

   ```
   rpm -Uvh jdk-7u80-linux-x64.rpm
   ```

3. Delete the RPM file.

   ```
   rm -f jdk-7u80-linux-x64.rpm
   ```

4. Edit `/etc/environment` and add the following line: `JAVA_HOME=/usr/java/default`

Customize the Virtual Machine

Run the installation scripts to customize the virtual machine.

Procedure

1. Register the RHEL operating system to enable the RHEL yum repositories. This allows the installation script to download packages from the yum repository. See "Registering from the Command Line" in the *Red Hat Enterprise Linux 6 Deployment Guide*, available on the Red Hat website.
2. Download the scripts from `https://deployed_serengeti_server_IP/custos/custos.tar.gz`.

3 Create the directory /tmp/custos, make this your working directory, and run tar xf to uncompress the tar file.

```
mkdir /tmp/custos
cd /tmp/custos
tar xf /tmp/custos/custos.tar.gz
```

4 Run the installer.sh script specifying the /usr/java/default directory path.

```
./installer.sh /usr/java/default
```

You must use the same version of the installer.sh script as your Big Data Extensions deployment.

5 Remove the /etc/udev/rules.d/70-persistent-net.rules file to prevent increasing the eth number during the clone operation.

If you do not remove the file, virtual machines that are cloned from the template cannot get IP addresses. If you power on the Hadoop template virtual machine to make changes, remove the file before you shut down this virtual machine.

Install VMware Tools for RHEL 6.x

Procedure

1 Right-click the RHEL 6 virtual machine in vSphere Client, then select **Guest > Install/Upgrade VMware Tools**.

2 Log in to the virtual machine and mount the CD-ROM to access the VMware Tools installation package.

```
mkdir /mnt/cdrom
mount /dev/cdrom /mnt/cdrom
mkdir /tmp/vmtools
cd /tmp/vmtools
```

3 Run the tar xf command to extract the VMware Tools package tar file.

```
tar xf /mnt/cdrom/VMwareTools-*.tar.gz
```

4 Make vmware-tools-distrib your working directory, and run the vmware-install.pl script.

```
cd vmware-tools-distrib
./vmware-install.pl
```

Press **Enter** to finish the installation.

5 Remove the vmtools temporary (temp) file that is created as an artifact of the installation process.

```
rm -rf /tmp/vmtools
```

6 Shut down virtual machine.

Synchronize the Time on the Hadoop Template Virtual Machine

Synchronize the time on the Hadoop template virtual machine with vCenter Server.

Procedure

1 In the vSphere Web Client, right-click the Hadoop Template virtual machine and select **Edit Settings**.

2 On the **VM Options** tab, click **VMware Tools > Synchronize guest time with host**.

Complete the Process of Creating a Hadoop Template Virtual Machine

To use the customized Hadoop Template virtual machine you replace the original Hadoop Template virtual machine and restart the Tomcat service to enable the custom RHEL virtual machine template.

Procedure

1 On the **Virtual Hardware** tab of the Edit Settings dialog, uncheck the **Connected** checkbox. If the CD/DVD Device is connected to the ISO file, the clone virtual machine process fails.

2 Replace the original Hadoop Template virtual machine with the customized virtual machine that you created.

   a Move the original Hadoop Template virtual machine out of the vApp.

   b Drag the new template virtual machine that you just created into the vApp.

3 Log in to the Serengeti Management Server as the user `serengeti`, and restart the Tomcat service.

```
sudo /sbin/service tomcat restart
```

   Restarting the Tomcat service enables the custom RHEL virtual machine template, making it your Hadoop Template virtual machine.

Maintain a Customized Hadoop Template Virtual Machine

You can modify or update the Hadoop Template virtual machine operating system. When you make updates, you must remove the snapshot that is created by the virtual machine.

If you create a custom Hadoop Template virtual machine that uses a version of RHEL 6.x, or modify the operating system, you must remove the serengeti-snapshot that Big Data Extensions creates. If you do not remove the serengeti-snapshot, changes you made to the Hadoop Template virtual machine will not take effect.

Prerequisites

- Deploy the Big Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24.
- Create a customized Hadoop Template virtual machine using RHEL 6.x. See "Create a Hadoop Template Virtual Machine using RHEL Server 6.x and VMware Tools," on page 69

.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Power on the Hadoop Template virtual machine and apply changes or updates.

3 Remove the `/etc/udev/rules.d/70-persistent-net.rules` file to prevent increasing the `eth` number during the clone operation.

   If you do not remove the file, virtual machines that are cloned from the template cannot get IP addresses. If you power on the Hadoop template virtual machine to make changes, remove the file before you shut down this virtual machine.

4 From the vSphere Web Client, shut down the Hadoop Template virtual machine.

5 Delete the snapshot labeled serengeti-snapshot from the customized Hadoop Template virtual machine.

   a In the vSphere Web Client, right-click the Hadoop Template virtual machine and select **Snapshot > Snapshot Manager**

   b Select the serengeti-snapshot, and click **Delete.**

   The generated snapshot is removed.

6 Synchronize the time on the Hadoop template virtual machine with vCenter Server.

   a In the vSphere Web Client, right-click the Hadoop template virtual machine and select **Edit Settings**.

   b On the **VM Options** tab, click **VMware Tools > Synchronize guest time with host**.

Managing the Big Data Extensions Environment

6

After you install Big Data Extensions, you can stop and start the Serengeti services, create user accounts, manage passwords, update SSL certificates, and log in to cluster nodes to perform troubleshooting.

This chapter includes the following topics:

Add Specific User Names to Connect to the Serengeti Management Server

You can add specific user names with which to login to the Serengeti Management Server. The user names you add are the only users who can connect to the Serengeti Management Server using the Serengeti CLI or the Big Data Extensions user interface for use with vSphere Web Client.

Passwords must be from 8 to 20 characters, use only visible lowerASCII characters (no spaces), and must contain at least one uppercase alphabetic character (A - Z), at least one lowercase alphabetic character (a - z), at least one digit (0 - 9), and at least one of the following special characters: _, @, #, $, %, ^, &, *

Prerequisites

- Deploy the Serengeti vApp.
- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.

Procedure

1 Right-click the Serengeti Management Server virtual machine and select **Open Console**.

The password for the Serengeti Management Server appears.

NOTE If the password scrolls off the console screen, press Ctrl+D to return to the command prompt.

2 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

   Use the IP address that appears in the **Summary** tab and the current password.

3 Edit the `/opt/serengeti/conf/Users.xml` file to add additional user names.

```
vi /opt/serengeti/conf/Users.xml
```

4 Edit the `<user name="*" />` attribute by replacing the asterisk (*) wildcard character with the user name you wish to use. You can add multiple user names by adding a new `<user name="name" />` attribute on its own line. The `User.xml` file supports multiple lines.

```
<user name="jsmith" />
<user name="sjones" />
<user name="jlydon" />
```

5 Restart the Tomcat service.

```
/sbin/service tomcat restart
```

Only the user names you add to the `User.xml` file can be used to login to the Serengeti Management Server using the Serengeti CLI or the Big Data Extensions user interface for use with vSphere Web Client.

Change the Password for the Serengeti Management Server

When you power on the Serengeti Management Server for the first time, it generates a random password that is used for the `root` and `serengeti` users. If you want an easier to remember password, you can use the virtual machine console to change the random password for the `root` and `serengeti` users.

NOTE You can change the password for the virtual machine of any node by using this procedure.

Passwords must be from 8 to 20 characters, use only visible lowerASCII characters (no spaces), and must contain at least one uppercase alphabetic character (A - Z), at least one lowercase alphabetic character (a - z), at least one digit (0 - 9), and at least one of the following special characters: _, @, #, $, %, ^, &, *

Prerequisites

- Deploy the Serengeti vApp.
- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.

Procedure

1 Right-click the Serengeti Management Server virtual machine and select **Open Console**.

   The password for the Serengeti Management Server appears.

   NOTE If the password scrolls off the console screen, press Ctrl+D to return to the command prompt.

2 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

   Use the IP address that appears in the **Summary** tab and the current password.

3 Use the `/opt/serengeti/sbin/set-password` command to change the password for the `root` user and the `serengeti` user.

```
sudo /opt/serengeti/sbin/set-password -u
```

4 Enter a new password, and enter it again to confirm.

The next time you log in to the Serengeti Management Server, use the new password.

What to do next

You can create a new user name and password for the Serengeti Command-Line Interface Client.

Create a User Name and Password for the Serengeti Command-Line Interface

The Serengeti Command-Line Interface Client uses the vCenter Server login credentials with read permissions on the Serengeti Management Server. If you do not create a user name and password for the Serengeti Command-Line Interface Client, it will use the default vCenter Server administrator credentials. However, for security reasons, it's best to create a user account specifically for use with the Serengeti Command-Line Interface Client.

Passwords must be from 8 to 20 characters, use only visible lowerASCII characters (no spaces), and must contain at least one uppercase alphabetic character (A - Z), at least one lowercase alphabetic character (a - z), at least one digit (0 - 9), and at least one of the following special characters: _, @, #, $, %, ^, &, *

Prerequisites

- Deploy the Big Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24.
- Install the Serengeti Command-Line Interface Client. See "Install the Serengeti Remote Command-Line Interface Client," on page 31.

Procedure

1 Open a Web browser and go to: `https://vc-hostname:port/vsphere-client`.

   The *vc-hostname* can be either the DNS host name or IP address of vCenter Server. By default the port is 9443, but this can change during the installation of the vSphere Web Client.

2 Type the user name and password that has administrative privileges on vCenter Server, and click **Login**.

   Note vCenter Server 5.5 users must use a local domain to perform SSO related operations.

3 From the vSphere Web Client Navigator panel, select **Administration**, **SSO Users and Groups**.

4 Change the login credentials.

The login credentials are updated. The next time you access the Serengeti Command-Line Interface use the new login credentials.

What to do next

You can change the password of the Serengeti Management Server. See "Change the Password for the Serengeti Management Server," on page 76.

Specify a Group of Users in Active Directory or LDAP to Use a Hadoop Cluster

You can specify an Active Directory or LDAP server for user authentication. This lets you manage users from a central point.

By defaultBig Data Extensions is installed with authentication only for local user accounts. If you want to use LDAP or Active Directory to authenticate users, you must configure Big Data Extensions for use with your LDAP or Active Directory service.

Big Data Extensions lets you authenticate local users, those managed by LDAP or Active Directory server, or a combination of these authentication methods.

Prerequisites

- Deploy the Big Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24.
- Use the Serengeti Management Server Administration Portal to enable SSO and update the certificate. See "Configure vCenter Single Sign-On Settings for the Serengeti Management Server," on page 30.

Procedure

1. Use the vSphere Web Clientto log in to vCenter Server.
2. Select Big Data Extensions and click the **Manage** tab.
3. Select User Mode and click **Edit**.

   The Configure User dialog box appears.
4. Choose the user authentication mode you wish to use for your Big Data Extensions environment.

   Table 6-1. User Authentication Modes

User Mode	Description
Local	Select **Local** to create and manage users and groups that are stored locally in your Big Data Extensions environment. Local is the default user management solution.
LDAP user	Select **LDAP user** to create and manage users and groups that are stored in your organization's identity source, such as Active Directory or LDAP. If you choose LDAP user you must configure Big Data Extensions to use an LDAP or Active Directory service.
Mixed mode	Select **Mixed mode** to use a combination of both local users and users stored in an external identity source. If you choose mixed mode you must configure Big Data Extensions to use AD as LDAP mode.

5. If you choose to use LDAP or Mixed mode, configure Big Data Extensions to use an LDAP or Active Directory service.

   Table 6-2. LDAP Connection Information

Base user DN	Specify the base user DN.
Base group DN	Specify the base group DN.
Primary server URL	Specify the primary server URL of your Active Directory or LDAP server.
Secondary server URL	Specify the secondary server URL of your Active Directory or LDAP server.
Username	Type the username of the Active Directory or LDAP server administrator account.
Password	Type the password of the Active Directory or LDAP server administrator account.

6. (Optional) Click **Test** to verify that user accounts are found.

Stop and Start Serengeti Services

You can stop and start Serengeti services to make a reconfiguration take effect, or to recover from an operational anomaly.

Procedure

1. Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.
2. Run the `serengeti-stop-services.sh` script to stop the Serengeti services.

   ```
   serengeti-stop-services.sh
   ```

3 Run the `serengeti-start-services.sh` script to start the Serengeti services.

```
serengeti-start-services.sh
```

Ports Used for Communication between Big Data Extensions and the vCenter Server

Big Data Extensions queries information from the vCenter Server, and uses the vCenter Server Single Sign-On service.

Big Data Extensions Management Server

The table below shows the published port for the management server.

Service	Port	Comments
Serengeti Rest API	8080, 8443	Open for Serengeti client and for BDE plugin registration called by VC
SSHD	22	Open for Serengeti client connection

Hadoop Ports

Serengeti deploys Hadoop and Hbase clusters using all default ports. The following lists all ports that are used by the Hadoop or HBase service, the production network.

	Daemon	Default Port
HDFS	Namenode Webpage	50070
	Namenode RPC	8020
	Datanode	50075 50010 50020
MapReduce	JobTracker Webpage	50030
	JobTracker RPC	8021
	TaskTracker	50060
Yarn	Resource Manager Webpage	8088
	Resource Manager RPC	8030, 8031, 8032, 8033
	Node Manager	8040, 8042
Hive	N/A	1000

Hbase Ports

The table below shows the ports used by HBase clusters, along with the default port numbers.

Service	Property Name	Port
ZooKeeper	hbase.zookeeper.property.clientPort	2181
Master	hbase.master.port	60000
Master	hbase.master.info.port	60010
Region server	hbase.regionserver.port	60020
Region server	hbase.regionserver.info.port	60030

Service	Property Name	Port
REST server	hbase.rest.port	8080
REST server	hbase.rest.info.port	8085
Thrift server	hbase.thrift.port	9090
Thrift server	hbase.thrift.info.port	9095

MapR Ports

The table below defines the ports used by a MapR cluster, along with the default port numbers.

Service	Port
CLDB	7222
CLDB JMX monitor port	7220
CLDB web port	7221
HBase Master	60000
HBase Master (for GUI)	60010
HBase RegionServer	60020
Hive Metastore	9083
JobTracker Webpage	50030
JobTracker RPC	8021
MFS server	5660
MySQL	3306
NFS	2049
NFS monitor (for HA)	9997
NFS management	9998
Port mapper	111
TaskTracker	50060
Web UI HTTPS	8443
ZooKeeper	5181

Verify the Operational Status of the Big Data Extensions Environment

To successfully provision a Hadoop cluster, your Big Data Extensions environment must meet certain criteria. You can verify that your environment meets these criteria prior to creating Hadoop clusters, as well as troubleshoot cluster creation issues you may encounter.

Operational Status of Big Data Extensions Services

Big Data Extensions consists of several services that you can verify are running.

Big Data Extensions consists of the following services: Tomcat service, Yum server, Chef server, and PostgreSQL server. You can verify that these services are running prior to creating Hadoop clusters.

Prerequisites

- Deploy the Serengeti vApp.

- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

2 Verify that the Tomcat service is running.

   a Run the command `pgrep -f org.apache.catalina.startup.Bootstrap -l`.

   ```
   pgrep -f org.apache.catalina.startup.Bootstrap -l
   ```

   b Run the command `wget https://bde_server_ip:8443 --no-check-certificate`

   ```
   wget https://bde_server_ip:8443 --no-check-certificate
   ```

3 Verify that the Yum server is running.

   Run the command `/sbin/service httpd status`.

   ```
   /sbin/service httpd status
   ```

   If the Yum server is operating properly it responds with the status message `running`.

4 Verify that the Chef server is running.

   Run the command `sudo /usr/bin/chef-server-ctl status`. The `status` subcommand displays the status of all services available to the Chef server.

   ```
   sudo /usr/bin/chef-server-ctl status
   ```

5 Verify that the PostgreSQL server is running.

   a Run the command `pgrep -f /opt/chef-server/embedded/bin/postgres -l` to verify that the postgres process is running. The `-l` option lists the available databases.

   ```
   pgrep -f /opt/chef-server/embedded/bin/postgres -l
   ```

   b Run the command `echo "\dt" | psql -U serengeti` to display the database tables created for Big Data Extensions. The `-dt` option specifies the name of the database to connect to, and turns off the display of the database column names in the resulting output. The `-U` option specifies the username with which to connect to the database.

   ```
   echo "\dt" | psql -U serengeti
   ```

   If the databases available to PostgreSQL and the tables owned by the **serengeti** user display, your PostgreSQL server is running as expected.

What to do next

If any of the above services is not running, you can view the initialization status of the Serengeti Management Server services, view error messages to help troubleshoot problems, and recover services that may not have successfully started using the Serengeti Management Server Administration Portal. See "View Serengeti Management Server Initialization Status," on page 124.

Verify Network Connectivity with vCenter Server

You can verify if your Big Data Extensions deployment can connect to vCenter Server, and identify possible causes that may be preventing a successful network connection.

Prerequisites

- Deploy the Serengeti vApp.

- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

2 Run the command `wget https://vcenter_server_ip:9443 --no-check-certificate`.

```
wget https://vcenter_server_ip:9443 --no-check-certificate
```

If this command retrieves the `index.html` file whose title is *vSphere Web Client*, vCenter Server is running, and there is connectivity between Big Data Extensions and vCenter Server.

If running this command fails to retrieve the `index.html` file, see Step. 3.

3 If the command returns the error message `Connecting to vcenter_server_ip:vcenter_server_port... failed: Connection refused`, the vCenter Server IP address you specified is reachable, but the vCenter Server network port number is incorrect.

4 If the vCenter Server IP address and port number are correct, check your Big Data Extensions deployment's network configuration and ensure that it is properly configured. For example, is Big Data Extensions using a valid IP address and gateway?

What to do next

If you are unable to verify a network connection between Big Data Extensions and vCenter Server, and cannot identify the cause of the problem, the troubleshooting topics provide solutions to problems you might encounter when using Big Data Extensions. See Chapter 13, “Troubleshooting,” on page 139

Verify vCenter Server User Authentication

You can verify if your vCenter Server user authentication is working properly, and identify possible causes that may be preventing a successful cluster creation.

Prerequisites

- Deploy the Serengeti vApp.
- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as the user `serengeti`.

2 Type `serengeti` to start the Serengeti Command-Line Interface.

3 Run the command `connect -host localhost:8443`, and at the prompt type your username and password, which might be different from your login credentials for the Serengeti Management Server. If you can log into Big Data Extensions your vCenter Server user authentication is working correctly.

What to do next

Before creating new virtual machines on hosts, the time on the target hosts is checked against the time on the Serengeti Management Server. If the time between the Serengeti Management Server and the hosts is not synchronized, the virtual machine creation will fail. See “Check Time Synchronization Between Serengeti Management Server and Hosts,” on page 83.

Check Time Synchronization Between Serengeti Management Server and Hosts

When you run the `cluster create` or `cluster create ... —-resume` command, the command can fail if there are time discrepancies in the environment. You can verify that the time is within allowable tolerances and synchronize the time between the Serengeti Management Server and the other hosts within your environment.

Before creating new virtual machines on hosts, the time on the target hosts is checked against the time on the Serengeti Management Server. If the time between the Serengeti Management Server and the hosts is not synchronized, the cluster creation might fail.

Prerequisites

- Deploy the Serengeti vApp.
- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.

Procedure

1. Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.
2. Run the command `date +%T` to see the time on the Serengeti Management Server.

   ```
   date +%T
   ```

3. From the vSphere Web Client, record the time of each host in the datacenter.
4. Compare the date and time from the Serengeti Management Server and each host to see if they are greater than the Maximum-Threshold. If there is HBase service in the cluster, the Maximum-Threshold is 20 seconds. Otherwise, the Maximum-Threshold is 4 minutes.

   If the times between hosts are not synchronized, login to each host and view the `/etc/ntp.conf` file to verify if the NTP configuration is correct.

5. From the vSphere Web Client, configure all ESXi hosts to synchronize their clocks with the same NTP server.

What to do next

After you synchronize the time between the Serengeti Management Server and the other ESXi hosts within your environment, try to create a cluster.

Verify Network Connectivity Between Compute Nodes and Isilon HDFS

If you are using EMC Isilon OneFS for your HDFS, you can verify the network connectivity from the compute nodes to the Isilon OneFS filesystem.

Procedure

1. Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.
2. For each compute node (TaskTracker or NodeManager), login and run the command `hadoop dfsadmin -report` to verify that the HDFS is running correctly. If the command returns the `Configured Capacity` and `Present Capacity`, the worker node can successfully access the HDFS.

   If the HDFS does not respond, see Step 3.

3 Ensure that the HDFS IP address and network port number is correct. Login to the Isilon Namenode (which may require a different username and password) and verify that the HDFS service is listening on port 8020.

   If the HDFS is listening on the correct network port, see Step 4.

4 Check the `fs.defaultFS` entry in the Hadoop configuration file `core-site.xml`. Ensure that the IP address, FQDN, and network port are configured to use the correct HDFS service.

Check Which Users and User Groups Exist in the Isilon OneFS

If you use EMC Isilon OneFS as the external HDFS cluster, you must create and configure users and user groups and prepare your Isilon OneFS environment. You can verify that you have created the correct users and user groups, and check which users and groups exist in your Isilon OneFS environment.

Prerequisites

Prepare the Isilon OneFS for use as an external HDFS cluster. See "Prepare the EMC Isilon OneFS as the External HDFS Cluster," on page 104.

Procedure

1 Open a command shell, such as Bash or PuTTY, and SSH to the Isilon OneFS node.

2 Run the command `isi auth users/groups list` to list the existing Isilon OneFS users and user groups.

3 Run the command `ls -al HDFS_ROOT_DIR` to verify which users and user groups are using the HDFS.

   When running the `ls` command in the Isilon filesystem, the `-al` option must come before the `HDFS_ROOT_DIR` directory name. Otherwise, the `-al` option will be regarded as a directory name by the `ls` command.

```
ls -al HDFS_ROOT_DIR
```

NOTE In the HDFS subdirectory there may be files and directories with permissions and ownership assigned to users or groups other than those using Big Data Extensions.

Check Storage Capacity

To successfully deploy a cluster you must have enough storage capacity in your Big Data Extensions environment.

The datastores you add to your Big Data Extensions environment are made available to the clusters you create within Big Data Extensions. If you do not add enough storage capacity cluster creation will fail.

In addition to overall storage capacity, you must ensure that you have enough Shared and Local storage. Shared stroage is recommended for master nodes, and enables you to use vMotion, HA, and Fault Tolerance. Local storage is recommended for worker nodes.

Prerequisites

You must have added a datastore to your Big Data Extensions environment. See "Add a Datastore in the vSphere Web Client," on page 92

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

2 Run the command `datastore list --detail` to see which vCenter Server datastores are in use by Big Data Extensions.

3 Using the configuration values specified in the cluster specification file, calculate how much storage capacity the cluster will require.

4 Use the vSphere Web Client to log in to vCenter Server, and verify that the datastores you identified as belonging to Big Data Extensions have enough storage capacity for the clusters you want to create. Additionally, ensure that the datastores are in an active state.

What to do next

If your Big Data Extensions environment does not have adequate storage capacity to create clusters, add additional datastores. See "Add a Datastore in the vSphere Web Client," on page 92.

Verify the Ambari Application Manager Installation

If you use Apache Ambari to manage your Hadoop cluster, you can verify that the Ambari service is running, has a network connection, and valid user credentials with which to connect to your cluster.

Prerequisites

- Deploy the Big Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24
- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.
- Add the Ambari application manager to your Big Data Extensions environment. See "Add an Application Manager by Using the vSphere Web Client," on page 45.

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as the user `serengeti`.

2 Run the `curl` command with the `-u` option to specify the username and password in use by the Ambari service, and the `-G` option to specify the URL of the Ambari system check service: `http://`*`ambari_server_ip`*`:8080/api/v1/check`

```
curl -u username:password -G http://ambari_server_ip:8080/api/v1/check
```

- If the system returns `RUNNING`, the Ambari server is running. If you receive a system message indicating that your Ambari service is not running, investigate the issue and confirm that you can successfully start Ambari before proceeding.
- If the system returns `Bad credentials`, the username and password are incorrect. Obtain the correct username and password for your Ambari installation.
- If the `curl` command hangs for 30 or more seconds, and the system returns the error message `curl: (7) Failed to connect to `*`ambari_server_ip`*` port `*`port_number`*`: Connection refused`, the IP/FQDN or port number is incorrect. Obtain the correct network address for your Ambari installation.

  This error message may also indicate that the Ambari server virtual machine in powered off. Verify that the Ambari virtual machine is powered on, and that the Ambari server is running.

What to do next

If your Ambari installation is not responding, confirm that it is properly installed and configured. See "Modify an Application Manager by Using the Web Client," on page 46.

Verify Cloudera Manager Installation

If you use Cloudera Manager to manage your Hadoop cluster, you can verify that Cloudera Manager is running, has a network connection, and valid user credentials with which to connect to your cluster.

Prerequisites

- Deploy the Big Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24
- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.
- Add the Cloudera Manager application to your Big Data Extensions environment. See "Add an Application Manager by Using the vSphere Web Client," on page 45.

Procedure

1. Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as the user `serengeti`.
2. Run the `curl` command with the `-u` option to specify the username and password in use by Cloudera Manager, and the `-G` option to specify the URL of the Cloudera ManagerAPI version number : `http://`*`cloudera_manager_server_ip`*`:7180/api/version`

   ```
   curl -u username:password -G http://cloudera_manager_server_ip:7180/api/version
   ```

   Record the API version number returned by Cloudera Manager.
3. Run the `curl` command with the `-u` option to specify the username and password in use by Cloudera Manager, and the `-G` option to specify the URL of the Cloudera Manager `/tools/echo` query: `http://`*`cloudera_manager_server_ip`*`:7180/api/`*`cloudera_manager_api_version`*`/tools/echo`

   ```
   curl -u username:password -G http://cloudera_manager_server_ip:
   7180/api/cloudera_manager_api_version/tools/echo
   ```

   This example specifies a Cloudera Manager installation using the username and password **`cloudera`**, whose network address is **`192.168.1.1`** using API version **`v5`**.

   ```
   curl -u cloudera:cloudera -G http://192.168.1.1:7180/api/v5/tools/echo
   ```

   - If the system returns `Hello world!`, Cloudera Manager is running. If you receive a system message indicating that your Cloudera Manager is not running, investigate the issue and confirm that you can successfully start Cloudera Manager before proceeding.
   - If the system returns `Error 401 Bad credentials`, the username and password are incorrect. Obtain the correct username and password for your Cloudera Manager installation.
   - If the system returns the error message `curl: (7) Failed to connect to `*`cloudera_manager_server_ip`*` port 7180: No route to host`, the IP address or FQDN is incorrect. Obtain the correct network address for your Cloudera Manager installation.

     This error message may also indicate that the Cloudera Manager virtual machine in powered off. Verify that the Cloudera Manager virtual machine is powered on, and that Cloudera Manager is running.

What to do next

If your Cloudera Manager installation is not responding, confirm that it is properly installed and configured. See "Modify an Application Manager by Using the Web Client," on page 46.

Check DNS Forward and Reverse Lookup

Big Data Extensions requires a properly configured network environment. You can verify that you have a properly configured forward and reverse address lookup for you DNS.

Reverse DNS lookup determines the hostname associated with a given IP address. Forward DNS lookup determines the determines the IP address associated with a given hostname.

Prerequisites

- Deploy the Big Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24
- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.

Procedure

1. Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as the user `serengeti`.
2. Run the `echo` command to retrieve the IP addresses in use by the cluster.

   ```
   echo ipv4_address_from_network_interface | psql
   ```

   Record the IP addresses for each network interface card in use by the cluster.
3. For each IP address you recorded in the previous step, run the `host` command to verify that DNS reverse lookup returns the fully qualified domain name (FQDN). If the system responds with a FQDN for each IP address, DNS reverse lookup is working.

   ```
   host IP_address
   ```

   Record the FQDN for each network address you check.
4. For each FQDN you recorded in the previous step, run the `host` command to verify that DNS forward lookup returns the IP address associated with th FQDN. If the system responds with an IP address for each FQDN, DNS forward lookup is working.
5. (Optional) If you are unable to resolve the IP addresses and FQDNs, open the file `/etc/resolv.conf`, and confirm that a DNS name server has been configured for use with your environment.
   - If there is no name server configured for use with your environment, ask you administrator for the correct DNS server name to use.
   - If a name server is configured, but your DNS does not provide forward or reverse lookup, investigate the cause and configure your DNS as required. Possible causes preventing your DNS from functioning correctly may include:
     - The name server cannot be reached due to an incorrect IP address.
     - The DNS service on that virtual machine may be shutdown, or unresponsive.
     - The virtual machine containing the DNS service may be shutdown.

What to do next

If your DNS is not functioning as expected, investigate the cause and make the necessary configuration or operational changes until you are able to verify that you have a properly configured forward and reverse address lookup for you DNS. See "Modify the DNS Type in the vSphere Web Client," on page 94.

Verify the Network Connection Between Big Data Extensions and the Cluster Nodes

The Serengeti Management Server must be able to connect to each of the nodes in a Hadoop cluster. You can verify that the Serengeti Management Server can contact each cluster node.

Prerequisites

- Deploy the Big Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24
- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.
- Add a network for use by Big Data Extensions. See "Add a Network in the vSphere Web Client," on page 93.

Procedure

1. Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as the user `serengeti`.
2. Run the `echo` command to retrieve the IP addresses in use by the cluster.

   ```
   echo "select ipv4_address_from_network_interface" | psql
   ```

   Record the IP addresses for each network interface card in use by the cluster.
3. Run the `ping` command to contact each IP address and verify that the Serengeti Management Server can contact each of the cluster nodes.

What to do next

If you are unable to establish a connection between the Serengeti Management Server and the Hadoop cluster nodes, investigate the cause and make the necessary changes until you are able to verify that you have a properly configured network.

Verify the Local Yum Repository

If you created a local yum repository from which to deploy your Hadoop distributions, you can verify that the repository is working properly.

Prerequisites

- Deploy the Big Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24
- Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.
- You created a local Yum repository from which to deploy your Hadoop distributions. See "Configuring Yum and Yum Repositories," on page 52.

Procedure

1. Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as the user **serengeti**.
2. Run the command `wget` *`local_repository_url`* to download the local repository Web page.
3. You can open and view the local repository Web page with a Web browser inside your network to verify that the local repository works.

What to do next

You can successfully create Hadoop clusters in your Big Data Extensions environment. See Chapter 8, "Creating Hadoop and HBase Clusters," on page 97

Managing vSphere Resources for Clusters 7

Big Data Extensions lets you manage the resource pools, datastores, and networks that you use in the clusters that you create.

This chapter includes the following topics:

Add a Resource Pool with the Serengeti Command-Line Interface

You add resource pools to make them available for use by Hadoop clusters. Resource pools must be located at the top level of a cluster. Nested resource pools are not supported.

When you add a resource pool to Big Data Extensions it symbolically represents the actual vSphere resource pool as recognized by vCenter Server. This symbolic representation lets you use the Big Data Extensions resource pool name, instead of the full path of the resource pool in vCenter Server, in cluster specification files.

Note After you add a resource pool to Big Data Extensions, do not rename the resource pool in vSphere. If you rename it, you cannot perform Serengeti operations on clusters that use that resource pool.

Procedure

1 Access the Serengeti Command-Line Interface client.

2 Run the `resourcepool add` command.

   The `--vcrp` parameter is optional.

   This example adds a Serengeti resource pool named `myRP` to the vSphere rp1 resource pool that is contained by the `cluster1` vSphere cluster.

```
resourcepool add --name myRP --vccluster cluster1 --vcrp rp1
```

Remove a Resource Pool with the Serengeti Command-Line Interface

You can remove resource pools from Serengeti that are not in use by a Hadoop cluster. You remove resource pools when you do not need them or if you want the Hadoop clusters you create in the Serengeti Management Server to be deployed under a different resource pool. Removing a resource pool removes its reference in vSphere. The resource pool is not deleted.

Procedure

1 Access the Serengeti Command-Line Interface client.

2 Run the `resourcepool delete` command.

If the command fails because the resource pool is referenced by a Hadoop cluster, you can use the `resourcepool list` command to see which cluster is referencing the resource pool.

This example deletes the resource pool named `myRP`.

```
resourcepool delete --name myRP
```

Add a Datastore in the vSphere Web Client

You can add datastores to Big Data Extensions to make them available to big data clusters.
Big Data Extensions supports both shared datastores and local datastores.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 From the Inventory Lists, select **Resources**.

4 Expand the **Inventory Lists**, and select **Datastores**.

5 Click the **Add** (+) icon.

6 In the **Name** text box, type a name with which to identify the datastore in Big Data Extensions.

Passwords must be from 8 to 20 characters, use only visible lowerASCII characters (no spaces), and must contain at least one uppercase alphabetic character (A - Z), at least one lowercase alphabetic character (a - z), at least one digit (0 - 9), and at least one of the following special characters: _, @, #, $, %, ^, &, *

7 From the **Type** list, select the datastore type in vSphere.

Type	Description
Shared	Recommended for master nodes. Enables you to leverage vMotion, HA, and Fault Tolerance. NOTE If you do not specify shared storage and try to provision a cluster using vMotion, HA, or Fault Tolerance, the provisioning fails.
Local	Recommended for worker nodes. Throughput is scalable and the cost of storage is lower.

8 Select one or more vSphere datastores to make available to the Big Data Extensions datastore that you are adding.

9 Click **OK** to save your changes.

The vSphere datastores are available for use by big data clusters deployed within Big Data Extensions.

Remove a Datastore in the vSphere Web Client

You remove a datastore from Big Data Extensions when you no longer want the Hadoop clusters you create to use that datastore.

Prerequisites

Remove all Hadoop clusters associated with the datastore. See "Delete a Cluster in the vSphere Web Client," on page 110.

Procedure

1. Use the vSphere Web Client to log in to vCenter Server.
2. Select **Big Data Extensions**.
3. From the Inventory Lists, select **Resources**.
4. Expand **Resources**, select **Inventory Lists**, and select **Datastores**.
5. Select the datastore that you want to remove, right-click, and select **Remove**.
6. Click **Yes** to confirm.

   If you did not remove the cluster that uses the datastore, you receive an error message indicating that the datastore cannot be removed because it is currently in use.

The datastore is removed from Big Data Extensions.

Add a Network in the vSphere Web Client

You add networks to Big Data Extensions to make the IP addresses contained by those networks available to big data clusters.

Procedure

1. Use the vSphere Web Client to log in to vCenter Server.
2. Select **Big Data Extensions**.
3. From the Inventory Lists, select **Resources**.
4. Expand **Resources**, click **Inventory Lists > Inventory Lists** and select **Networks**.
5. Click the **Add** (+) icon.
6. In the **Name** text box, type a name with which to identify the network resource in Big Data Extensions.
7. From the **Port group name** list, select the vSphere port group that you want to add to Big Data Extensions.

8 Select a DNS type.

Option	Description
Normal	The DNS server provides both forward and reverse FQDN to IP resolution. Reverse DNS is IP address to domain name mapping. The opposite of forward (normal) DNS which maps domain names to IP addresses. Normal is the default DNS type.
Dynamic	Dynamic DNS (DDNS or DynDNS) is a method of automatically updating a name server in the Domain Name System (DNS) with the active DNS configuration of its configured hostnames, addresses or other information. Big Data Extensions integrates with a Dynamic DNS server in its network through which it provides meaningful host names to the nodes in a Hadoop cluster. . The cluster will then automatically register with the DNS server.
Others	There is no DNS server in the VLAN, or the DNS server doesn't provide normal DNS resolution or Dynamic DNS services. In this case, you must add FQDN/IP mapping for all nodes in the `/etc/hosts` file for each node in the cluster. Through this mapping of hostnames to IP addresses each node can contact another node in the cluster.

9 Choose the type of addressing to use for the network: **Use DHCP to obtain IP addresses** or **Use static IP addresses**.

10 (Optional) If you chose **Use static IP addresses** in Step 9, enter one or more IP address ranges.

11 Click **OK** to save your changes.

The IP addresses of the network are available to big data clusters that you create within Big Data Extensions.

Modify the DNS Type in the vSphere Web Client

DHCP selects the IP address for the IP pool randomly. The FQDN and the IP address of the nodes in a cluster are random. The Hadoop user or application cannot identify where the master nodes are unless they do a query to Big Data Extensions. Even if the user knows the original address, the address might change when the cluster is restarted. Therefore, it is difficult for the Hadoop user or application to access the cluster.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 From the Inventory Lists, select **Resources**.

4 Expand **Resources**, select **Inventory Lists > Networks**.

5 Select a single network to modify, right-click, and select **Modify DNS Type**.

6 Select a DNS type.

Option	Description
Normal	The DNS server provides both forward and reverse FQDN to IP resolution. Reverse DNS is IP address to domain name mapping. The opposite of forward (normal) DNS which maps domain names to IP addresses. Normal is the default DNS type.
Dynamic	Dynamic DNS (DDNS or DynDNS) is a method of automatically updating a name server in the Domain Name System (DNS) with the active DNS configuration of its configured hostnames, addresses or other information. Big Data Extensions integrates with a Dynamic DNS server in its network through which it provides meaningful host names to the nodes in a Hadoop cluster. . The cluster will then automatically register with the DNS server.
Others	There is no DNS server in the VLAN, or the DNS server doesn't provide normal DNS resolution or Dynamic DNS services. In this case, you must add FQDN/IP mapping for all nodes in the `/etc/hosts` file for each node in the cluster. Through this mapping of hostnames to IP addresses each node can contact another node in the cluster.

7 Click **OK** to save your changes.

Reconfigure a Static IP Network in the vSphere Web Client

You can reconfigure a Big Data Extensions static IP network by adding IP address segments to it. You might need to add IP address segments so that there is enough capacity for a cluster that you want to create.

Prerequisites

If your network uses static IP addresses, be sure that the addresses are not occupied before you add the network.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 From the Inventory Lists, select **Resources**.

4 Expand **Resources**, select **Inventory Lists > Networks**.

5 Select the static IP network to reconfigure, right-click, and select **Add IP Range**.

6 Click **Add IP range**, and enter the IP address information.

7 Click **OK** to save your changes.

IP address segments are added to the network.

Remove a Network in the vSphere Web Client

You can remove an existing network from Big Data Extensions when you no longer need it. Removing an unused network frees the IP addresses for use by other services.

Prerequisites

Remove clusters assigned to the network. See "Delete a Cluster in the vSphere Web Client," on page 110.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 From the Inventory Lists, click **Resources**.

4 Expand **Resources**, select **Inventory Lists > Networks**.

5 Select the network to remove, right-click, and select **Remove**.

6 Click **Yes** to confirm.

  If you have not removed the cluster that uses the network, you receive an error message indicating that the network cannot be removed because it is currently in use.

The network is removed, and the IP addresses are available for use.

Creating Hadoop and HBase Clusters 8

Big Data Extensions you can create and deploy Hadoop and HBase clusters. A big data cluster is a type of computational cluster designed for storing and analyzing large amounts of unstructured data in a distributed computing environment.

Restrictions

- When you create an HBase only cluster, you must use the default application manager because the other application managers do not support HBase only clusters.
- You cannot rename a cluster that was created with Cloudera Manager or Ambari application manager.
- Temporarily powering off hosts will cause Big Data clusters to fail during cluster creation.

  When creating Big Data clusters, Big Data Extensions calculates virtual machine placement according to available resources, Hadoop best practices, and user defined placement policies prior to creating the virtual machines. When performing placement calculations, if some hosts are powered off or set to stand-by, either manually, or automatically by VMware Distributed Power Management (VMware DPM), those hosts will not be considered as available resources when Big Data Extensions calculates virtual machine placement for use with a Big Data cluster.

  If a host is powered off or set to stand-by after Big Data Extensions calculates virtual machine placement, but before it creates the virtual machines, the cluster fails to create until you power on those hosts. The following workarounds can help you both prevent and recover from this issue.

  - Disable VMware DPM on those vSphere clusters where you deploy and run Big Data Extensions.
  - Put hosts in maintenance mode before you power them off.
  - If a Big Data cluster fails to create due to its assigned hosts being temporarily unavailable, resume the cluster creation after you power-on the hosts.

Requirements

The resource requirements are different for clusters created with the Serengeti Command-Line Interface and the Big Data Extensions plug-in for the vSphere Web Client because the clusters use different default templates. The default clusters created by using the Serengeti CLI are targeted for Project Serengeti users and proof-of-concept applications, and are smaller than the Big Data Extensions plug-in templates, which are targeted for larger deployments for commercial use.

Some deployment configurations require more resources than other configurations. For example, if you create a Greenplum HD 1.2 cluster, you cannot use the small size virtual machine. If you create a default MapR or Greenplum HD cluster by using the Serengeti CLI, at least 550 GB of storage and 55 GB of memory are recommended. For other Hadoop distributions, at least 350 GB of storage and 35 GB of memory are recommended.

⚠ **CAUTION** When you create a cluster with Big Data Extensions, Big Data Extensions disables the virtual machine automatic migration on the cluster. Although this prevents vSphere from automatically migrating the virtual machines, it does not prevent you from inadvertently migrating cluster nodes to other hosts by using the vCenter Server user interface. Do not use the vCenter Server user interface to migrate clusters. Performing such management functions outside of the Big Data Extensions environment can make it impossible for you to perform some Big Data Extensions operations, such as disk failure recovery.

Passwords must be from 8 to 20 characters, use only visible lowerASCII characters (no spaces), and must contain at least one uppercase alphabetic character (A - Z), at least one lowercase alphabetic character (a - z), at least one digit (0 - 9), and at least one of the following special characters: _, @, #, $, %, ^, &, *

This chapter includes the following topics:

About Hadoop and HBase Cluster Deployment Types

With Big Data Extensions, you can create and use several types of big data clusters.

Basic Hadoop Cluster

Simple Hadoop deployment for proof of concept projects and other small-scale data processing tasks. The Basic Hadoop cluster contains HDFS and the MapReduce framework. The MapReduce framework processes problems in parallel across huge datasets in the HDFS.

HBase Cluster

Runs on top of HDFS and provides a fault-tolerant way of storing large quantities of sparse data.

Data and Compute Separation Cluster

Separates the data and compute nodes, or clusters that contain compute nodes only. In this type of cluster, the data node and compute node are not on the same virtual machine.

Compute Only Cluster

You can create a cluster that contain only compute nodes, for example Jobtracker, Tasktracker, ResourceManager and NodeManager nodes, but not Namenode and Datanodes. A compute only cluster is used to run MapReduce jobs on an external HDFS cluster.

Compute Workers Only Cluster

Contains only compute worker nodes, for example, Tasktracker and NodeManager nodes, but not Namenodes and Datanodes. A compute workers only cluster is used to add more compute worker nodes to an existing Hadoop cluster.

HBase Only Cluster

Contains HBase Master, HBase RegionServer, and Zookeeper nodes, but not Namenodes or Datanodes. Multiple HBase only clusters can use the same external HDFS cluster.

Customized Cluster

Uses a cluster specification file to create clusters using the same configuration as your previously created clusters. You can edit the cluster specification file to customize the cluster configuration.

Hadoop Distributions Supporting MapReduce v1 and MapReduce v2 (YARN)

If you use either Cloudera CDH4 or CDH5 Hadoop distributions, which support both MapReduce v1 and MapReduce v2 (YARN), the default Hadoop cluster configurations are different. The default hadoop cluster configuration for CDH4 is a MapReduce v1 cluster. The default hadoop cluster configuration for CDH5 is a MapReduce v2 cluster. All other distributions support either MapReduce v1 or MapReduce v2 (YARN), but not both.

About Cluster Topology

You can improve workload balance across your cluster nodes, and improve performance and throughput, by specifying how Hadoop virtual machines are placed using topology awareness. For example, you can have separate data and compute nodes, and improve performance and throughput by placing the nodes on the same set of physical hosts.

To get maximum performance out of your big data cluster, configure your cluster so that it has awareness of the topology of your environment's host and network information. Hadoop performs better when it uses within-rack transfers, where more bandwidth is available, to off-rack transfers when assigning MapReduce tasks to nodes. HDFS can place replicas more intelligently to trade off performance and resilience. For example, if you have separate data and compute nodes, you can improve performance and throughput by placing the nodes on the same set of physical hosts.

CAUTION When you create a cluster with Big Data Extensions, Big Data Extensions disables the virtual machine automatic migration of the cluster. Although this prevents vSphere from migrating the virtual machines, it does not prevent you from inadvertently migrating cluster nodes to other hosts by using the vCenter Server user interface. Do not use the vCenter Server user interface to migrate clusters. Performing such management functions outside of the Big Data Extensions environment might break the placement policy of the cluster, such as the number of instances per host and the group associations. Even if you do not specify a placement policy, using vCenter Server to migrate clusters can break the default ROUNDROBIN placement policy constraints.

You can specify the following topology awareness configurations.

Hadoop Virtualization Extensions (HVE)

Enhanced cluster reliability and performance provided by refined Hadoop replica placement, task scheduling, and balancer policies. Hadoop clusters implemented on a virtualized infrastructure have full awareness of the topology on which they are running when using HVE.

To use HVE, your Hadoop distribution must support HVE and you must create and upload a topology rack-hosts mapping file.

RACK_AS_RACK

Standard topology for Apache Hadoop distributions. Only rack and host information are exposed to Hadoop. To use RACK_AS_RACK, create and upload a server topology file.

HOST_AS_RACK

Simplified topology for Apache Hadoop distributions. To avoid placing all HDFS data block replicas on the same physical host, each physical host is treated as a rack. Because data block replicas are never placed on a rack, this avoids the worst case scenario of a single host failure causing the complete loss of any data block.

Use HOST_AS_RACK if your cluster uses a single rack, or if you do not have rack information with which to decide about topology configuration options.

None

No topology is specified.

About HBase Database Access

Serengeti supports several methods of HBase database access.

- Log in to the client node virtual machine and run `hbase shell` commands.
- Log in to the client node virtual machine and run HBase jobs by using the `hbase` command.

```
hbase org.apache.hadoop.hbase.PerformanceEvaluation —-nomapred randomWrite 3
```

The default Serengeti-deployed HBase cluster does not contain Hadoop JobTracker or Hadoop TaskTracker daemons. To run an HBase MapReduce job, you must deploy a customized cluster that includes JobTracker and TaskTracker nodes.

- Use the client node Rest-ful Web Services, which listen on port 8080, by using the curl command.

  ```
  curl -I http://client_node_ip:8080/status/cluster
  ```

- Use the client node Thrift gateway, which listens on port 9090.

Create a Big Data Cluster in the vSphere Web Client

After you complete deployment of the Hadoop distribution, you can create big data clusters to process data. You can create multiple clusters in your Big Data Extensions environment but your environment must meet all prerequisites and have adequate resources.

Prerequisites

- Start the Big Data Extensions vApp.
- Install theBig Data Extensions plug-in.
- Connect to a Serengeti Management Server.
- Configure one or more Hadoop distributions.
- Understand the topology configuration options that you want to use with your cluster.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions > Big Data Clusters**.

3 In the **Objects** tab, click **New Big Data Cluster**.

4 Follow the prompts to create the new cluster. The table describes the information to enter for the cluster that you want to create.

Option	Description
Hadoop cluster name	Type a name to identify the cluster. The only valid characters for cluster names are alphanumeric and underscores. When you choose the cluster name, also consider the applicable vApp name. Together, the vApp and cluster names must be < 80 characters.
Application manager	Select an application manager. The list contains the default application manager and the application managers that you added to your Big Data Extensions environment. For example, Cloudera Manager and Ambari.
Hadoop distro	Select the Hadoop distribution. The list contains the default Apache Bigtop distribution for Big Data Extensions and the distributions that you added to your Big Data Extensions environment. The distribution names match the value of the `--name` parameter that was passed to the `config-distro.rb` script when the Hadoop distribution was configured. For example, **`cdh5`** and **`mapr`**. NOTE To create an Apache Bigtop, Cloudera CDH4 and CDH5, Hortonworks HDP 2.x, or Pivotal PHD 1.1 or later cluster, you must configure a valid DNS and FQDN for the cluster's HDFS and MapReduce network traffic. If the DNS server cannot provide valid forward and reverse FQDN/IP resolution, the cluster creation process might fail or the cluster is created but does not function.

Option	Description
Local repository URL	Type a local repository URL. This is an optional item for all of application managers. If you specify a local repository URL, the Cloudera Manager or Ambari application manager downloads the required Red Hat Package Managers (RPMs) from the local repository that you specify instead of from a remote repository, which could affect your system performance.
Deployment type	Select the type of cluster you want to create. ■ Basic Hadoop Cluster ■ Basic HBase Cluster ■ Compute Only Hadoop Cluster ■ Compute Workers Only Cluster ■ HBase Only Cluster ■ Data/Compute Separation Hadoop Cluster ■ Customized The type of cluster you create determines the available node group selections. If you select **Customize**, you can load an existing cluster specification file.
DataMaster Node Group	The DataMaster node is a virtual machine that runs the Hadoop NameNode service. This node manages HDFS data and assigns tasks to Hadoop TaskTracker services deployed in the worker node group. Select a resource template from the drop-down menu, or select **Customize** to customize a resource template. For the master node, use shared storage so that you protect this virtual machine with vSphere HA and vSphere FT.
ComputeMaster Node Group	The ComputeMaster node is a virtual machine that runs the Hadoop JobTracker service. This node assigns tasks to Hadoop TaskTracker services deployed in the worker node group. Select a resource template from the drop-down menu, or select **Customize** to customize a resource template. For the master node, use shared storage so that you protect this virtual machine with vSphere HA and vSphere FT.
HBaseMaster Node Group (HBase cluster only)	The HBaseMaster node is a virtual machine that runs the HBase master service. This node orchestrates a cluster of one or more RegionServer slave nodes. Select a resource template from the drop-down menu, or select **Customize** to customize a resource template. For the master node, use shared storage so that you protect this virtual machine with vSphere HA and vSphere FT.
Worker Node Group	Worker nodes are virtual machines that run the Hadoop DataNode, TaskTracker, and HBase HRegionServer services. These nodes store HDFS data and execute tasks. Select the number of nodes and the resource template from the drop-down menu, or select **Customize** to customize a resource template. For worker nodes, use local storage. NOTE You can add nodes to the worker node group by using **Scale Out Cluster**. You cannot reduce the number of nodes.
Client Node Group	A client node is a virtual machine that contains Hadoop client components. From this virtual machine you can access HDFS, submit MapReduce jobs, run Pig scripts, run Hive queries, and HBase commands. Select the number of nodes and a resource template from the drop-down menu, or select **Customize** to customize a resource template. NOTE You can add nodes to the client node group by using **Scale Out Cluster**. You cannot reduce the number of nodes.

Option	Description
Hadoop Topology	Select the topology configuration that you want the cluster to use. ■ RACK_AS_RACK ■ HOST_AS_RACK ■ HVE ■ NONE If you do not see the topology configuration that you want, define it in a topology rack-hosts mapping file, and use the Serengeti Command-Line Interface to upload the file to the Serengeti Management Server. See "About Cluster Topology," on page 100
Network	Select one or more networks for the cluster to use. For optimal performance, use the same network for HDFS and MapReduce traffic in Hadoop and Hadoop+HBase clusters. HBase clusters use the HDFS network for traffic related to the HBase Master and HBase RegionServer services. IMPORTANT You cannot configure multiple networks for clusters that use the MapR Hadoop distribution. ■ To use one network for all traffic, select the network from the **Network** list. ■ To use separate networks for the management, HDFS, and MapReduce traffic, select **Customize the HDFS network and MapReduce network**, and select a network from each network list.
Resource Pools	Select one or more resource pools that you want the cluster to use.
VM Password	Choose how initial administrator passwords are assigned to the virtual machine nodes of the cluster. ■ Use random password. ■ Set password. To assign a custom initial administrator password to all the nodes in the cluster, choose **Set password**, and type and confirm the initial password. Passwords must be from 8 to 20 characters, use only visible lowerASCII characters (no spaces), and must contain at least one uppercase alphabetic character (A - Z), at least one lowercase alphabetic character (a - z), at least one digit (0 - 9), and at least one of the following special characters: _, @, #, $, %, ^, &, * IMPORTANT If you set an initial administrator password, it is used for nodes that are created by future scaling and disk failure recovery operations. If you use the random password, nodes that are created by future scaling and disk failure recovery operations will use new, random passwords.
Local repository URL	Type a local repository URL. This is an optional item for all application managers. If you specify a local repository URL, the Cloudera Manager or Ambari application manager downloads the required Red Hat Package Managers (RPMs) from the local repository that you specify instead of from a remote repository, which could affect your system performance.

The Serengeti Management Server clones the template virtual machine to create the nodes in the cluster. When each virtual machine starts, the agent on that virtual machine pulls the appropriate Big Data Extensions software components to that node and deploys the software.

Create an HBase Only Cluster in Big Data Extensions

With Big Data Extensions, you can create an HBase only cluster, which contain only HBase Master, HBase RegionServer, and Zookeeper nodes, but not Namenodes and Datanodes. The advantage of having an HBase only cluster is that multiple HBase clusters can use the same external HDFS.

Procedure

1. Prerequisites for Creating an HBase Only Cluster on page 104

   Before you can create an HBase only cluster, you must verify that your system meets all of the prerequisites.

2. Prepare the EMC Isilon OneFS as the External HDFS Cluster on page 104

   If you use EMC Isilon OneFS as the external HDFS cluster to the HBase only cluster, you must create and configure users and user groups, and prepare your Isilon OneFS environment.

3. Create an HBase Only Cluster by Using the vSphere Web Client on page 105

   You can use the vSphere Web Client to create an HBase only cluster.

Prerequisites for Creating an HBase Only Cluster

Before you can create an HBase only cluster, you must verify that your system meets all of the prerequisites.

Prerequisites

- Verify that you started the Serengeti vApp.
- Verify that you have more than one distribution if you want to use a distribution other than the default distribution.
- Verify that you have an existing HDFS cluster to use as the external HDFS cluster.

  To avoid conflicts between the HBase only cluster and the external HDFS cluster, the clusters should use the same Hadoop distribution and version.
- If the external HDFS cluster was not created using Big Data Extensions, verify that the HDFS directory `/hadoop/hbase`, the group `hadoop`, and the following users exist in the external HDFS cluster:
  - hdfs
  - hbase
  - serengeti
- If you use the EMC Isilon OneFS as the external HDFS cluster, verify that your Isilon environment is prepared.

  For information about how to prepare your environment, see “Prepare the EMC Isilon OneFS as the External HDFS Cluster,” on page 104.

Prepare the EMC Isilon OneFS as the External HDFS Cluster

If you use EMC Isilon OneFS as the external HDFS cluster to the HBase only cluster, you must create and configure users and user groups, and prepare your Isilon OneFS environment.

Procedure

1. Log in to one of the Isilon HDFS nodes as `user root`.
2. Create the users.
   - hdfs

- hbase
- serengeti
- mapred

The `yarn` and `mapred` users should have write, read, and execute permissions to the entire exported HDFS directory.

3. Create the user group `hadoop`.
4. Create the directory `tmp` under the root HDFS directory.
5. Set the owner as `hdfs:hadoop` with the read and write permissions set as `777`.
6. Create the directory `hadoop` under the root HDFS directory.
7. Set the owner as `hdfs:hadoop` with the read and write permissions set as `775`.
8. Create the directory `hbase` under the directory `hadoop`.
9. Set the owner as `hbase:hadoop` with the read and write permissions set as `775`.
10. Set the owner of the root HDFS directory as `hdfs:hadoop`.

Example: Configuring the EMC Isilon OneFS Environment

```
isi auth users create --name="hdfs"
isi auth users create --name="hbase"
isi auth users create --name="serengeti"
isi auth groups create --name="hadoop"
pw useradd mapred -G wheel
pw useradd yarn -G wheel
chown hdfs:hadoop /ifs
mkdir /ifs/tmp
chmod 777 /ifs/tmp
chown hdfs:hadoop /ifs/tmp
mkdir -p /ifs/hadoop/hbase
chmod -R 775 /ifs/hadoop
chown hdfs:hadoop /ifs/hadoop
chown hbase:hadoop /ifs/hadoop/hbase
```

What to do next

You are now ready to create the HBase only cluster with the EMC Isilon OneFS as the external cluster.

Create an HBase Only Cluster by Using the vSphere Web Client

You can use the vSphere Web Client to create an HBase only cluster.

You must use the default application manager because the other application managers do not support HBase only clusters.

Procedure

1. In the Big Data Clusters page, click **New Big Data Cluster**.
2. On the General page, enter a name for the cluster.
3. Select **Default** from the **Application Manager** drop-down menu.
4. Select a distribution from the **Hadoop Distribution** drop-down menu.
5. On the Set Node Groups page, select **HBase Only Cluster** from the **Deployment Type** drop-down menu.

6 In the **NameNode URI** text box, enter the external HDFS NameNode URI.

   The NameNode URI is the URI of the NameNode, for example ***`hdfs://namenode_hostname:8020`***.

7 Follow the prompts to complete the HBase cluster creation process.

Create a Cluster with an Application Manager by Using the vSphere Web Client

To create and manage a cluster with an application manager other than the default application manager, you must specify the application manager to use before you create the cluster.

NOTE If you want to use a local yum repository, after you select either Cloudera Manager or Ambari for your application manager, a text box appears where you can enter the URL of the local repository you want to use. It is important that you have created the repository before you create the cluster. For more information about setting up a yum repository, see "Set Up a Local Yum Repository for Ambari Application Manager," on page 64 or "Set Up a Local Yum Repository for Cloudera Manager Application Manager," on page 61.

Prerequisites

- Connect to an application manager.
- Ensure that you have adequate resources allocated to run the Hadoop cluster. For information about resource requirements, see the documentation for your application manager.
- Configure one or more Hadoop distributions.

Procedure

1 In the Big Data Clusters page, click **New Big Data Cluster**.

2 Follow the prompts to create the new cluster.

What to do next

To view the new cluster, from theBig Data Extensions navigation pane, under **Inventory Lists**, click **Big Data Clusters**.

If you do not specify an application manager, the default application manager is used.

Create a Compute-Only Cluster with a Third Party Application Manager by Using vSphere Web Client

You can create compute-only clusters to run MapReduce jobs on existing HDFS clusters, including storage solutions that serve as an external HDFS.

If you use EMC Isilon OneFS as the external HDFS cluster to the HBase only cluster, you must create and configure users and user groups, and prepare your Isilon OneFS environment. See "Prepare the EMC Isilon OneFS as the External HDFS Cluster," on page 104.

Prerequisites

- Deploy the Serengeti vApp.
- Ensure that you have adequate resources allocated to run the Hadoop cluster.
- To use any Hadoop distribution other than the default distribution, add one or more Hadoop distributions. See the *VMware vSphere Big Data Extensions Administrator's and User's Guide*.

Procedure

1 From Big Data Extensions, select **New Big Data Cluster**.

2 On the General panel, select from the pull-down list the application manager you want to use to manage the cluster.

3 To customize the cluster for the Cloudera Manager or Ambari application managers, select **Customize** from the pull down list.

4 Click **Load** to select the specification file.

5 Complete the steps of the wizard to finish the cluster creation process.

Create a Compute Workers Only Cluster by Using the vSphere Web Client

If you already have a physical Hadoop cluster and want to do more CPU or memory intensive operations, you can increase the compute capacity by provisioning a workers only cluster. The workers only cluster is a part of the physical Hadoop cluster and can be scaled out elastically.

With the compute workers only clusters, you can "burst out to virtual." It is a temporary operation that involves borrowing resources when you need them and then returning the resources when you no longer need them. With "burst out to virtual," you spin up compute only workers nodes and add them to either an existing physical or virtual Hadoop cluster.

Worker only clusters are not supported on Ambari and Cloudera Manager application managers.

Prerequisites

- Ensure that you have an existing Hadoop cluster.
- Verify that you have the IP addresses of the NameNode and ResourceManager node.

Procedure

1 Click **Create Big Data Cluster** on the objects pane.

2 In the Create Big Data Cluster wizard, choose the same distribution as the Hadoop cluster.

3 Set the DataMaster URL HDFS:*namenode ip or fqdn*:8020.

4 Set the ComputeMaster URL *nodeManager ip or fqdn*.

5 Follow the steps in the wizard and add the other resources.

There will be three node managers in the cluster. The three new node managers are registered to the resource manager.

Managing Hadoop and HBase Clusters 9

You can use the vSphere Web Client to start and stop your big data cluster and modify the cluster configuration. You can also manage a cluster using the Serengeti Command-Line Interface.

⚠ CAUTION Do not use vSphere management functions such as migrating cluster nodes to other hosts for clusters that you create with Big Data Extensions. Performing such management functions outside of the Big Data Extensions environment can make it impossible for you to perform some Big Data Extensions operations, such as disk failure recovery.

This chapter includes the following topics:

Stop and Start a Cluster in the vSphere Web Client

You can stop a running Hadoop cluster and start a stopped Hadoop cluster from the vSphere Web Client.

Prerequisites

- To stop a cluster it must be running.
- To start a cluster it must be stopped.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 From the Inventory Lists, click **Big Data Clusters**.

4 Select the cluster to stop or start from the Hadoop Cluster Name column, and right-click to display the Actions menu.

5 Select **Shut Down Big Data Cluster** to stop a running cluster, or select **Start Big Data Cluster** to start a cluster.

Delete a Cluster in the vSphere Web Client

You can delete a cluster by using the vSphere Web Client. When you delete a cluster, it is removed from the inventory and the datastore.

When you create a cluster, Big Data Extensions creates a folder and a resource pool for each cluster, and resource pools for each node group in the cluster. When you delete a cluster all of these organizational folders and resource pools are also removed.

When you delete a cluster, it is removed from the inventory and the datastore.

You can delete a running cluster, a stopped cluster, or a cluster that is in an error state.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 In the object navigator, select **Big Data Extensions**.

3 In Inventory Lists, select **Big Data Clusters**.

4 From the Objects Name column, select the cluster to delete.

5 Click the **All Actions** icon, and select **Delete Big Data Cluster**.

The cluster and all the virtual machines it contains are removed from your Big Data Extensions environment.

About Resource Usage and Elastic Scaling

Scaling lets you adjust the compute capacity of Hadoop data-compute separated clusters. When you enable elastic scaling for a Hadoop cluster, the Serengeti Management Server can stop and start compute nodes to match resource requirements to available resources. You can use manual scaling for more explicit cluster control.

Manual scaling is appropriate for static environments where capacity planning can predict resource availability for workloads. Elastic scaling is best suited for mixed workload environments where resource requirements and availability fluctuate.

When you select manual scaling, Big Data Extensions disables elastic scaling. You can configure the target number of compute nodes for manual scaling. If you do not configure the target number of compute nodes, Big Data Extensions sets the number of active compute nodes to the current number of active compute nodes. If nodes become unresponsive, they remain in the cluster and the cluster operates with fewer functional nodes. In contrast, when you enable elastic scaling, Big Data Extensions manages the number of active TaskTracker nodes according to the range that you specify, replacing unresponsive or faulty nodes with live, responsive nodes.

For both manual and elastic scaling, Big Data Extensions, not vCenter Server, controls the number of active nodes. However, vCenter Server applies the usual reservations, shares, and limits to the resource pool of a cluster according to the vSphere configuration of the cluster. vSphere DRS operates as usual, allocating resources between competing workloads, which in turn influences how Big Data Extensions dynamically adjusts the number of active nodes in competing Hadoop clusters while elastic scaling is in effect.

Big Data Extensions also lets you adjust the access priority for the datastores of cluster nodes by using the vSphere Storage I/O Control feature. Clusters configured for HIGH I/O shares receive higher priority access than clusters with NORMAL priority. Clusters configured for NORMAL I/O shares receive higher priority access than clusters with LOW priority. In general, higher priority provides better disk I/O performance.

Scaling Modes

To change between manual and elastic scaling, you change the scaling mode.

- MANUAL. Big Data Extensions disables elastic scaling. When you change to manual scaling, you can configure the target number of compute nodes. If you do not configure the target number of compute nodes, Big Data Extensions sets the number of active compute nodes to the current number of active compute nodes.
- AUTO. Enables elastic scaling. Big Data Extensions manages the number of active compute nodes, maintaining the number of compute nodes in the range from the configured minimum to the configured maximum number of compute nodes in the cluster. If the minimum number of compute nodes is undefined, the lower limit is 0. If the maximum number of compute nodes is undefined, the upper limit is the number of available compute nodes.

  Elastic scaling operates on a per-host basis, at a node-level granularity. That is, the more compute nodes a Hadoop cluster has on a host, the finer the control that Big Data Extensions elasticity can exercise. The tradeoff is that the more compute nodes you have, the higher the overhead in terms of runtime resource cost, disk footprint, I/O requirements, and so on.

  When resources are overcommitted, elastic scaling reduces the number of powered on compute nodes. Conversely, if the cluster receives all the resources it requested from vSphere, and Big Data Extensions determines that the cluster can make use of additional capacity, elastic scaling powers on additional compute nodes.

  Resources can become overcommitted for many reasons, such as:

  - The compute nodes have lower resource entitlements than a competing workload, according to how vCenter Server applies the usual reservations, shares, and limits as configured for the cluster.
  - Physical resources are configured to be available, but another workload is consuming those resources.

  In elastic scaling, Big Data Extensions has two different behaviors for deciding how many active compute nodes to maintain. In both behaviors, Big Data Extensions replaces unresponsive or faulty nodes with live, responsive nodes.

  - Variable. The number of active, healthy TaskTracker compute nodes is maintained from the configured minimum number of compute nodes to the configured maximum number of compute nodes. The number of active compute nodes varies as resource availability fluctuates.
  - Fixed. The number of active, healthy TaskTracker compute nodes is maintained at a fixed number when the same value is configured for the minimum and maximum number of compute nodes.

Default Cluster Scaling Parameter Values

When you create a cluster, the scaling configuration of the cluster is as follows.

- The scaling mode is MANUAL, for manual scaling.
- The minimum number of compute nodes is -1. It appears as "Unset" in the Serengeti CLI displays. Big Data Extensions elastic scaling treats a `minComputeNodeNum` value of -1 as if it were zero (0).
- The maximum number of compute nodes is -1. It appears as "Unset" in the Serengeti CLI displays. Big Data Extensions elastic scaling treats a `maxComputeNodeNum` value of -1 as if it were unlimited.

- The target number of nodes is not applicable. Its value is -1. Big Data Extensions manual scaling operations treat a `targetComputeNodeNum` value of -1 as if it were unspecified upon a change to manual scaling.

Interactions Between Scaling and Other Cluster Operations

Some cluster operations cannot be performed while Big Data Extensions is actively scaling a cluster.

If you try to perform the following operations while Big Data Extensions is scaling a cluster in MANUAL mode, Big Data Extensions warns you that in the current state of the cluster, the operation cannot be performed.

- Concurrent attempt at manual scaling
- Switch to AUTO mode while manual scaling operations are in progress

If a cluster is in AUTO mode for elastic scaling when you perform the following cluster operations on it, Big Data Extensions changes the scaling mode to MANUAL and changes the cluster to manual scaling. You can re-enable the AUTO mode for elastic scaling after the cluster operation finishes, except if you delete the cluster.

- Delete the cluster
- Repair the cluster
- Stop the cluster

If a cluster is in AUTO mode for elastic scaling when you perform the following cluster operations on it, Big Data Extensions temporarily switches the cluster to MANUAL mode. When the cluster operation finishes, Big Data Extensions returns the scaling mode to AUTO, which re-enables elastic scaling.

- Resize the cluster
- Reconfigure the cluster

If Big Data Extensions is scaling a cluster when you perform an operation that changes the scaling mode to MANUAL, your requested operation waits until the scaling finishes, and then the requested operation begins.

Optimize Cluster Resource Usage with Elastic Scaling in the vSphere Web Client

You can specify the scaling mode of a cluster. Scaling lets you specify the number of nodes that the cluster can use, and whether it adds nodes or uses nodes within a targeted range.

When you enable elastic scaling for a cluster, Big Data Extensions optimizes cluster performance and use of nodes that have a Hadoop TaskTracker role.

When you set a cluster's scaling mode to AUTO, configure the minimum number of compute nodes. If you do not configure the minimum and maximum number of compute nodes, the previous settings are retained. When you set a cluster's scaling mode to MANUAL, configure the target number of compute nodes. If you do not configure the target number of compute nodes, Big Data Extensions sets the number of active compute nodes to the current number of active compute nodes.

In elastic scaling, Big Data Extensions has two different behaviors for deciding how many active compute nodes to maintain. In both behaviors, Big Data Extensions replaces unresponsive or faulty nodes with live, responsive nodes.

- Variable. The number of active, healthy TaskTracker compute nodes is maintained from the configured minimum number of compute nodes to the configured maximum number of compute nodes. The number of active compute nodes varies as resource availability fluctuates.
- Fixed. The number of active, healthy TaskTracker compute nodes is maintained at a fixed number when the same value is configured for the minimum and maximum number of compute nodes.

Prerequisites

- Understand how elastic scaling and resource usage work. See "About Resource Usage and Elastic Scaling," on page 110.
- Verify that the cluster you want to optimize is data-compute separated. See "About Hadoop and HBase Cluster Deployment Types," on page 99

Procedure

1. Use the vSphere Web Client to log in to vCenter Server.
2. In the object navigator select **Big Data Extensions**.
3. Under Inventory Lists click **Big Data Clusters**.
4. Select the cluster whose elasticity mode you want to set from the Hadoop Cluster Name column.
5. Click the **All Actions** icon, and select **Set Elasticity Mode**.
6. Specify the elasticity settings for the cluster that you want to modify.

Option	Description
Elasticity mode	Select the type of elasticity mode you want to use. You can choose manual or automatic.
Target compute nodes	Specify the number of compute nodes the cluster should target for use. This option is applicable only to manual scaling (manual elasticity mode). If you do not specify the target number of compute nodes, the node setting remains unconfigured, and Big Data Extensions sets the number of active compute nodes to the current number of active compute nodes. **NOTE** A value of "Unset" or "-1" means that the node setting has not been configured and is not applicable.
Min compute nodes	Specify the minimum number (the lower limit) of active compute nodes to maintain in the cluster. This option is applicable only to elastic scaling (automatic elasticity mode). To ensure that under contention elasticity keeps a cluster operating with more than a cluster's initial default setting of zero compute nodes, configure the minimum number of compute nodes to a nonzero number.
Max compute nodes	Specify the maximum number (the upper limit) of active compute nodes to maintain in the cluster. This option is applicable only to elastic scaling (automatic elasticity mode).

What to do next

Specify the cluster's access priority for datastores. See "Use Disk I/O Shares to Prioritize Cluster Virtual Machines in the vSphere Web Client," on page 116.

Schedule Fixed Elastic Scaling for a Hadoop Cluster

You can enable fixed, elastic scaling according to a preconfigured schedule. Scheduled fixed, elastic scaling provides more control than variable, elastic scaling while still improving efficiency, allowing explicit changes in the number of active compute nodes during periods of predictable usage.

For example, in an office with typical workday hours, there is likely a reduced load on a VMware View resource pool after the office staff goes home. You could configure scheduled fixed, elastic scaling to specify a greater number of compute nodes from 8 PM to 4 AM, when you know that the workload would otherwise be very light.

Prerequisites

From the Serengeti Command-Line Interface, enable the cluster for elastic scaling, and set the `minComputeNodeNum` and `MaxComputeNodeNum` parameters to the same value: the number of active TaskTracker nodes that you want during the period of scheduled fixed elasticity.

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

2 Use any scheduling mechanism that you want to call the `/opt/serengeti/sbin/set_compute_node_num.sh` script to set the number of active TaskTracker compute nodes that you want.

```
/opt/serengeti/sbin/set_compute_node_num.sh --name cluster_name
--computeNodeNum num_TT_to_maintain
```

After the scheduling mechanism calls the `set_compute_node_num.sh` script, fixed, elastic scaling remains in effect with the configured number of active TaskTracker compute nodes until the next scheduling mechanism change or until a user changes the scaling mode or parameters in either the vSphere Web Client or the Serengeti Command-Line Interface.

This example shows how to use a crontab file on the Serengeti Management Server to schedule specific numbers of active TaskTracker compute nodes.

```
# cluster_A: use 20 active TaskTracker compute nodes from 11:00 to 16:00, and 30 compute
nodes the rest of the day
00 11   *  *  *   /opt/serengeti/sbin/set_compute_node_num.sh --name cluster_A
--computeNodeNum 20 >> $HOME/schedule_elasticity.log 2>&1
00 16   *  *  *   /opt/serengeti/sbin/set_compute_node_num.sh --name cluster_A
--computeNodeNum 30 >> $HOME/schedule_elasticity.log 2>&1

# cluster_B: use 3 active TaskTracker compute nodes beginning at 10:00 every weekday
0  10   *  *  1-5 /opt/serengeti/sbin/set_compute_node_num.sh --name cluster_B
--computeNodeNum 3  >> $HOME/schedule_elasticity.log 2>&1

# cluster_C: reset the number of active TaskTracker compute nodes every 6 hours to 15
0  */6  *  *  *   /opt/serengeti/sbin/set_compute_node_num.sh --name cluster_B
--computeNodeNum 15 >> $HOME/schedule_elasticity.log 2>&1
```

Scale a Cluster in or out by using the vSphere Web Client

When you create Hadoop clusters you must specify the number of nodes to use. After the cluster is created, you can resize the cluster by changing the number of worker nodes and client nodes. You can increase the number of nodes to scale out a node group. You can also decrease the number of nodes to scale in a compute-only node group. A node group is considered to be a compute-only node group if it only contains compute roles such as tasktracker or nodemanger.

You can resize the cluster using the vSphere Web Client or the Serengeti CLI Client. However, the CLI provides more configuration options than the vSphere Web Client. See the *VMware vSphere Big Data Extensions Command-Line Interface Guide*.

By default you can only scale in compute nodes. To scale in node groups containing other roles (for example role A and role B), you need to login to the Big Data Extensions server and remove role A and role B in related blacklist files. The blacklist file name is `scale_in_roles_blacklist.json`, and is located in the directory `/opt/serengeti/conf/`*`application_manager_type`*. The *application_manager_type* can be Ambari, Cloudera Manager, or Default.

IMPORTANT Even if you changed the user password on the nodes, the changed password is not used for the new nodes that are created when you resize a cluster. If you set the initial administrator password when you created the cluster, that initial administrator password is used for the new nodes. If you did not set the initial administrator password when you created the cluster, new random passwords are used for the new nodes.

Prerequisites

- Verify that the cluster is running. See "Stop and Start a Cluster in the vSphere Web Client," on page 109.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 From the Inventory List, select **Big Data Clusters**.

4 From the Hadoop Cluster Name column, select the cluster to resize.

5 Click the **All Actions** icon, and select **Scale Out/In**.

6 From the **Node Group** list, select the worker or client node group to scale out or to scale in.

If a node group does not have any nodes, it does not appear in the **Node group** list.

7 In the **Instance number** text box, type the target number of node instances to add, and click **OK**.

The cluster is scaled to the specified number of nodes.

Scale CPU and RAM in the vSphere Web Client

You can increase or decrease the compute capacity of a cluster to prevent CPU or memory resource contention among running jobs.

You can adjust compute resources without increasing the workload on the Master node. If increasing or decreasing the CPU or RAM of a cluster is unsuccessful for a node, which is commonly because of insufficient resources being available, the node is returned to its original CPU or RAM setting.

All node types support CPU and RAM scaling, but do not scale the master node CPU or RAM of a cluster because Big Data Extensions powers down the virtual machine during the scaling process.

When you scale the CPU or RAM of a cluster, the number of CPUs must be a multiple of the number of cores per socket, and you must scale the amount of RAM as a multiple of 4, allowing a minimum of 3748 MB.

Prerequisites

- Verify that the cluster that you want to scale is running. See "Stop and Start a Cluster in the vSphere Web Client," on page 109.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 From the Inventory Lists, select **Big Data Clusters**.

4 From the Hadoop Cluster Name column, select the cluster that you want to scale up or down.

5 Click the **All Actions** icon, and select **Scale Up/Down**.

6 From the **Node group** drop-down menu, select the ComputeMaster, DataMaster, Worker, Client, or Customized node group whose CPU or RAM you want to scale up or down.

7 Enter the number of vCPUs to use and the amount of RAM and click **OK**.

After applying new values for CPU and RAM, the cluster is placed into Maintenance mode as it applies the new values. You can monitor the status of the cluster as the new values are applied.

Use Disk I/O Shares to Prioritize Cluster Virtual Machines in the vSphere Web Client

You can set the disk I/O shares for the virtual machines running a cluster. Disk shares distinguish high-priority virtual machines from low-priority virtual machines.

Disk shares is a value that represents the relative metric for controlling disk bandwidth to all virtual machines. The values are compared to the sum of all shares of all virtual machines on the server and, on an ESXi host, the service console. Big Data Extensions can adjust disk shares for all virtual machines in a cluster. Using disk shares you can change a cluster's I/O bandwidth to improve the cluster's I/O performance.

For more information about using disk shares to prioritize virtual machines, see the VMware vSphere ESXi and vCenter Server documentation.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 In the object navigator select **Big Data Extensions**.

3 In the Inventory Lists click **Big Data Clusters**.

4 Select the cluster whose disk IO shares you want to set from the Hadoop Cluster Name column.

5 Click the **Actions** icon, and select **Set Disk IO Share**.

6 Specify a value to allocate a number of shares of disk bandwidth to the virtual machine running the cluster.

   Clusters configured for HIGH I/O shares receive higher priority access than those with NORMAL and LOW priorities, which provides better disk I/O performance. Disk shares are commonly set LOW for compute virtual machines and NORMAL for data virtual machines. The master node virtual machine is commonly set to NORMAL.

7 Click **OK** to save your changes.

About vSphere High Availability and vSphere Fault Tolerance

The Serengeti Management Server leverages vSphere HA to protect the Hadoop master node virtual machine, which can be monitored by vSphere.

When a Hadoop NameNode or JobTracker service stops unexpectedly, vSphere restarts the Hadoop virtual machine in another host, reducing unplanned downtime. If vsphere Fault Tolerance is configured and the master node virtual machine stops unexpectedly because of host failover or loss of network connectivity, the secondary node is used, without downtime.

Change the User Password on All of the Nodes of a Cluster

You can change the user password for all nodes in a cluster. The user password that you can change includes the `serengeti` and `root` users.

Passwords must be from 8 to 20 characters, use only visible lowerASCII characters (no spaces), and must contain at least one uppercase alphabetic character (A - Z), at least one lowercase alphabetic character (a - z), at least one digit (0 - 9), and at least one of the following special characters: _, @, #, $, %, ^, &, *

Important If you scale out or perform disk recovery operations on a cluster after you change the user password for the cluster's original nodes, the changed password is not used for the new cluster nodes that are created by the scale out or disk recovery operation. If you set the cluster's initial administrator password when you created the cluster, that initial administrator password is used for the new nodes. If you did not set the cluster's initial administrator password when you created the cluster, new random passwords are used for the new nodes.

Prerequisites

- Deploy the Big Data Extensions vApp. See "Deploy the Big Data Extensions vApp in the vSphere Web Client," on page 24 .
- Configure a Hadoop distribution to use with Big Data Extensions.
- Create a cluster.

Procedure

1. Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.
2. Run the `serengeti-ssh.sh` script..

   ```
   serengeti-ssh.sh cluster_name 'echo new_password | sudo passwd username --stdin'
   ```

   This example changes the password for all nodes in the cluster labeled `mycluster` for the user `serengeti` to `mypassword`.

   ```
   serengeti-ssh.sh mycluster  'echo mypassword | sudo passwd serengeti --stdin'
   ```

The password for the user account that you specify changes on all the nodes in the cluster.

Reconfigure a Cluster with the Serengeti Command-Line Interface

You can reconfigure any big data cluster that you create with Big Data Extensions.

The cluster configuration is specified by attributes in Hadoop distribution XML configuration files such as: `core-site.xml`, `hdfs-site.xml`, `mapred-site.xml`, `hadoop-env.sh`, `yarn-env.sh`, `yarn-site.sh`, and `hadoop-metrics.properties`.

Note Always use the `cluster config` command to change the parameters specified by the configuration files. If you manually modify these files, your changes will be erased if the virtual machine is rebooted, or you use the `cluster config`, `cluster start`, `cluster stop`, or `cluster resize` commands.

Procedure

1 Use the `cluster export` command to export the cluster specification file for the cluster that you want to reconfigure.

```
cluster export --name    cluster_name --specFile file_path/cluster_spec_file_name
```

Option	Description
cluster_name	Name of the cluster that you want to reconfigure.
file_path	The file system path to which to export the specification file.
cluster_spec_file_name	The name with which to label the exported cluster specification file.

2 Edit the configuration information located near the end of the exported cluster specification file.

If you are modeling your configuration file on existing Hadoop XML configuration files, use the `convert-hadoop-conf.rb` conversion tool to convert Hadoop XML configuration files to the required JSON format.

```
…
"configuration": {
    "hadoop": {
      "core-site.xml": {
        // check for all settings at http://hadoop.apache.org/common/docs/stable/core-
default.html
        // note: any value (int, float, boolean, string) must be enclosed in double quotes
and here is a sample:
        // "io.file.buffer.size": "4096"
      },
      "hdfs-site.xml": {
        // check for all settings at http://hadoop.apache.org/common/docs/stable/hdfs-
default.html
      },
      "mapred-site.xml": {
        // check for all settings at http://hadoop.apache.org/common/docs/stable/mapred-
default.html
      },
      "hadoop-env.sh": {
        // "HADOOP_HEAPSIZE": "",
        // "HADOOP_NAMENODE_OPTS": "",
        // "HADOOP_DATANODE_OPTS": "",
        // "HADOOP_SECONDARYNAMENODE_OPTS": "",
        // "HADOOP_JOBTRACKER_OPTS": "",
        // "HADOOP_TASKTRACKER_OPTS": "",
        // "HADOOP_CLASSPATH": "",
        // "JAVA_HOME": "",
        // "PATH": "",
      },
      "log4j.properties": {
        // "hadoop.root.logger": "DEBUG, DRFA ",
        // "hadoop.security.logger": "DEBUG, DRFA ",
      },
      "fair-scheduler.xml": {
        // check for all settings at
http://hadoop.apache.org/docs/stable/fair_scheduler.html
        // "text": "the full content of fair-scheduler.xml in one line"
      },
      "capacity-scheduler.xml": {
        // check for all settings at
http://hadoop.apache.org/docs/stable/capacity_scheduler.html
      }
    }
  }
…
```

3 (Optional) If the JAR files of your Hadoop distribution are not in the `$HADOOP_HOME/lib` directory, add the full path of the JAR file in `$HADOOP_CLASSPATH` to the cluster specification file.

This action lets the Hadoop daemons locate the distribution JAR files.

For example, the Cloudera CDH3 Hadoop Fair Scheduler JAR files are in `/usr/lib/hadoop/contrib/fairscheduler/`. Add the following to the cluster specification file to enable Hadoop to use the JAR files.

```
…
"configuration": {
  "hadoop": {
    "hadoop-env.sh": {
      "HADOOP_CLASSPATH": "/usr/lib/hadoop/contrib/fairscheduler/*:$HADOOP_CLASSPATH"
    },
    "mapred-site.xml": {
      "mapred.jobtracker.taskScheduler": "org.apache.hadoop.mapred.FairScheduler"
      …
    },
    "fair-scheduler.xml": {
      …
    }
  }
}
…
```

4 Access the Serengeti CLI.

5 Run the `cluster config` command to apply the new Hadoop configuration.

```
cluster config --name cluster_name --specFile file_path/cluster_spec_file_name
```

6 (Optional) Reset an existing configuration attribute to its default value.

a Remove the attribute from the configuration section of the cluster configuration file or comment out the attribute using double back slashes (//).

b Re-run the `cluster config` command.

Recover from Disk Failure with the Serengeti Command-Line Interface Client

If there is a disk failure in a cluster, and the disk does not perform management roles such as NameNode, JobTracker, ResourceManager, HMaster, or ZooKeeper, you can recover by running the Serengeti `cluster fix` command.

Big Data Extensions uses a large number of inexpensive disk drives for data storage (configured as JBOD). If several disks fail, the Hadoop data node might shutdown. Big Data Extensions enables you to recover from disk failures.

Serengeti supports recovery from swap and data disk failure on all supported Hadoop distributions. Disks are recovered and started in sequence to avoid the temporary loss of multiple nodes at once. A new disk matches the storage type and placement policies of the corresponding failed disk.

The MapR distribution does not support recovery from disk failure by using the `cluster fix` command.

IMPORTANT Even if you changed the user password on the nodes of the cluster, the changed password is not used for the new nodes that are created by the disk recovery operation. If you set the initial administrator password of the cluster when you created the cluster, that initial administrator password is used for the new nodes. If you did not set the initial administrator password of the cluster when you created the cluster, new random passwords are used for the new nodes.

Procedure

1 Access the Serengeti CLI.

2 Run the `cluster fix` command.

The `nodeGroup` parameter is optional.

```
cluster fix --name cluster_name --disk [--nodeGroup nodegroup_name]
```

Enter Maintenance Mode to Perform Backup and Restore with the Serengeti Command-Line Interface Client

Before performing backup and restore operations, or other maintenance tasks, you must place Big Data Extensions into maintenance mode.

Prerequisites

- Deploy the Serengeti vApp.
- Ensure that you have adequate resources allocated to run the Hadoop cluster.
- To use any Hadoop distribution other than the default distribution, add one or more Hadoop distributions. See the *VMware vSphere Big Data Extensions Administrator's and User's Guide*.

Procedure

1 Log into the Serengeti Management Server.

2 Run the script `/opt/serengeti/sbin/serengeti-maintenance.sh` to place Big Data Extensions into maintenance mode, or check maintenance status.

```
serengeti-maintenance.sh on | off | status
```

Option	Description
on	Turns on maintenance mode. Upon entering maintenance mode, Big Data Extensions continues executing jobs that have already been started, but will not respond to any new requests.
off	Turn off maintenance mode, and returns Big Data Extensions to its normal operating state.
status	Displays the maintenance status of Big Data Extensions. ■ A status of `safe` means it is safe to backup or perform other maintenance tasks on your Big Data Extensions deployment. ■ A status of `off` means maintenance mode has been turned off, and it is not safe to perform maintenance tasks such as backup and restore. ■ A status of `on` means Big Data Extensions has entered maintenance mode, but it is not yet safe to perform back and restore operations. You must wait until the system returns the `safe` status message.

To place your Big Data Extensions deployment into maintenance mode, run the `serengeti-maintenance.sh` script with the `on` option.

```
serengeti-maintenance.sh on
```

3 Verify that Big Data Extensions is in maintenance mode.

   When Big Data Extensions completes all jobs that have been submitted, the maintenance status will enter safe mode. Run the `serengeti-maintenance.sh` with the `status` parameter repeatedly until it returns the `safe` system status message.

```
serengeti-maintenance.sh status
safe
```

4 Perform the necessary system maintenance tasks.

5 Once you have completed the necessary system maintenance tasks, return Big Data Extensions to its normal operating state by manually exiting maintenance mode.

```
serengeti-maintenance.sh off
```

Log in to Hadoop Nodes with the Serengeti Command-Line Interface Client

To perform troubleshooting or to run your management automation scripts, log in to Hadoop master, worker, and client nodes with SSH from the Serengeti Management Server using SSH client tools such as SSH, PDSH, ClusterSSH, and Mussh, which do not require password authentication.

To connect to Hadoop cluster nodes over SSH, you can use a user name and password authenticated login. All deployed nodes are password-protected with either a random password or a user-specified password that was assigned when the cluster was created.

Prerequisites

Use the vSphere Web Client to log in to vCenter Server, and verify that the Serengeti Management Server virtual machine is running.

Procedure

1 Right-click the Serengeti Management Server virtual machine and select **Open Console**.

   The password for the Serengeti Management Server appears.

   NOTE If the password scrolls off the console screen, press Ctrl+D to return to the command prompt.

2 Use the vSphere Web Client to log in to the Hadoop node.

   The password for the `root` user appears on the virtual machine console in the vSphere Web Client.

3 Change the password of the Hadoop node by running the `set-password -u` command.

```
sudo /opt/serengeti/sbin/set-password -u
```

Monitoring the Big Data Extensions Environment 10

You can monitor the status of Serengeti-deployed clusters, including their datastores, networks, and resource pools through the Serengeti Command-Line Interface. You can also view a list of available Hadoop distributions. Monitoring capabilities are also available in the vSphere Web Client.

This chapter includes the following topics:

Enable the Big Data Extensions Data Collector

If you did not enable the Big Data Extensions data collector during installation, you can enable it at a later time. The Customer Experience Improvement Program collects product usage data from your Big Data Extensions environment for analysis and troubleshooting.

The data collector collects four types of data including the Big Data Extensions footprint, operations information, environmental information, and cluster snapshots.

Prerequisites

- Review the Customer Experience Improvement Program description, and determine if you wish to collect data and send it to VMware help improve your user experience using Big Data Extensions. See "The Customer Experience Improvement Program," on page 23.
- Install Big Data Extensions. See Chapter 2, "Installing Big Data Extensions," on page 19

Procedure

1. Use the vSphere Web Client to log in to vCenter Server.
2. Select Big Data Extensions and click the **Manage** tab.
3. In the Customer Experience Improvement Program pane click **Edit**.

   The Customer Experience Improvement Program dialog box appears.

4 Select the **Enable Customer Experience Improvement Program** check box.

What to do next

You can disable the data collector at a later time if you wish to discontinue your use of the Customer Experience Improvement Program. See "Disable the Big Data Extensions Data Collector," on page 124.

Disable the Big Data Extensions Data Collector

The Customer Experience Improvement Program collects product usage data from your Big Data Extensions environment for analysis and troubleshooting if necessary. If you do not want to use the Customer Experience Improvement Program feature, you can disable the Big Data Extensions data collector.

The data collector collects four types of data including the Big Data Extensions footprint, operations information, environmental information, and cluster snapshots. If you disable the Customer Experience Improvement Program, this data is not available for use in troubleshooting and problem resolution.

Procedure

1 Use the vSphere Web Client to log in to Big Data Extensions.

2 Select Big Data Extensions and click the **Manage** tab.

3 In the Customer Experience Improvement Program pane click **Edit**.

The Customer Experience Improvement Program dialog box appears.

4 Deselect the **Enable Customer Experience Improvement Program** check box.

What to do next

You can enable the data collector at a later time if you choose to use of the Customer Experience Improvement Program. See "Enable the Big Data Extensions Data Collector," on page 123.

View Serengeti Management Server Initialization Status

You can you view the initialization status of the Serengeti Management Server services, view error messages to help troubleshoot problems, and recover services that may not have successfully started.

Big Data Extensions may not successfully start for many reasons. The Serengeti Management Server Administration Portal lets you view the initialization status of the Serengeti services, view error messages for individual services to help troubleshoot problems, and recover services that may not have successfully started.

Prerequisites

- Ensure that you know the IP address of the Serengeti Management Server to which you want to connect.
- Ensure that you have login credentials for the Serengeti Management Server `root` user.

Procedure

1 Open a Web browser and go the URL of the Serengeti Management Server Administration Portal.

```
https://management-server-ip-address:5480
```

2 Type **`root`** for the user name, type the password, and click **Login**.

3 Click the **Summary** tab.

The Serengeti Management Server services and their operational status is displayed in the Summary page.

4 Do one of the following.

Option	Description
View Initialize Status	Click **Details**. The Serengeti Server Setup dialog box lets you view the initialization status of the Serengeti Management Server. If the Serengeti Management Server fails to initialize, an error message with troubleshooting information displays. Once you resolve the error, a **Retry** button lets you restart the failed service.
View Chef Server Services	Click the **Chef Server** tree control to expand the list of Chef services.
Recover a Stopped or Failed Service	Click **Recover** to restart a stopped or failed service. If a service fails due to a configuration error, you must first resolve the problem that caused the service to fail before you can successfully recover the failed service.
Refresh	Click **Refresh** to update the information displayed in the Summary page.

What to do next

If there is an error that you need to resolve, the troubleshooting topics provide solutions to problems you might encounter when using Big Data Extensions. See Chapter 13, "Troubleshooting," on page 139.

View Provisioned Clusters in the vSphere Web Client

You can view the clusters deployed within Big Data Extensions, including information about whether the cluster is running, the type of Hadoop distribution used by a cluster, and the number and type of nodes in the cluster.

Prerequisites

- Create one or more clusters whose information you can view.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 In the Inventory Lists, select **Big Data Clusters**.

4 Select **Big Data Clusters**.

Information about all provisioned clusters appears in the right pane.

Table 10-1. Cluster Information

Option	Description
Name	Name of the cluster.
Status	Status of the cluster.
Distribution	Hadoop distribution in use by the cluster.
Elasticity Mode	The elasticity mode in use by the cluster.
Disk IO Shares	The disk I/O shares in use by the cluster.
Resources	The resource pool or vCenter Server cluster in use by the Big Data cluster.
Managed by	The application manager that manages the cluster.
Information	Number and type of nodes in the cluster.
Progress	Status messages of actions being performed on the cluster.

View Cluster Information in the vSphere Web Client

Use the vSphere Web Client to view virtual machines running each node, resource allocation, IP addresses, and storage information for each node in the Hadoop cluster.

Prerequisites

- Create one or more Hadoop clusters.
- Start the Hadoop cluster.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 From the Inventory Lists, click **Big Data Clusters**.

4 Click a Big Data cluster.

Information about the cluster appears in the right pane, in the **Nodes** tab.

Table 10-2. Cluster Information

Column	Description
Node Group	Lists all nodes by type in the cluster.
VM Name	Name of the virtual machine on which a node is running.
Management Network	IP address of the virtual machine.
Host	Host name, IP address, or Fully Qualified Domain Name (FQDN) of the ESXi host on which the virtual machine is running.
Status	The virtual machine reports the following status types: ■ Not Exist. Status before you create a virtual machine instance in vSphere. ■ Powered On. The virtual machine is powered on after virtual disks and network are configured. ■ VM Ready. A virtual machine is started and IP is ready. ■ Service Ready. Services inside the virtual machine have been provisioned. ■ Bootstrap Failed. A service inside the virtual machine failed to provision. ■ Powered Off. The virtual machine is powered off. ■ Service Alert. There is critical issue reported for the services inside of the virtual machine.* ■ Service Unhealthy. There is an unhealthy issue reported for the services inside of the virtual machine.* * Check the details from the corresponding application manager.
Task	Status of in-progress Serengeti operations.

5 From the **Nodes** tab, select a node group.

Information about the node group appears in the Node details panel of the **Nodes** tab.

Table 10-3. Cluster Node Details

Field	Description
Node Group	Name of the selected node group.
VM Name	Name of the node group's virtual machine.
Management network	Network used for management traffic.
HDFS Network	Network used for HDFS traffic.

Table 10-3. Cluster Node Details (Continued)

Field	Description
MapReduce Network	Network used for MapReduce traffic.
Host	Host name, IP address, or Fully Qualified Domain Name (FQDN) of the ESXi host on which the virtual machine is running.
vCPU	Number of virtual CPUs assigned to the node.
RAM	Amount of RAM used by the node. **Note** The RAM size that appears for each node shows the allocated RAM, not the RAM that is in use.
Storage	The amount of storage allocated for use by the virtual machine running the node.
Error	Indicates a node failure.

Monitor the HDFS Status in the vSphere Web Client

When you configure a Hadoop distribution to use with Big Data Extensions, the Hadoop software includes the Hadoop Distributed File System (HDFS). You can monitor the health and status of HDFS from the vSphere Web Client. The HDFS page lets you browse the Hadoop file system, view NameNode logs, and view cluster information including live, dead, and decommissioning nodes, and NameNode storage information.

HDFS is the primary distributed storage used by Hadoop applications. A HDFS cluster consists of a NameNode that manages the file system metadata and DataNodes that store the actual data.

Prerequisites

- Create one or more Hadoop clusters.

Procedure

1. Use the vSphere Web Client to log in to vCenter Server.
2. Select **Big Data Extensions**.
3. In the Inventory Lists, select **Big Data Clusters**.
4. Select the cluster whose HDFS status you want to view from the **Big Data Cluster List** tab.
5. Select **Open HDFS Status Page** from the **Actions** menu.

   The HDFS status information appears in a new Web page.

Note If you use Big Data Extensions in a vCenter Server environment using IPv6, the vSphere Web Client is unable to access the HDFS Status Page, which uses an IPv4 address. To view the HDFS Status Page, open a Web browser and go to the URL that displays in the error message when you attempt to access the status page as instructed in this procedure.

Monitor MapReduce Status in the vSphere Web Client

The Hadoop software includes MapReduce, a software framework for distributed data processing. You can monitor MapReduce status vSphere Web Client. The MapReduce Web page includes information about scheduling, running jobs, retired jobs, and log files.

Prerequisites

- Create one or more Hadoop clusters whose MapReduce status you can monitor.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 From the Inventory Lists, click **Big Data Clusters**.

4 Select the cluster whose MapReduce status you want to view from the **Big Data Cluster List** tab.

5 Select **Open MapReduce Status Page** from the **Actions** menu.

The MapReduce status information appears in a new Web page.

NOTE If you use Big Data Extensions in a vCenter Server environment using IPv6, the vSphere Web Client is unable to access the MapReduce Status Page, which uses an IPv4 address. To view the MapReduce Status Page, open a Web browser and go to the URL that displays in the error message when you attempt to access the status page as instructed in this procedure.

Monitor HBase Status in the vSphere Web Client

HBase is the Hadoop database. You can monitor the health and status of your HBase cluster, as well as the tables that it hosts, from the vSphere Web Client.

Prerequisites

Create one or more HBase clusters.

Procedure

1 Use the vSphere Web Client to log in to vCenter Server.

2 Select **Big Data Extensions**.

3 In the Inventory Lists, click **Big Data Clusters**.

4 In the **Big Data Cluster List** tab, select the cluster whose HBase status you want to view.

5 From the **Actions** menu, select **Open HBase Status Page**.

The HBase status information appears in a new Web page.

NOTE If you use Big Data Extensions in a vCenter Server environment using IPv6, the vSphere Web Client is unable to access the HBase Status Page, which uses an IPv4 address. To view the HBase Status Page, open a Web browser and go to the URL that displays in the error message when you attempt to access the status page as instructed in this procedure.

Accessing Hive Data with JDBC or ODBC 11

You can run Hive queries from a Java Database Connectivity (JDBC) or Open Database Connectivity (ODBC) application leveraging the Hive JDBC and ODBC drivers.

You can access data from Hive using either JDBC or ODBC.

Hive JDBC Driver

Hive provides a Type 4 (pure Java) JDBC driver, defined in the class `org.apache.hadoop.hive.jdbc.HiveDriver`. When configured with a JDBC URI of the form `jdbc:hive://host:port/dbname`, a Java application can connect to a Hive server running at the specified host and port. The driver makes calls to an interface implemented by the Hive Thrift Client using the Java Thrift bindings.

You can choose to connect to Hive through JDBC in embedded mode by using the URI `jdbc:hive://`. In embedded mode, Hive runs in the same JVM as the application that invokes it. You do not have to launch it as a standalone server, because it does not use the Thrift service or the Hive Thrift Client.

Hive ODBC Driver

The Hive ODBC driver allows applications that support the ODBC protocol to connect to Hive. Like the JDBC driver, the ODBC driver uses Thrift to communicate with the Hive server.

This chapter includes the following topics:

- "Configure Hive to Work with JDBC," on page 129
- "Configure Hive to Work with ODBC," on page 131

Configure Hive to Work with JDBC

The Hive JDBC driver lets you access Hive from a Java program that you write, or from a Business Intelligence or similar application that uses JDBC to communicate with database products.

The default JDBC 2.0 port is 21050. Hive accepts JDBC connections through port 21050 by default. Make sure this port is available for communication with other hosts on your network. For example, ensure that the port is not blocked by firewall software.

Prerequisites

You must have an application that can use the Hive JDBC driver to connect to a Hive server.

Procedure

1 Open a command shell, such as Bash or PuTTY, and log in to the Hive server node.

2 Create the file `HiveJdbcClient.java` with the Java code to connect to the Hive Server.

```
import java.sql.SQLException;
import java.sql.Connection;
import java.sql.ResultSet;
import java.sql.Statement;
import java.sql.DriverManager;
public class HiveJdbcClient {
    private static String driverName = "org.apache.hadoop.hive.jdbc.HiveDriver";
    /**
    * @param args
    * @throws SQLException
    **/
    public static void main(String[] args) throws SQLException {
        try {
            Class.forName(driverName);
        } catch (ClassNotFoundException e){
            // TODO Auto-generated catch block
            e.printStackTrace();
            System.exit(1);
        }
        Connection con = DriverManager.getConnection("jdbc:hive://localhost:10000/default", 
"", "");
        Statement stmt = con.createStatement();
        String tableName = "testHiveDriverTable";
        stmt.executeQuery("drop table " + tableName);
        ResultSet res = stmt.executeQuery("create table " + tableName + " (key int, value 
string)");
        // show tables
        String sql = "show tables '" + tableName + "'";
        System.out.println("Running: " + sql);
        res = stmt.executeQuery(sql);
        if (res.next()) {
            System.out.println(res.getString(1));
        }
        // describe table
        sql = "describe " + tableName;
        System.out.println("Running: " + sql);
        res = stmt.executeQuery(sql);
        while (res.next()) {
            System.out.println(res.getString(1) + "\t" + res.getString(2));
        }
        // load data into table
        // NOTE: filepath has to be local to the hive server
        // NOTE: /tmp/test_hive_server.txt is a ctrl-A separated file with two fields per 
line
        String filepath = "/tmp/test_hive_server.txt";
        sql = "load data local inpath '" + filepath + "' into table " + tableName;
        System.out.println("Running: " + sql);
        res = stmt.executeQuery(sql);
        // select * query
        sql = "select * from " + tableName;
        System.out.println("Running: " + sql);
        res = stmt.executeQuery(sql);
        while (res.next()){
            System.out.println(String.valueOf(res.getInt(1)) + "\t" + res.getString(2));
        }
        // regular hive query
        sql = "select count(1) from " + tableName;
        System.out.println("Running: " + sql);
        res = stmt.executeQuery(sql);
        while (res.next()){
            System.out.println(res.getString(1));
        }
    }
}
```

3 Run the JDBC code using one of the following methods.

- Run the `javac` command identifying the Java code containing the JDBC code.`javac HiveJdbcClient.java`
- Run a shell script to populate the data file, define the classpath, and invoke the JDBC client.

The example below uses Apache Hadoop 1.1.2 distribution. If you are using a different Hadoop distribution, you must update the value of the `HADOOP_CORE` variable to correspond to the version of the distribution you are using.

```
#!/bin/bash
HADOOP_HOME=/usr/lib/hadoop
HIVE_HOME=/usr/lib/hive
echo -e '1\x01foo' > /tmp/test_hive_server.txt
echo -e '2\x01bar' >> /tmp/test_hive_server.txt
HADOOP_CORE=`ls /usr/lib/hadoop-1.1.2/hadoop-core-*.jar`
CLASSPATH=.:$HADOOP_CORE:$HIVE_HOME/conf
for jar_file_name in ${HIVE_HOME}/lib/*.jar
do
CLASSPATH=$CLASSPATH:$jar_file_name
done
java -cp $CLASSPATH HiveJdbcClient
```

Either of these methods establishes a JDBC connection with the Hive server using the host and port information that you specify in the Java application or shell script.

Configure Hive to Work with ODBC

The Hive ODBC driver allows you to access Hive from a program that you write, or a Business Intelligence or similar application that uses ODBC to communicate with database products.

To access Hive data using ODBC, use the ODBC driver recommended for use with your Hadoop distribution.

Prerequisites

- Verify that the Hive ODBC driver supports the application or the third-party product that you intend to use.
- Download an appropriate ODBC connector and configure it for use with your environment.
- Configure a Data Source Name (DSN).

  DSNs specify how an application connects to Hive or other database products. Refer to your particular application's documentation to understand how it connects to Hive and other database products using ODBC.

Procedure

1 Open the **ODBC Data Source Administrator** from the Windows **Start** menu.

2 Click the **System DSN** tab, and click **Add**.

3 Select the ODBC driver that you want to use with your Hadoop distribution, and click **Finish**.

4 Enter values for the following fields.

Option	Description
Data Source Name	Type a name by which to identify the DSN.
Host	Fully qualified hostname or IP address of the node running the Hive service.
Port	Port number for the Hive service. The default is 21000.
Hive Server Type	Set to HiveServer1 or HiveServer2.
Authentication	If you are using Hiveserver2, specify the following. ■ **Mechanism**. Set to User Name. ■ **User Name**. User name with which to run Hive queries.

5 Click **OK**.

6 Click **Test** to test the ODBC connection.

7 After you verify that the connection works, click **Finish**.

The new ODBC connector appears in the User Data Sources list.

What to do next

Configure the application to work with your Hadoop distribution's Hive service. See your particular application's documentation to understand how it connects to Hive and other database products that use ODBC.

Big Data Extensions Security Reference

12

Use the Security Reference to learn about the security features of your Big Data Extensions installation and the measures that you can take to safeguard your environment from attack.

- Services, Network Ports, and External Interfaces on page 133

  The operation of Big Data Extensions depends on certain services, ports, and external interfaces.

- Big Data Extensions Configuration Files on page 135

  Some Big Data Extensions configuration files contain settings that may affect your environment's security.

- Big Data Extensions Public Key, Certificate, and Keystore on page 136

  The Big Data Extensions public key, certificate, and keystore are located on the Serengeti Management Server.

- Big Data Extensions Log Files on page 136

  The files that contain system messages are located on the Serengeti Management Server.

- Big Data Extensions User Accounts on page 137

  You must set up an administrative user and a **root** user account to administer Big Data Extensions.

- Security Updates and Patches on page 137

  You can apply security updates and patches as they are made available by either VMware, or the vendors of operating systems and Hadoop distributions.

Services, Network Ports, and External Interfaces

The operation of Big Data Extensions depends on certain services, ports, and external interfaces.

Big Data Extensions Services

The operation of Big Data Extensions depends on several services that run on the Big Data Extensions vApp.

Table 12-1. Big Data Extensions Services

Service Names	Startup Type	Description
http	Automatic	Apache Web Server Secure remote console access.
sshd	Automatic	Secure remote console access.
rsyslog	Automatic	The `rsyslog` service is an enhanced, multi-threaded `syslog` daemon
Tomcat	Automatic	Tomcat Server which runs the Big Data Extensions Web Service

Table 12-1. Big Data Extensions Services (Continued)

Service Names	Startup Type	Description
Thrift Service	Automatic	The communication broker between Big Data Extensions Web Service and Chef Server's `knife` process.
Chef Server	Automatic	Chef is an open source configuration management framework and tool. The Chef Server is the primary component of the Chef framework.
Nginx	Automatic	Nginx is part of the Chef Server, and acts as the proxy for handling all requests to Chef Server API.
Postgres	Automatic	The database server is use by the Chef Server and Big Data Extensions Web Service.

Big Data Extensions Communication Ports

Big Data Extensions uses several communication ports and protocols.

The table below shows the ports listening on the Serengeti Management Server (also called the Big Data Extensions Management Server) for all local and external network addresses.

Table 12-2. Serengeti Management Server Services and Network Ports

Service Name	Ports	Protocol	Listen on Local Port?
httpd	433/TCP	HTTP	No
sshd	22/TCP	SSH	No
Tomcat	8080/TCP, 8443/TCP	HTTP, HTTPS	No
nginx	9080/TCP, 9443/TCP	HTTP, HTTPS	No
Thrift Service	9090	TCP	Yes
postgres	5432	Postgres	Yes

Big Data Extensions Hadoop and HBase Node Communication Ports

Big Data Extensions deploys Hadoop and HBase clusters which use their default ports for the cluster nodes they deploy.

Table 12-3. Ports in use by Hadoop clusters created with Big Data Extensions

Service Name	Daemon Name	Ports	Protocol
HDFS	Namenode Web page	50070/TCP	HTTP
	Namenode RPC	8020/TCP	RPC
	Datanode	50075/TCP, 50010/TCP, 50020/TCP	RPC
MapReduce	JobTracker Web page	50030/TCP	HTTP
	JobTracker RPC	8021/TCP	RPC
	TaskTracker	50060/TCP	RPC
Yarn	Resource Manager Web page	8088/TCP	HTTP
	Resource Manager RPC	8030/TCP, 8031/TCP, 8032/TCP, 8033/TCP	RPC

Table 12-3. Ports in use by Hadoop clusters created with Big Data Extensions (Continued)

Service Name	Daemon Name	Ports	Protocol
	NodeManager	8040/TCP, 8042/TCP	RPC
Hive	Hive Server	10000/TCP	RPC

Table 12-4. Ports in use by HBase clusters created with Big Data Extensions

Service Name	Ports	Protocol
Zookeeper	2181/TCP	Zookeeper
HBase Master	60000/TCP, 60010/TCP	RPC
HBase RegionServer	60020/TCP, 60030/TCP	RPC
HBase Thrift Service	9090/TCP, 9095/TCP	RPC
HBase REST Service	8080/TCP, 8085/TCP	HTTP

Table 12-5. Ports in use by MapR clusters created with Big Data Extensions

Service Name	Ports	Protocol
CLDB	7222	
CLDB JMX monitor port	7220	
CLDB web port	7221	
HBase Master	60000	
HBase Master (for GUI)	60010	
HBase RegionServer	60020	
Hive Metastore	9083	
JobTracker Webpage	50030	
JobTracker RPC	8021	RPC
MFS server	5660	
MySQL	3306	
NFS	2049	
NFS monitor (for HA)	9997	
NFS management	9998	
Port mapper	111	
TaskTracker	50060	
Web UI HTTPS	8443	
Zookeeper	5180	

Big Data Extensions Configuration Files

Some Big Data Extensions configuration files contain settings that may affect your environment's security.

Big Data Extensions Configuration Files Containing Security-Related Resources

All security-related resources are accessible by the **serengeti** and **root** user accounts. Protecting these user accounts is critical to the security of Big Data Extensions.

Table 12-6. Configuration Files Containing Security-Related Resources

File	Description
`/opt/serengeti/tomcat/conf/server.xml`	Configuration file for theTomcat server, which includes network ports and SSL key store file locations and passwords.
`/opt/serengeti/conf/vc.properties`	Key store configuration file for Big Data Extensions Web Service.
`/var/opt//chef-server/nginx/etc/nginx.conf`	Configuration file for the Nginx Web server, which includes network ports and certificate information.
`/etc/httpd/conf.d/ssl.conf`	Configuration file for thehttpd Web server.

Big Data Extensions Public Key, Certificate, and Keystore

The Big Data Extensions public key, certificate, and keystore are located on the Serengeti Management Server.

All security-related resources are accessible by the **serengeti** and **root** user accounts. Protecting these user accounts is critical to the security of Big Data Extensions.

Table 12-7. Big Data Extensions Public Key, Certificate, and Keystore

File Location	Service
`/opt/serengeti/.certs/`	Tomcat
`/var/opt/chef-server/nginx/ca/`	Nginx
`/etc/pki/tls/private/`	httpd
`/etc/pki/tls/certs/`	httpd

Big Data Extensions Log Files

The files that contain system messages are located on the Serengeti Management Server.

Big Data Extensions uses the following log files to track and record system messages and events. The log files are located on the Serengeti Management Server and Chef Server.

Table 12-8. Big Data Extensions Log Files

File	Description
`/opt/serengeti/logs/serengeti.log`	Tracks and records events for the Big Data Extensions Web Service
`/opt/serengeti/logs/ironfan.log`	Tracks and records events when provisioning new clusters using the default application manager.
`/opt/serengeti/logs/serengeti-boot.log`	Tracks and records events when the Big Data Extensions Server boots up.
`/opt/serengeti/logs/serengeti-upgrade.log`	Tracks and records events when upgrading Big Data Extensions and cluster nodes.
`/opt/serengeti/logs/provision-hook.log`	Tracks and records events when executing hooks during cluster provisioning.
`sudo chef-server-ctl tail`	To track the Chef Server log files run the `tail` command on the `chef-server-ctl` service.

Security-Related Log Messages

Big Data Extensions does not provide any security-related log messages.

Big Data Extensions User Accounts

You must set up an administrative user and a **root** user account to administer Big Data Extensions.

Big Data Extensions Root User Account

The root password of Serengeti Management Server is a random password generated when powering on the Big Data Extensions vApp for the first time. You can see the password in the virtual machine console for Big Data Extensions in the vSphere Web Client.

The root password of Big Data Extensions nodes in a cluster is a random password generated when creating the cluster, or specified by a user before creating the cluster.

Passwords must be from 8 to 20 characters, use only visible lowerASCII characters (no spaces), and must contain at least one uppercase alphabetic character (A - Z), at least one lowercase alphabetic character (a - z), at least one digit (0 - 9), and at least one of the following special characters: _, @, #, $, %, ^, &, *

Only visible lower ASCII characters (No spaces)

Big Data Extensions Administrative User Account

Big Data Extensions administrative user is the user account **serengeti**, which has `sudo` privileges. The **serengeti** user password is the same as that of the **root** user. You can change the password by running the command `sudo /opt/serengeti/sbin/set-password -u` on the Serengeti Management Server.

You can specify a password for the **serengeti** user by running the command `passwd serengeti`. The password for the **serengeti** user can be a different password from that assigned to the **root** user.

To manage Big Data Extensions you must login to the Serengeti Management Server as the **serengeti** user. Once logged in a the **serengeti** user you can change to the **root** user account if necessary.

Support for Active Directory and OpenLDAP

Big Data Extensions supports integration with Active Directory and OpenLDAP. When configured to work with Active Directory or OpenLDAP, the Serengeti Management Server and cluster nodes can authenticate or authorize users against your Active Directory or OpenLDAP user directory.

Security Updates and Patches

You can apply security updates and patches as they are made available by either VMware, or the vendors of operating systems and Hadoop distributions.

Big Data Extensions Operating System Versions

Big Data Extensions uses the following operating systems and versions.

- The Big Data Extensions virtual appliance uses CentOS 5.11 (x86_64) and CentOS 6.6 (x86_64) as guest operating systems.
- The Serengeti Management Server uses CentOS 5.11.
- The Big Data Extensions cluster nodes use CentOS 6.6.

Applying Patches and Security Updates

You apply security patches and updates using conventional upgrade procedures. For example, using yum or rpm upgrade. See Chapter 3, "Upgrading Big Data Extensions," on page 35.

Troubleshooting 13

The troubleshooting topics provide solutions to problems that you might encounter when using Big Data Extensions.

This chapter includes the following topics:

- "MapReduce Job Fails to Run and Does Not Appear In the Job History," on page 157
- "Cannot Submit MapReduce Jobs for Compute-Only Clusters with External Isilon HDFS," on page 157
- "MapReduce Job Stops Responding on a PHD or CDH4 YARN Cluster," on page 158
- "Cannot Download the Package When Using Downloadonly Plugin," on page 158
- "Cannot Find Packages When You Use Yum Search," on page 158
- "Remove the HBase Rootdir in HDFS Before You Delete the HBase Only Cluster," on page 159

Log Files for Troubleshooting

Big Data Extensions and Serengeti create log files that provide system and status information that you can use to troubleshoot deployment and operation problems.

Table 13-1. Log Files

Category	File Name	Information	Location
Serengeti vApp boot-up log	■ `serengeti-boot.log`	Deployment time messages, which you can use to troubleshoot an unsuccessful deployment.	`/opt/serengeti/logs`
Serengeti server service log	■ `serengeti.log`	Web service component logs.	`/opt/serengeti/logs`
Serengeti server installation and configuration log	■ `ironfan.log`	Software installation and configuration information.	`/opt/serengeti/logs`
Serengeti server elastic scaling log	■ `vhm.log`	Elastic scaling logs.	`/opt/serengeti/logs`

VMware vSphere ESXi and vCenter Server Log Files

In addition to the Big Data Extensions and Serengeti log files, vSphere ESXi and vCenter Server also create log files that provide system and status information that you can use to troubleshoot deployment and operation problems.

If you encounter error messages which begin with the statement `Failed to execute vCenter Server command:`, check your vSphere ESXi and vCenter Server log files for additional troubleshooting information. There are a number of ways in which you can view log files depending on whether they are for vCenter Server or an ESXi host. See the *VMware vSphere ESXi and vCenter Server* documentation for your ESXi and vCenter Server product version.

Configure Serengeti Logging Levels

The Serengeti system and back-end tasks use Apache log4j, with the default logging level INFO, to log messages. You can configure the logging level to customize the amount and type of information shown in the system and event logs.

Enabling logging at a given level also enables logging at all higher levels.

The levels in descending order are:

- SEVERE (highest value)
- WARNING
- INFO
- CONFIG

- FINE
- FINER
- FINEST (lowest value)

In addition there is a level OFF that can be used to turn off logging, and a level ALL that can be used to enable logging of all messages.

Procedure

1. Open the `/opt/serengeti/conf/log4j.properties` file for editing.
2. Change the logging level.
3. Save your changes and close the file.
4. Stop and restart the Serengeti services.

Collect Log Files for Troubleshooting

You can collect log files from the Serengeti Management Server or from the nodes of a cluster to help you and the VMware support team with troubleshooting.

If you include a cluster name when you run the command, the following log files are collected from each node in the specified cluster.

- `/var/log/hadoop`
- `/var/log/hbase`
- `/var/log/zookeeper`
- `/var/log/gphd`
- `/opt/mapr/logs`
- `/opt/mapr/hadoop/hadoop/logs`
- `/var/chef/cache/chef-stacktrace.out`

If you do not include a cluster name when you run the command, the following log files are collected from the Serengeti Management Server.

- `/opt/serengeti/logs`
- `/opt/serengeti/conf`
- `/var/log/messages`

NOTE The log files that are collected from each node or the Serengeti Management Server are configured in the `/opt/serengeti/etc/support/cluster.files` and `/opt/serengeti/etc/support/serengeti.files` files, respectively. To change which log files are collected, edit the applicable FILES file.

Procedure

1. Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.
2. Change to the directory where you want the log files stored.
3. Run the `serengeti-support.sh` script.

   ```
   serengeti-support.sh cluster_name
   ```

Big Data Extensions collects the log files and saves them in a single tarball in the Serengeti Management Server directory from which the command was run.

Troubleshooting Cluster Creation Failures

The cluster creation process can fail for many reasons. If cluster creation fails, try to resume the process.

You can use one of these methods to resume the cluster creation process.

- If you created the cluster with the Serengeti Command-Line Interface, run the `cluster create ... --resume` command.
- If you created the cluster with the vSphere Web Client, select the cluster, right-click, and select **Resume**.

If you cannot resume the process and successfully create the cluster, see the troubleshooting topics in this section.

Bootstrap Failed 401 Unauthorized Error

When you run the `cluster create` or `cluster create ... --resume` command, the command can fail. The reason it failed is logged to the associated Serengeti server installation and configuration log file, `/opt/serengeti/logs/ironfan.log`.

Problem

The `cluster create` or `cluster create ... --resume` command fails.

- On the Command-Line Interface, an error message appears:

  ```
  Bootstrap Failed
  ```

- In the Serengeti server installation and configuration log file, `/opt/seregeti/logs/ironfan.log`, an error message appears:

  ```
  [Fri, 09 Aug 2013 01:24:01 +0000] INFO: *** Chef 11.X.X *** [Fri, 09 Aug 2013 01:24:01
  +0000] INFO: Client key /home/ubuntu/chef-repo/client.pem is not present - registering [Fri,
  09 Aug 2013 01:24:01 +0000] INFO: HTTP Request Returned 401 Unauthorized: Failed to
  authenticate. Please synchronize the clock on your client [Fri, 09 Aug 2013 01:24:01 +0000]
  FATAL: Stacktrace dumped to /var/chef/cache/chef-stacktrace.out [Fri, 09 Aug 2013 01:24:01
  +0000] FATAL: Net::HTTPServerException: 401 "Unauthorized"
  ```

Cause

This error occurs if the Serengeti Management Server and the failed virtual machine clocks are not synchronized.

Solution

From the vSphere Client, configure all ESXi hosts to synchronize their clocks with the same NTP server.

After you correct the clocks, you can run the `cluster create ... --resume` command to complete the cluster provisioning process.

Cannot Create a Cluster with the hdfs-hbase-template-spec.json File

If you use the `/opt/serengeti/conf/hdfs-hbase-template-spec.json` from the Serengeti server virtual machine to create a cluster, cluster creation fails.

Problem

The `cluster create` or `cluster create ... --resume` command fails, and the Command-Line Interface displays an error message:

```
cluster cluster_name create failed: Unrecognized field "groups" (Class
com.vmware.bdd.apitypes.ClusterCreate), not marked as ignorable at [Source:
java.io.StringReader@7563a320; line: 3, column: 13] (through reference chain:
com.vmware.bdd.apitypes.ClusterCreate["groups"])
```

Cause

The `/opt/serengeti/conf/hdfs-hbase-template-spec.json` file is for Serengeti Management Server internal use only. It is not a valid cluster specification file.

Solution

Create your own cluster specification file.

Sample cluster specification files are in the `/opt/serengeti/samples` directory.

Insufficient Storage Space

If sufficient storage resources are not available when you run the `cluster create` or `cluster create ... --resume` command, cluster creation fails.

Problem

The `cluster create` or `cluster create ... --resume` command fails, and the Command-Line Interface or Big Data Extensions plug-in interface displays the following error message:

```
cluster $CLUSTER_NAME create failed: Cannot find a host with enough storage to place base nodes
[$NODE_NAME].
Node $NODE_NAME placed on host $HOST_NAME. Node $NODE_NAME placed on host $HOST_NAME. You must
add datastores on these hosts [$HOST_NAMES] to use them with the node group [$GROUP_NAME].
```

Cause

This error occurs if sufficient datastore space is not available.

Solution

1 Review the `/opt/serengeti/logs/serengeti.log` file, and search for the phrase `cannot find host with enough`.

  This information shows the Serengeti server snapshot of the vCenter Server cluster environment immediately after the placement failure.

  You can also find information about the datastore name and its capacity. Additionally, you can find the cluster specification file that you used, and information for the nodes that have been successfully placed.

2 Review your cluster specification file.

The cluster specification file defines the cluster's datastore requirements and determines the available space on the datastore that you added to Serengeti. Use this information to determine which storage is insufficient.

For example, if there is insufficient LOCAL datastore capacity for worker nodes, you must add additional LOCAL datastores to the Serengeti server and assign them to the cluster.

Distribution Download Failure

If the server for the Hadoop distribution is down when you run the `cluster create` or `cluster create ... --resume` command, cluster creation fails.

Problem

The reason the command failed is logged.

- For tarball-deployed distributions, the following error message appears on the Command-Line Interface or the Big Data Extensions plug-in interface:

```
Unable to run command 'execute[install hadoop-1.2.1 from tarball]' on node xftest-client-0.
SSH to this node and run the command 'sudo chef-client' to view error messages.
```

- For Yum-deployed distributions, the following error message appears on the Command-Line Interface or the Big Data Extensions plug-in interface:

```
Cannot bootstrap node xfbigtop-master-0.
remote_file[/etc/yum.repos.d/bigtop2.repo] (hadoop_common::add_repo line 85) had an error:
Net::HTTPServerException: 404 "Not Found"
SSH to this node and view the log file /var/chef/cache/chef-stacktrace.out, or run the
command 'sudo chef-client' to view error messages.
```

Cause

The package server is down.

- For tarball-deployed distributions, the package server is the Serengeti Management Server.
- For Yum-deployed distributions, the package server is the source of the Yum-deployed distribution: either the official Yum repository or your local Yum server.

Solution

1 Ensure that the package is reachable.

Distribution Type	Requirement
tarball-deployed	Ensure that the `httpd` service on the Serengeti Management Server is running.
Yum-deployed	Ensure that the Yum repository file URLs are correctly configured in the manifest file.

2 Ensure that you can download the necessary file from the failed node.

Distribution Type	Necessary File
tarball-deployed	tarball
Yum-deployed	Yum repository file

Serengeti Management Server IP Address Unexpectedly Changes

The IP address of the Serengeti Management Server changes unexpectedly.

Problem

When you create a cluster after the Serengeti Management Server IP address changes, the cluster creation process fails with a bootstrap failure.

Cause

The network setting is DHCP.

Solution

Restart the Serengeti Management Server virtual machine.

After Disconnecting a Host from vCenter Server the Cluster Resume Process Fails

If you disconnect a host from vCenter Server after a failed cluster creation attempt, you cannot successfully resume the cluster creation.

Problem

If cluster creation fails, and then you disconnect the affected host from vCenter Server and try to resume the cluster creation process, it fails and you receive the following error message: `cluster` *`cluster-name`* `resume failed: Failed to create virtual machine cluster` *`cluster-name`*`.`

Cause

When you disconnect the host from vCenter Server, the host's virtual machines become unavailable. When you try to resume the cluster creation, the Serengeti Management Server cannot remove the unavailable virtual machines from the disconnected host.

Solution

1. Manually remove the affected hosts from vCenter Server.
2. Repeat the cluster create resume process.

Cluster Provisioning Stops Responding if Virtual Machines are Powered Off or Reset During Bootstrapping

When you create, configure, or resume creating or configuring a cluster, the process stops responding.

Problem

If you create, configure, or resume creating or configuring a cluster, and then power off or reset a virtual machine while it is bootstrapping, the cluster provisioning process stops responding.

Cause

When a virtual machine is powered off or reset during bootstrapping, its SSH connection stops responding.

Solution

1. Do one of the following:
   - If you are using the Serengeti Command-Line Interface, press Ctrl+C.
   - If you are using the vSphere Web Client, no action is required.

2 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

3 Kill the failed cluster provisioning process.

```
ps ax | grep knife | grep cluster-name | head -1 | awk '{print $1}' | xargs kill -9
```

4 Force the cluster's status to PROVISION_ERROR.

```
set-cluster-status.sh cluster-name PROVISION_ERROR
```

5 Use the vSphere Web Client to log in to vCenter Server.

6 Power on any virtual machines in the cluster that are powered off.

7 Resume the cluster creation process.

- If you created the cluster with the Serengeti Command-Line Interface, run the `cluster create ... --resume` command.
- If you created the cluster with the vSphere Web Client, select the cluster, right-click, and select **Resume**.

HBase Cluster Creation Job Fails When Time Difference Among Nodes is More Than 20 Seconds

If the time difference among nodes is more than 20 seconds, you must synchronize the times before you can create an HBase cluster or run jobs.

Problem

If you attempt to create an HBase cluster with nodes whose times are more than 20 seconds apart, the cluster creation might fail. If it succeeds, any HBase jobs that you run will fail.

Cause

HBase requires that the time difference between its master-server and region-server nodes be 20 seconds or less.

Solution

1 Make sure that the NTP server is running on all ESXi hosts and that the time difference among all ESXi hosts is less than 20 seconds.

Wait a few minutes to let the nodes synchronize their time with their ESXi hosts.

2 Make sure that the time difference among nodes in the cluster is less than 20 seconds.

a Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

b Run the `serengeti-ssh.sh` script.

```
serengeti-ssh.sh hbase_cluster_name date
```

c If the times are more than 20 seconds apart, repeat steps 1 and 2.

3 Start the failed process or services.

- If the original cluster creation failed, try to resume the cluster creation process.
  - If you created the cluster with the Serengeti Command-Line Interface, run the `cluster create ... --resume` command.

  - If you created the cluster with the vSphere Web Client, select the cluster, right-click, and select **Resume**.
- If the cluster resume process failed, try again to resume it.
- If the cluster creation succeeded but running a job failed, start the failed services.
  - If you are using the Serengeti Command-Line Interface, run the following commands.

    ```
    cluster export --name cluster_name --specFile /tmp/1
    cluster config --name cluster_name --specFile /tmp/1 --yes
    ```

  - If you are using the vSphere Web Client, stop and restart the cluster.

Creating a Large Scale Cluster in Big Data Extensions Results In a Bootstrap Failed Error

When you create a large scale cluster, for example, of 300 or more nodes per cluster, in Big Data Extensions, you might get a `Bootstrap failed` error.

Problem

Generally, one database connection can serve two nodes at the same time, so for a cluster with 300 or more nodes, 150 database connections are required. To avoid receiving a `Bootstrap failed` error, increase the size of the database connection pool.

Cause

The size of the database connection pool was not large enough to handle creating a large scale cluster with 300 or more nodes.

Solution

1 After the Big Data Extensions vApp is deployed, log in to the Big Data Extensions Management Server as user serengeti.

2 Increase the database connection pool size.

Option	Description
/etc/chef-server/chef-server.rb	Indicates the location on the Management Server to configure the database connection pool size.
postgresql['max_connections']	Indicates the maximum number of connections for the postgresql database. This value should usually be `erchef['db_pool_size'] + 100`.
erchef['db_pool_size']	Indicates the database connection pool size.

```
sudo sed -i -e "s|erchef\['db_pool_size'\] .*|erchef['db_pool_size'] = 150|"
/etc/chef-server/chef-server.rb
sudo sed -i -e "s|postgresql\['max_connections'\] .*|postgresql['max_connections'] = 250|"
/etc/chef-server/chef-server.rb
sudo chef-server-ctl reconfigure
```

Cannot Create a Cluster for Which the Time Is Not Synchronized

When you run the `cluster create` or `cluster create ... --resume` command, the command can fail if there are time discrepancies in the environment.

Problem

The `cluster create` or `cluster create ... --resume` command fails, and the Command-Line Interface or Big Data Extensions plug-in interface displays the following error message:

```
You must synchronize the time of the following hosts [$HOST_NAMES] with the Serengeti Management
Server to use them.
```

Cause

Before creating new virtual machines on hosts, the time on the target hosts is checked against the time on the Serengeti Management Server. If the time between the Serengeti Management Server and the hosts is not synchronized, the virtual machine creation will fail.

Solution

◆ From the vSphere Web Client, configure all ESXi hosts to synchronize their clocks with the same NTP server.

What to do next

After you synchronize the time between the Serengeti Management Server and the other ESXi hosts within your environment, try to create a cluster.

Big Data Extensions Virtual Appliance Upgrade Fails

The upgrade of the Big Data Extensions virtual appliance might fail. If the upgrade process fails, you can try the upgrade again.

Problem

The upgrade of the Big Data Extensions virtual appliance does not succeed.

Solution

1. Revert to the prior state for both of the virtual machines for the Big Data Extensions virtual appliance based on the snapshots that vSphere Update Manager took.

   Use the virtual machine's snapshot manager and select the snapshot created by vSphere Update Manager.

2. Reboot the virtual appliance.
3. Resolve the blocking issue.
4. Restart the remediation task.

   Click **Remediate** in the vSphere Update Manager user interface to redo the upgrade process.

Upgrade Cluster Error When Using Cluster Created in Earlier Version of Big Data Extensions

To enable the Serengeti Management Server to manage clusters that you created in a previous version of Big Data Extensions, you must upgrade the components in each cluster's virtual machines. The Serengeti Management Server uses these components to control the cluster nodes.

Problem

When you upgrade from an earlier version of Big Data Extensions, clusters that you need to upgrade are shown with an alert icon next to the cluster name. When you click the alert icon the error message "Upgrade the cluster to the latest version" displays as a tool tip. See "View Provisioned Clusters in the vSphere Web Client," on page 125.

You can also identify clusters you need to upgrade using the `cluster list` command. When you run the `cluster list` command the message "Need Upgrade" displays where the cluster version normally appears.

Solution

1. For each cluster that you created in a previous version of Big Data Extensions, make sure that all of the cluster's nodes are powered on and have valid IP addresses.

   If a node does not have a valid IP address, it cannot be upgraded to the new version of Big Data Extensions virtual machine tools.

   a. Log into the vSphere Web Client connected to vCenter Server and navigate to **Hosts and Clusters**.

   b. Select the cluster's resource pool, select the **Virtual Machines** tab, and power on the cluster's virtual machines.

   IMPORTANT It might take up to five minutes for vCenter Server to assign valid IP addresses to the big data cluster nodes. Do not perform the remaining upgrade steps until the nodes have received their IP addresses.

2. Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

3. Run the `cluster upgrade` command for each cluster created in a previous version of Big Data Extensions.

   ```
   cluster upgrade --name cluster-name
   ```

4. If the upgrade fails for a node, make sure that the failed node has a valid IP address, and then rerun the `cluster upgrade` command.

   You can rerun the command as many times as you need to upgrade all the nodes.

5. Stop and restart your Hadoop and HBase clusters.

Virtual Update Manager Does Not Upgrade the Hadoop Template Virtual Machine Under Big Data Extensions vApp

When you use the Virtual Update Manager (VUM) to upgrade the Big Data Extensions 2.0 vApp to Big Data Extensions 2.1 vApp, VUM reports that the upgrade completed successfully, but VUM only upgrades the Big Data Extensions Management Server, not the Big Data Extensions hadoop template.

Cause

The hadoop template virtual machine in Big Data Extensions 2.1 vApp contains CentOS 6.5 and the hadoop template virtual machine in Big Data Extensions 2.0 vApp contains CentOS 6.4. The current versions of Virtual Update Manager (versions 5.1 and 5.5) do not support upgrading from CentOS 6.4 to CentOS 6.5.

Solution

To work around this issue, you must manually replace the hadoop template virtual machine after you upgrade the Big Data Extensions vApp.

Procedure

1 Download the Big Data Extensions template OVA file from the Big Data Extensions download page.

2 Login to the vSphere Client.

3 Delete the original hadoop-template virtual machine under the Big Data Extensions vApp or move it out of the Big Data Extensions vApp which has been upgraded.

4 Select **File > Deploy OVA Template**.

5 Enter the downloaded template OVA path in the **Deploy OVA Template** dialog.

6 Complete the steps in the OVA Deployment wizard.

   On the step to configure the name and location, enter **`hadoop-template`** or any other valid name. On the step to configure the resource pool, select the Big Data Extensions vApp which has been upgraded.

7 Login to the Big Data Extensions server via SSH.

8 Restart Big Data Extensions Web services: `sudo service tomcat restart`

Unable to Connect the Big Data Extensions Plug-In to the Serengeti Server

When you install Big Data Extensions on vSphere 5.1 or later, the connection to the Serengeti Management Server fails to authenticate.

Problem

The Big Data Extensions plug-in is unable to connect to the Serengeti server.

Cause

During the deployment, the Single Sign-On (SSO) link was not entered. The Serengeti Management Server cannot authenticate the connection from the plug-in.

Solution

Use the Serengeti Management Server Administration Portal to configure the SSO settings. See "Configure vCenter Single Sign-On Settings for the Serengeti Management Server," on page 30.

If you still cannot connect the Big Data Extensions plug-in to the Serengeti Server, use the `EnableSSOAuth` utility.

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `root`.

2 Configure the SSO settings.

```
EnableSSOAuth https://vCenter-server-IP-address:7444/lookupservice/sdk
```

3 Restart the tomcat service.

```
/sbin/service tomcat restart
```

4 Connect the Big Data Extensions plug-in to the Serengeti Management Server.

vCenter Server Connections Fail to Log In

The Serengeti Management Server tries but fails to connect to vCenter Server.

Problem

The Serengeti Management Server tries and fails to connect to vCenter Server.

Cause

vCenter Server is unreachable for any reason, such as network issues or too many running tasks.

Solution

Ensure that vCenter Server is reachable.

- Connect to vCenter Server with the vSphere Web Client or the VMware Infrastructure Client (VI Client) .
- Ping the vCenter Server IP address to verify that the Serengeti Management Server is connecting to the correct IP address.

Management Server Cannot Connect to vCenter Server

If you enable an IPv6 connection with vCenter Server and then change the IP address, the management server cannot connect to vCenter Server. You cannot fix the problem by restarting the management server.

Solution

1. Use the vSphere Web Client to log in to vCenter Server.
2. Power off the Management Server.
3. Navigate to the Management Server Network 2 Settings section.
4. Under vApp Options, select **Edit Settings > Options > Properties**.
5. Enter the new IPv6 address for vCenter Server to item in the **vCenter's IPv6 Address to connect** text box.
6. Power on the Management Server.

Cannot Perform Serengeti Operations after Deploying Big Data Extensions

If the Big Data Extensions plug-in cannot communicate with vCenter Server during initialization, internal errors occur and you cannot perform Serengeti operations.

Problem

When you run a command, the Serengeti CLI or Big Data Extensions plug-in displays the following error.

```
Internal error: REST API transport layer error.
```

Cause

The Serengeti Management Server must communicate with the vCenter Extension vService through the vCenter Server FQDN during initialization. The following scenarios can prevent you from performing Serengeti operations.

- (vSphere 5.1) If the host name is not correctly set during vSphere installation, vCenter Server cannot receive the correct host name or FQDN, nor provide the correct information to the Serengeti Management Server. As a result, the Big Data Extensions plug-in cannot successfully communicate with vCenter Server.

- (vSphere 5.5) If the vCenter Server Appliance is deployed into an OVF environment that has a static IP network configuration and a blank host name, the reverse lookup from the IP address cannot be performed. Without the reverse lookup, the host name is incorrectly set for the vCenter Server Appliance.
- If vCenter Server is configured with an FQDN, but the Big Data Extensions vApp instance's DNS server cannot communicate with vCenter Server or cannot resolve the vCenter Server FQDN, the host name cannot be set.

Solution

For vSphere 5.1, set the host name.

1. Use the vSphere Web Client to log in to vCenter Server.
2. From the Inventory pane, click **Big Data Extensions**, and click the **Summary** tab.
3. In the Connected Server dialog, click the name of the connected vCenter Server.
4. Click the **Manage** tab, click **Settings**, and from the tab's left navigation pane, click **Advanced Settings**.
5. Find the keys related to the FQDN, and change their values from the incorrect FQDN to the correct FQDN or IP address.
6. Delete the existing Big Data Extensions vApp instance and install a new Big Data Extensions vApp instance.

For vSphere 5.5, reinstall your environment.

1. Delete and reinstall the vCenter Server Appliance. During reinstallation, set the vCenter Server Appliance host name in the OVF properties.
2. Delete the existing Big Data Extensions vApp instance and install a new Big Data Extensions vApp instance.

For DNS server or network problems, delete the existing Big Data Extensions vApp instance and install a new Big Data Extensions vApp instance that includes network connectivity and a DNS server that can resolve the vCenter Server FQDN.

SSL Certificate Error When Connecting to Non-Serengeti Server with the vSphere Console

From the vSphere Web Client, you cannot connect to a non-Serengeti server.

Problem

When you use the Big Data Extensions plug-in to vCenter Server and try to connect to a non-Serengeti server, you receive an error message:

```
SSL error:
Check certificate failed.
Please select a correct serengeti server.
```

Cause

When you use the Big Data Extensions plug-in, you can connect only to Serengeti servers.

Solution

Connect only to Serengeti servers. Do not perform certificate-related operations.

Cannot Restart or Reconfigure a Cluster For Which the Time Is Not Synchronized

When the time on the hosts and the Serengeti Management Server drifts apart, the cluster cannot be restarted or reconfigured.

Problem

The cluster fails to start, and the Command-Line Interface or Big Data Extensions plug-in interface displays the following error message:

```
Nodes in cluster $CLUSTER_NAME start failure: Synchronize the time of the host [$HOST_NAME(S)]
with the Serengeti Management Server running on $HOST_NAME.
```

Cause

This error occurs if the Serengeti Management Server and the failed virtual machine clocks are not synchronized. The time on all hosts within a cluster is checked against the time on the Serengeti Management Server. If the time between the Serengeti Management Server and the hosts is not synchronized, the virtual machine fails to start.

Solution

◆ From the vSphere Web Client, configure all ESXi hosts to synchronize their clocks with the same NTP server.

After you correct the clocks, you can try to start or reconfigure the cluster.

Cannot Restart or Reconfigure a Cluster After Changing Its Distribution

After you change a cluster's distribution vendor or distribution version, but not the distribution name, the cluster cannot be restarted or reconfigured.

Problem

When you try to restart or reconfigure a cluster after changing its distribution vendor or distribution version in the manifest, you receive the following error message:

```
Bootstrap Failed
```

Cause

When you manually change an existing distribution's vendor or version in the manifest file and reuse a distribution's name, the Serengeti server cannot start the node.

Solution

1. Revert the manifest file.
2. Use the `config-distro.rb` tool to add a new distribution, with a unique name, for the distribution vendor and version that you want.

Virtual Machine Cannot Get IP Address and Command Fails

A Serengeti command fails, and the CLI displays the following error message: `Virtual Machine Cannot Get IP Address`.

Cause

This error occurs when a network configuration error occurs.

For static IP, the cause is typically an IP address conflict.

For DHCP, common causes include:

- The number of virtual machines that require IPs exceeds the available DHCP addresses.
- The DHCP server fails to allocate sufficient addresses.
- The DHCP renew process failed after an IP address expires.

Solution

- Verify that the vSphere port group has enough available ports for the new virtual machine.
- If the network is using static IP addressing, ensure that the IP address range is not used by another virtual machine.
- If the network is using DHCP addressing, ensure that an IP address is available to allocate for the new virtual machine.

Cannot Change the Serengeti Server IP Address From the vSphere Web Client

When you attempt to change the Serengeti server IP address from the vSphere Web Client, the procedure fails.

Solution

Prerequisites

Get a static IP address.

Procedure

1 On the Serengeti Management Server, edit the following configuration file /etc/sysconfig/network-scripts/ifcfg-eth0 by replacing the contents of the file with following contents:

```
DEVICE=eth0
BOOTPROTO=static
ONBOOT=yes
TYPE=Ethernet
IPADDR=your_static_ip
PREFIX=your_prefix
GATEWAY=your_gateway
DNS1=your_dns1
DNS2=your_dns2
```

2 Reboot the Serengeti Management Server.

When the operating system starts, it configures the IP address according to the contents of the new configuration file.

A New Plug-In Instance with the Same or Earlier Version Number as a Previous Plug-In Instance Does Not Load

When you install a new Big Data Extensions plug-in instance that has the same or earlier version as a previous Big Data Extensions plug-in instance, the previous version is loaded instead of the new version.

Problem

When you install a new Big Data Extensions plug-in that has the same or lower version number as a previous Big Data Extensions plug-in, the previous version is loaded instead of the new version. This happens regardless of whether you uninstall the previous plug-in.

Cause

When you uninstall a plug-in instance, the vSphere Web client does not remove the plug-in instance package from the Serengeti server.

After you install a plug-in instance with the same or earlier version number as the previous plug-in instance, and try to load the new plug-in instance, vSphere finds the previous plug-in instance package in its local directory. vSphere does not download the new plug-in instance package from the remote Serengeti server.

Solution

1 Uninstall the old plug-in instance.

2 Remove the old plug-in instance.

- For vCenter Server Appliance instances, delete the `/var/lib/vsphere-client/vc-packages/vsphere-client-serenity/vsphere-bigdataextensions-`*`version`* folder.

- For vSphere Web client servers on Windows, delete the `%ProgramData%/vmware/vSphere Web Client/vc-packages/vsphere-client-serenity/vsphere-bigdataextensions-`*`version`* folder.

3 Restart the vSphere Web Client.

- For vCenter Server Appliance instances, restart the vSphere Web Client service at the vCenter Server Appliance Web console, `http://$`*`vCenter-Server-Appliance-IP`*`:5480`

- For vSphere Web Client servers on Windows, restart the vSphere Web Client service from the services console.

4 Install the new plug-in instance.

Host Name and FQDN Do Not Match for Serengeti Management Server

The Serengeti Management Server host name and Fully Qualified Domain Name (FQDN) must match before you can perform some Big Data Extensions operations, such as an upgrade.

Problem

The Serengeti Management Server's host name and FQDN are not the same.

Cause

Certain sequences of deployment steps can cause the Serengeti Management Server's host name and FQDN to be different.

Solution

1 Open a command shell, such as Bash or PuTTY, and log in to the Serengeti Management Server as user `serengeti`.

2 Create a new file for the `set_hostname.sh` script.

```
touch /tmp/set_hostname.sh
```

3 Open the `/tmp/set_hostname.sh` file for editing, and add the following lines.

```
ETHIP=`/sbin/ifconfig eth0 | grep "inet addr" | awk '{print $2}' | sed 's/addr://'`
FQDN=$ETHIP
RET=`/bin/ipcalc --silent --hostname $ETHIP`
if [ "$?" = "0" ]; then
  FQDN=`echo $RET | awk -F= '{print $2}'`
fi
echo "set hostname to ${FQDN}"
`hostname ${FQDN}`
```

4 Save your changes and close the file.

5 Run the `set_hostname.sh` script.

```
sudo bash /tmp/set_hostname.sh
```

Serengeti Operations Fail After You Rename a Resource in vSphere

After you use vSphere to rename a resource, Serengeti commands fail for all Serengeti clusters that use that resource.

Problem

If you use vSphere to rename a Serengeti resource that is used by provisioned Serengeti clusters, Serengeti operations fail for the clusters that use that resource. This problem occurs for vCenter Server resource pools, datastores, and networks that you add to Serengeti, and their related hosts, vCenter Server clusters, and so on. The error message depends on the type of resource, but generally indicates that the resource is inaccessible.

Cause

The Serengeti resource mapping requires that resource names do not change.

Solution

Use vSphere to revert the resource to its original name.

Big Data Extensions Server Does Not Accept Resource Names With Two or More Contiguous White Spaces

If you include two or more contiguous white space characters in the name for a Big Data Extensions resource pool, datastore, or network, the add process fails.

Solution

No workarounds or patches are available for this issue.

Non-ASCII characters are not displayed correctly

When you work with the CLI on a Windows platform, if you run a script command on a file that contains non-ASCII characters, it returns messages that are not displayed correctly.

Cause

It is a known issue that non-ASCII characters are not recognized on Windows platforms.

Solution

No workarounds or patches are available for this issue

MapReduce Job Fails to Run and Does Not Appear In the Job History

A submitted MapReduce job fails to run and does not appear in the job history.

Problem

When you submit a MapReduce job and the workload is heavy, the MapReduce job does not run, and it does not appear in the MapReduce job history.

Cause

During heavy workloads, the JobTracker or NameNode service might be too busy to respond to vSphere HA monitoring within the configured timeout value. When a service does not respond to vSphere HA request, vSphere restarts the affected service.

Solution

1 Stop the HMonitor service.

   When you stop the HMonitor service, vSphere HA failover is disabled.

   a Open a command shell, such as Bash or PuTTY, and log in to the affected cluster node.

   b Stop the HMonitor service.

   ```
   sudo /etc/init.d/hmonitor-*-monitor stop
   ```

2 Increase the JobTracker vSphere timeout value.

   a Open the `/user/lib/hadoop/monitor/vm-jobtracker.xml` file for editing.

   b Find the `service.monitor.probe.connect.timeout` property.

   c Change the value of the `<value>` element.

   d Save your changes and close the file.

3 Increase the NameNode vSphere timeout value.

   a Open the `/user/lib/hadoop/monitor/vm-namenode.xml` file for editing.

   b Find the `service.monitor.portprobe.connect.timeout` property.

   c Change the value of the `<value>` element.

   d Save your changes and close the file.

4 Start the HMonitor service.

   ```
   sudo /etc/init.d/hmonitor-*-monitor start
   ```

Cannot Submit MapReduce Jobs for Compute-Only Clusters with External Isilon HDFS

You cannot submit MapReduce Jobs for compute-only clusters that point to an external Isilon HDFS.

Problem

If you deploy a compute-only cluster with an external HDFS pointing to Isilon, the deployment appears to be successful. However, the JobTracker is in safe mode, which does not let you submit MapReduce jobs.

Cause

JobTracker requires a user named `mapred`.

Solution

1 SSH into the Isilon cluster.

2 Add the `mapred` user to the Isilon system's wheel group.

```
pw useradd mapred –G wheel
```

MapReduce Job Stops Responding on a PHD or CDH4 YARN Cluster

A MapReduce job stops responding on a PHD or CDH4 YARN cluster with one DataNode and one NodeManager agent, each with 378MB of memory.

Problem

MapReduce jobs stop responding when you run them on a PHD or CDH4 YARN cluster with one data node and one NodeManager agent.

Cause

Insufficient memory resources.

Solution

1 Create a PHD or CDH4 YARN cluster with two DataNodes and two NodeManagers.

2 Rerun the MapReduce job.

Cannot Download the Package When Using Downloadonly Plugin

When you try to set up a local yum repository, you might find that, when you use the downloadonly plugin, you can find the package that you need but you cannot download the package.

Solution

1 Perform the following command to check if the package has been installed on the machine:

```
yum remove <package_name>
```

2 If the package has been installed on the machine, remove the package and try to download it again.

Cannot Find Packages When You Use Yum Search

When you try to set up a local yum repository, you must download packages for either the Cloudera Manager or Ambari application manager. The packages have been put on the http server and can display in a browser but, when you use yum search, you cannot find the specific packages that you need.

Cause

If the repo file is not set correctly or you have data in the yum cache on your system, you might encounter this issue.

Solution

1 Make sure the yum repository server URL in the repo file is correct for the location and version.

2 Use the createrepo tool to make sure that you have created the repodata directory.

3 Use the `yum clean all` command to clean the yum cache.

4 Run the yum search again to locate the packages.

Remove the HBase Rootdir in HDFS Before You Delete the HBase Only Cluster

After you delete an HBase only cluster, the HBase data still exists on the external HDFS. It is important that you remove the HBase rootdir in HDFS before you delete the HBase only cluster.

Cause

The HBase rootdir was not removed before the HBase only cluster was deleted.

Solution

You can keep the data or remove the data.

Procedure

1. Log in to the HBase master node in the HBase only cluster.
2. Open the `hbase-site.xml` file and find the value for the property `hbase.rootdir`.

   ```
   /etc/hbase/conf/hbase-site.xml
   ```

3. Run the following command:

   ```
   hadoop fs -rmr <value_of_hbase.rootdir>
   ```

4. Delete the HBase only cluster in Big Data Extensions.

Index