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
Contents

About This Book 7

Updated Information 9

1 About VMware vSphere Big Data Extensions 11
   Getting Started with Big Data Extensions 11
   Big Data Extensions and Project Serengeti 13
   About Big Data Extensions Architecture 14
   Big Data Extensions Support for Hadoop Features By Distribution 14
   Hadoop Feature Support By Distribution 16

2 Installing Big Data Extensions 19
   System Requirements for Big Data Extensions 20
   Internationalization and Localization 22
   Deploy the Big Data Extensions vApp in the vSphere Web Client 23
   Install RPMs in the Serengeti Management Server Yum Repository 26
   Install the Big Data Extensions Plug-In 27
   Connect to a Serengeti Management Server 28
   Install the Serengeti Remote Command-Line Interface Client 29
   Access the Serengeti Command-Line Interface Using the Remote Command-Line Interface Client 30

3 Upgrading Big Data Extensions 33
   Prepare to Upgrade Big Data Extensions 34
   Upgrade Big Data Extensions Virtual Appliance 35
   Upgrade the Big Data Extensions Plug-in 38
   Upgrade the Serengeti Command-Line Interface 38
   Upgrade the CentOS 6.x Template 39
   Upgrade Big Data Extensions Virtual Machine Components 40

4 Managing Hadoop Distributions 43
   Hadoop Distribution Deployment Types 43
   Configure a Tarball-Deployed Hadoop Distribution 44
   Configuring Yum and Yum Repositories 46
   Create a Hadoop Template Virtual Machine using RHEL Server 6.x and VMware Tools 57
   Maintain a Customized Hadoop Template Virtual Machine 60

5 Managing the Big Data Extensions Environment 63
   Add Specific User Names to Connect to the Serengeti Management Server 64
   Change the Password for the Serengeti Management Server 64
   Configure vCenter Single Sign-On Settings for the Serengeti Management Server 65
   Create a User Name and Password for the Serengeti Command-Line Interface 66
Stop and Start Serengeti Services  66

6 Managing vSphere Resources for Hadoop and HBase Clusters  69
   Add a Resource Pool with the Serengeti Command-Line Interface  70
   Remove a Resource Pool with the Serengeti Command-Line Interface  70
   Add a Datastore in the vSphere Web Client  70
   Remove a Datastore in the vSphere Web Client  71
   Add a Network in the vSphere Web Client  72
   Reconfigure a Static IP Network in the vSphere Web Client  72
   Remove a Network in the vSphere Web Client  73

7 Creating Hadoop and HBase Clusters  75
   About Hadoop and HBase Cluster Deployment Types  76
   About Cluster Topology  77
   About HBase Database Access  77
   Create a Hadoop or HBase Cluster in the vSphere Web Client  78
   Create a Cluster with Topology Awareness with the Serengeti Command-Line Interface  80

8 Managing Hadoop and HBase Clusters  83
   Stop and Start a Hadoop Cluster in the vSphere Web Client  84
   Scale Out a Hadoop Cluster in the vSphere Web Client  84
   Scale CPU and RAM in the vSphere Web Client  85
   Reconfigure a Hadoop or HBase Cluster with the Serengeti Command-Line Interface  86
   Delete a Hadoop Cluster in the vSphere Web Client  88
   About Resource Usage and Elastic Scaling  88
   Use Disk I/O Shares to Prioritize Cluster Virtual Machines in the vSphere Web Client  92
   About vSphere High Availability and vSphere Fault Tolerance  93
   Recover from Disk Failure with the Serengeti Command-Line Interface Client  93
   Log in to Hadoop Nodes with the Serengeti Command-Line Interface Client  94
   Change the User Password on All of a Cluster’s Nodes  94

9 Monitoring the Big Data Extensions Environment  97
   View Serengeti Management Server Initialization Status  97
   View Provisioned Clusters in the vSphere Web Client  98
   View Cluster Information in the vSphere Web Client  99
   Monitor the Hadoop Distributed File System Status in the vSphere Web Client  100
   Monitor MapReduce Status in the vSphere Web Client  101
   Monitor HBase Status in the vSphere Web Client  101

10 Using Hadoop Clusters from the Serengeti Command-Line Interface  103
    Run HDFS Commands with the Serengeti Command-Line Interface  103
    Run MapReduce Jobs with the Serengeti Command-Line Interface  104
    Run Pig and PigLatin Scripts with the Serengeti Command-Line Interface  104
    Run Hive and Hive Query Language Scripts with the Serengeti Command-Line Interface  105

11 Accessing Hive Data with JDBC or ODBC  107
    Configure Hive to Work with JDBC  107
    Configure Hive to Work with ODBC  109
12 Troubleshooting 111
   Log Files for Troubleshooting 112
   Configure Serengeti Logging Levels 113
   Collect Log Files for Troubleshooting 113
   Big Data Extensions Virtual Appliance Upgrade Fails 114
   Troubleshooting Cluster Creation Failures 114
   Cannot Restart or Reconfigure a Cluster After Changing Its Distribution 121
   Cannot Restart or Reconfigure a Cluster Whose Time Is Not Synchronized 121
   Virtual Machine Cannot Get IP Address 122
   vCenter Server Connections Fail to Log In 122
   SSL Certificate Error When Connecting to Non-Serengeti Server with the vSphere Console 123
   Serengeti Operations Fail After You Rename a Resource in vSphere 123
   A New Plug-In Instance with the Same or Earlier Version Number as a Previous Plug-In Instance Does Not Load 123
   MapReduce Job Fails to Run and Does Not Appear in the Job History 124
   Cannot Submit MapReduce Jobs for Compute-Only Clusters with External Isilon HDFS 125
   MapReduce Job Stops Responding on a PHD or CDH4 YARN Cluster 125
   Unable to Connect the Big Data Extensions Plug-In to the Serengeti Server 125
   Cannot Perform Serengeti Operations after Deploying Big Data Extensions 126
   Host Name and FQDN Do Not Match for Serengeti Management Server 127
   Upgrade Cluster Error When Using Cluster Created in Earlier Version of Big Data Extensions 128

Index 129
**About This Book**

*VMware vSphere Big Data Extensions Administrator’s and User’s Guide* describes how to install 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 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](http://www.vmware.com/support/pubs).
Updated Information

This VMware vSphere Big Data Extensions Administrator’s and User’s Guide is updated with each release of the product or when necessary.

This table provides the update history of the VMware vSphere Big Data Extensions Administrator’s and User’s Guide.

<table>
<thead>
<tr>
<th>Revision</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-001514-02</td>
<td>Enhanced the description of how to access the upgrade source in the topic “Download the Upgrade Source and Accept the License Agreement,” on page 36.</td>
</tr>
<tr>
<td>EN-001514-01</td>
<td>Expanded the information on preparing to upgrade and upgrading the Big Data Extensions vApp in Chapter 3, “Upgrading Big Data Extensions,” on page 33.</td>
</tr>
<tr>
<td>EN-001514-00</td>
<td>Initial release.</td>
</tr>
</tbody>
</table>
About VMware vSphere Big Data Extensions

VMware® vSphere™ Big Data Extensions lets you deploy and centrally operate Hadoop and HBase 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 Hadoop and HBase cluster, and incorporates a full range of tools to help you optimize cluster performance and utilization.

- **Getting Started with Big Data Extensions** on page 11
  VMware vSphere Big Data Extensions lets you deploy Hadoop and HBase clusters. The tasks in this section describe how to set up vSphere for use with Big Data Extensions, deploy the Big Data Extensions vApp, access the vCenter Server and command-line interface (CLI) administrative consoles, and configure a Hadoop distribution for use with Big Data Extensions.

- **Big Data Extensions and Project Serengeti** on page 13
  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.

- **About Big Data Extensions Architecture** on page 14
  The Serengeti Management Server and Hadoop Template virtual machine work together to configure and provision Hadoop and HBase clusters.

- **Big Data Extensions Support for Hadoop Features By Distribution** on page 14
  Big Data Extensions provides different levels of feature support depending on the Hadoop distribution and version that you use.

- **Hadoop Feature Support By Distribution** on page 16
  Each Hadoop distribution and version provides differing feature support. Learn which Hadoop distributions support which features.

### Getting Started with Big Data Extensions

VMware vSphere Big Data Extensions lets you deploy Hadoop and HBase clusters. The tasks in this section describe how to set up vSphere for use with Big Data Extensions, deploy the Big Data Extensions vApp, access the 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. See “Big Data Extensions and Project Serengeti,” on page 13.
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. See “Big Data Extensions Support for Hadoop Features By Distribution,” on page 14.

Understand which features are supported by your Hadoop distribution. See “Hadoop Feature Support By Distribution,” on page 16.

**Procedure**

1. Do one of the following.

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

   Apache Hadoop is included in the Serengeti Management Server, but you can use any Hadoop distribution that Big Data Extensions supports. See Chapter 4, “Managing Hadoop Distributions,” on page 43.

**What to do next**

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

- Stop and start the Serengeti services, create user accounts, manage passwords, and log in to cluster nodes to perform troubleshooting. See Chapter 5, “Managing the Big Data Extensions Environment,” on page 63.

- Manage the vSphere resource pools, datastores, and networks that you use to create Hadoop and HBase clusters. See Chapter 6, “Managing vSphere Resources for Hadoop and HBase Clusters,” on page 69.

- Create, provision, and manage Hadoop and HBase clusters. See Chapter 7, “Creating Hadoop and HBase Clusters,” on page 75 and Chapter 8, “Managing Hadoop and HBase Clusters,” on page 83.

- 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. See Chapter 9, “Monitoring the Big Data Extensions Environment,” on page 97.

- On your Big Data clusters, run HDFS commands, Hive and Pig scripts, and MapReduce jobs, and access Hive data. See Chapter 10, “Using Hadoop Clusters from the Serengeti Command-Line Interface,” on page 103.

- If you encounter any problems when using Big Data Extensions, see Chapter 12, “Troubleshooting,” on page 111.
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 Hadoop and 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.

**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.
- Hadoop distribution from the Apache community.

**NOTE** VMware provides the Hadoop distribution as a convenience but does not provide enterprise-level support. The Apache Hadoop 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 Hadoop and HBase clusters.

Big Data Extensions performs the following steps to deploy a Hadoop or HBase 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 Hadoop or HBase 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 Hadoop virtual machines, at which point you have a running cluster based on your configuration settings.

Big Data Extensions Support for Hadoop Features By Distribution

Big Data Extensions provides different levels of feature support depending on the Hadoop distribution and version that you use.

Support for Hadoop MapReduce v1 Distribution Features

Big Data Extensions provides differing levels of feature support depending on the Hadoop distribution and version that you configure for use. Table 1-1 lists the supported Hadoop MapReduce v1 distributions and indicates which features are supported when using the distribution with Big Data Extensions.

| Table 1-1. Big Data Extensions Feature Support for Hadoop MapReduce v1 Distributions |
|-------------------------------------------------|-----------------|-----------------|-----------------|------------------|------------------|
| Version                                         | Apache Hadoop   | Cloudera CDH4, CDH5 | Hortonworks HDP 1.3 | Intel 2.5.1, 3.0.2 | MapR 3.0.2-3.1.0 |
| Automatic Deployment                            | Yes             | Yes             | Yes             | Yes              | Yes              |
| Scale Out                                       | Yes             | Yes             | Yes             | Yes              | Yes              |
| Create Cluster with Multiple Networks           | Yes             | Yes             | Yes             | Yes              | No               |
| Data-Compute Separation                         | Yes             | Yes             | Yes             | Yes              | Yes              |
| Compute-only                                    | Yes             | Yes             | Yes             | Yes              | No               |
| Elastic Scaling of Compute Nodes                | Yes             | Yes             | Yes when using MapReduce v1 | Yes | No |
| Hadoop Configuration                            | Yes             | Yes             | Yes             | Yes              | No               |
Table 1-1. Big Data Extensions Feature Support for Hadoop MapReduce v1 Distributions (Continued)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Apache Hadoop</th>
<th>Cloudera</th>
<th>Hortonworks</th>
<th>Intel</th>
<th>MapR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadoop Topology Configuration</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Run Hadoop Commands from the CLI</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Hadoop Virtualization Extensions (HVE)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>vSphere HA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Service Level vSphere HA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Hadoop Topology Configuration</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Support for Hadoop MapReduce v2 (YARN) Distribution Features

Big Data Extensions provides differing levels of feature support depending on the Hadoop distribution and version that you configure for use. Table 1-2 lists the supported Hadoop MapReduce v2 distributions and indicates which features are supported when using the distribution with Big Data Extensions.

Table 1-2. Big Data Extensions Feature Support for Hadoop MapReduce v2 (YARN) Distributions

<table>
<thead>
<tr>
<th>Feature</th>
<th>Apache Bigtop</th>
<th>Apache Hadoop</th>
<th>Cloudera</th>
<th>Cloudera</th>
<th>Hortonworks</th>
<th>Pivotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version</td>
<td>0.7.0</td>
<td>2.0</td>
<td>CDH4</td>
<td>CDH5</td>
<td>HDP 2.0, 2.1</td>
<td>PHD 1.1, 2.0</td>
</tr>
<tr>
<td>Automatic Deployment</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Scale Out</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Create Cluster with Multiple Networks</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Data-Compute Separation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Compute-only</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Elastic Scaling of Compute Nodes</td>
<td>Yes</td>
<td>Yes</td>
<td>No when using MapReduce 2</td>
<td>No when using MapReduce 2</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Hadoop Configuration</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Hadoop Topology Configuration</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Hadoop Feature Support By Distribution</td>
<td>Distribution</td>
<td>Feature Support</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>--------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Apache Hadoop</td>
<td>Apache Bigtop</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hadoop Commands from the CLI</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDFS2 Support</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDFS1 Support</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDFS1 Support</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MapReduce v1 Support</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MapReduce v2 (YARN) Support</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pig Support</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hive Support</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hive Server</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**About Service Level vSphere HA 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.

**Hadoop Feature Support By 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.

**Table 1-3. Hadoop Feature Support**

<table>
<thead>
<tr>
<th>Version</th>
<th>Apache Hadoop</th>
<th>Apache Bigtop</th>
<th>Cloudera</th>
<th>Hortonworks</th>
<th>Hortonworks</th>
<th>Intel</th>
<th>MapR</th>
<th>Pivotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>2.0</td>
<td>CDH4, CDH5</td>
<td>HDP 1.3</td>
<td>HDP 2.0-2.1</td>
<td>2.5.1, 3.0.2</td>
<td>3.0.2-3.1.0</td>
<td>PHD 1.1, 2.0</td>
<td>No</td>
</tr>
<tr>
<td>HDFS1</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>HDFS2</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>MapReduce v1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>MapReduce v2 (YARN)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pig</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Hive</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Hive Server</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Table 1-3. Hadoop Feature Support (Continued)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Apache Hadoop</th>
<th>Apache Bigtop</th>
<th>Cloudera</th>
<th>Hortonworks</th>
<th>Hortonworks</th>
<th>Intel</th>
<th>MapR</th>
<th>Pivotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBase</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>ZooKeeper</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Installing Big Data Extensions

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

Procedure

1. **System Requirements for Big Data Extensions** on page 20
   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.

2. **Internationalization and Localization** on page 22
   Big Data Extensions supports internationalization (I18N) level 1. 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.

3. **Deploy the Big Data Extensions vApp in the vSphere Web Client** on page 23
   Deploying the Big Data Extensions vApp is the first step in getting your Hadoop cluster up and running with vSphere Big Data Extensions.

4. **Install RPMs in the Serengeti Management Server Yum Repository** on page 26
   Install the wsdl4j and mailx RPM packages within the Serengeti Management Server's internal Yum repository.

5. **Install the Big Data Extensions Plug-In** on page 27
   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 using vSphere Web Client 5.1 and later. If you install Big Data Extensions on vSphere 5.0, perform all administrative tasks using the Serengeti Command-Line Interface Client.

6. **Connect to a Serengeti Management Server** on page 28
   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.

7. **Install the Serengeti Remote Command-Line Interface Client** on page 29
   Although the Big 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 Command-line Interface Client.

8. **Access the Serengeti Command-Line Interface Using the Remote Command-Line Interface Client** on page 30
   You can access the Serengeti Command-Line Interface using the Serengeti Remote Command-Line Interface Client. The Serengeti Remote Command-Line Interface Client lets you access the Serengeti Management Server to deploy, manage, and use Hadoop.
What to do next

If you want to create clusters on any Hadoop distribution other than Apache Hadoop, which is included in the Serengeti Management Server, install and configure the distribution for use with Big Data Extensions. See Chapter 4, "Managing Hadoop Distributions," on page 43.

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 5.0 (or later) Enterprise or Enterprise Plus.

**NOTE** The Big Data Extensions graphical user interface is supported only when using vSphere Web Client 5.1 and later. If you install Big Data Extensions on vSphere 5.0, perform all administrative tasks using the Serengeti Command-Line Interface.

- When installing Big Data Extensions on vSphere 5.1 or later, use VMware® vCenter™ Single Sign-On to provide user authentication. When logging in to vSphere 5.1 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 configuring 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.

**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 128 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.

**Internationalization and Localization**

Big Data Extensions supports internationalization (I18N) level 1. 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 VMware recommends configuring 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 creating 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
- cluster name
- nodeGroup name
- node name
- virtual machine name

The following attributes in the Serengeti cluster specification file.
- 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.

Deploy the Big Data Extensions vApp in the vSphere Web Client

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

Prerequisites
- Install and configure vSphere. See “System Requirements for Big Data Extensions,” on page 20.
  - 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.

See “System Requirements for Big Data Extensions,” on page 20 for a complete list.

**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.
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.

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, data store, and network when you deploy the vApp, use the vSphere Web Client or the Serengeti Command-Line Interface Client to specify the resource pool, data store, and network information before you create a Hadoop cluster.

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

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 27.

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` RPM packages within the Serengeti Management Server's internal Yum repository.

**Prerequisites**

Deploy the Big Data Extensions vApp. See “Deploy the Big Data Extensions vApp in the vSphere Web Client,” on page 23.

**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.

     ```bash
     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 can not connect to the Internet, you must manually download the RPMs, copy the files to the required directory, and create a repository.

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

        ```bash
        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.

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

**What to do next**

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 using vSphere Web Client 5.1 and later. If you install Big Data Extensions on vSphere 5.0, perform all administrative tasks using the Serengeti Command-Line Interface Client.

The Big Data Extensions plug-in provides a graphical user interface 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 Command-Line Interface 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**


- 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 can not 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 the vSphere Web Client 5.1 or later.
   
   `https://hostname-or-ip-address:port/vsphere-client`

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

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

3. Using the vSphere Web Client Navigator panel, locate the Serengeti Management Server that you want to register with the plug-in.

   You can find the Serengeti Management Server under the datacenter and resource pool into which you deployed it in the previous task.

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.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register or Unregister Radio Button</td>
<td>Select the <strong>Install</strong> radio button to install the plug-in. Select <strong>Uninstall</strong> to uninstall the plug-in.</td>
</tr>
<tr>
<td><strong>vCenter Server host name or IP address</strong></td>
<td>Type the server host name or IP address of vCenter Server. <strong>Note</strong>: Do not include <code>http://</code> or <code>https://</code> when you type the host name or IP address.</td>
</tr>
<tr>
<td><strong>User Name and Password</strong></td>
<td>Type the user name and password with administrative privileges that you use to connect to vCenter Server. The user name and password can not contain characters whose UTF-8 encoding is greater than 0x8000.</td>
</tr>
<tr>
<td><strong>Big Data Extensions Package URL</strong></td>
<td>The URL with the IP address of the management-server virtual machine where the Big Data Extensions plug-in package is located: <code>https://management-server-ip-address/vcplugin/serengeti-plugin.zip</code></td>
</tr>
</tbody>
</table>

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 From the Inventory pane, click **Big Data Extensions**.

**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 its Serengeti Management Server.

**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 using this procedure, 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 within 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. See “Deploy the Big Data Extensions vApp in the vSphere Web Client,” on page 23.

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

- Ensure that vCenter Single Sign-On is enabled and configured for use by Big Data Extensions for vSphere 5.1 and later. See “Deploy the Big Data Extensions vApp in the vSphere Web Client,” on page 23.

- Install the Big Data Extensions plug-in. See “Install the Big Data Extensions Plug-In,” on page 27.
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 dialog, click the Connect Server link.
5. Navigate to the Serengeti Management Server virtual machine within the Big Data Extensions vApp to which you want 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 in the Summary tab of the Big Data Extensions Home.

What to do next

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

Install the Serengeti Remote Command-Line Interface Client

Although the Big 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 Command-line Interface Client.

**IMPORTANT** You can only run Hadoop commands from the Serengeti CLI on a cluster running the Apache Hadoop 1.2.1 distribution. To use the command-line to run Hadoop administrative commands for clusters running other Hadoop distributions, such as `cfg`, `fs`, `mr`, `pig`, and `hive`, use a Hadoop client node to run these commands.

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 Command-line Interface Client.
  - If you are deploying on vSphere 5.1 or later, the Serengeti Command-line Interface Client uses your vCenter Single Sign-On credentials.
  - If you are deploying on vSphere 5.0, the Serengeti Command-line Interface Client uses the default vCenter Server administrator 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 Command-line Interface 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 Command-line Interface 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 Command-line Interface Client download package.

6 Change to the cli directory, and run the following command to open the Serengeti Command-line Interface Client:

   java -jar serengeti-cli-version.jar

What to do next
To learn more about using the Serengeti Command-line Interface Client, see the VMware vSphere Big Data Extensions Command-line Interface Guide.

Access the Serengeti Command-Line Interface Using the Remote Command-Line Interface Client

You can access the Serengeti Command-Line Interface using the Serengeti Remote Command-Line Interface Client. The Serengeti Remote Command-Line Interface Client lets you access the Serengeti Management Server to deploy, manage, and use Hadoop.

IMPORTANT  You can only run Hadoop commands from the Serengeti CLI on a cluster running the Apache Hadoop 1.2.1 distribution. To use the command-line to run Hadoop administrative commands for clusters running other Hadoop distributions, such as cfg, fs, mr, pig, and hive, use a Hadoop client node to run these commands.

Prerequisites
- Use the vSphere Web Client to log in to the vCenter Server instance on which you deployed the Serengeti vApp.
- Verify that the Big Data Extensions vApp deployment was successful and that the Serengeti Management Server is running.
- Verify that you have the correct password to log in to Serengeti Command-Line Interface Client. See “Create a User Name and Password for the Serengeti Command-Line Interface,” on page 66.

   The Serengeti Command-Line Interface Client 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 Open a Web browser to connect to the Serengeti Management Server cli directory.

   http://ip_address/cli

2 Download the ZIP file for your version and build.

   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 Command-Line Interface 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.
   java -jar serengeti-cli-version_number.jar
6 Connect to the Serengeti service.
   You must run the connect host command every time you begin a command-line session, and after the
   30 minute session timeout. You must run this command, or 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.
   c Type your password.

A command shell opens, and the Serengeti Command-Line Interface prompt appears.

- 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 the Tab key to complete a command.
You can use VMware vSphere® Update Manager™ to upgrade Big Data Extensions from earlier versions.

Procedure

1. Prepare to Upgrade Big Data Extensions on page 34
   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.

2. Upgrade Big Data Extensions Virtual Appliance on page 35
   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 Command-Line Interface, and the Big Data Extensions plug-in must all be the same, it is important that you upgrade all components to the new version.

3. Upgrade the Big Data Extensions Plug-in on page 38
   You must use the same version of the Serengeti Management Server and the Big Data Extensions plug-in.

4. Upgrade the Serengeti Command-Line Interface on page 38
   You must upgrade the Serengeti Command-Line Interface because the Serengeti Command-Line Interface must be the same version as your Big Data Extensions deployment.

5. Upgrade the CentOS 6.x Template on page 39
   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 can upgrade the previous CentOS 6.x template so that it is compatible with the Big Data Extensions upgrade.

6. Upgrade Big Data Extensions Virtual Machine Components on page 40
   To enable the Serengeti Management Server to manage clusters 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.
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. See Chapter 2, “Installing Big Data Extensions,” on page 19.

**IMPORTANT** Do not delete any files in 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.

**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. For each cluster that is in AUTO scaling mode, change the scaling mode to MANUAL.
   - Open a command shell and log in to the Serengeti Management Server as user serengeti.
   - Set the scaling mode of the cluster to MANUAL and the `--targetComputeNodeNum` parameter value to the number of provisioned compute nodes in the cluster.
     ```bash
     cluster setParam --name cluster-name --elasticityMode manual --targetComputeNodeNum num-provisioned-compute-nodes
     ```

6. 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.
7 Make sure that the host name of the Serengeti Management Server matches its fully qualified domain name (FQDN).

8 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 Command-Line Interface, 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. See “Prepare to Upgrade Big Data Extensions,” on page 34.

**Procedure**

1 **Configure Proxy Settings** on page 35
   
   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 36
   
   To start the Big Data Extensions upgrade process, you download the upgrade source from the URL that was supplied to you, either in an email notification or by your VMware representative, and accept the license agreement (EULA).

3 **Create an Upgrade Baseline** on page 36
   
   When you upgrade Big Data Extensions virtual appliances, you must create a custom virtual appliance upgrade baseline.

4 **Specify Upgrade Compliance Settings** on page 37
   
   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 37
   
   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.

**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 36.
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 from the URL that was supplied to you, either in an email notification or by your VMware representative, 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
Upgraded 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. See “Configure the Upgrade Remediation Task and Run the Upgrade Process,” on page 37.

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.

**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.

**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 Command-Line Interface**

You must upgrade the Serengeti Command-Line Interface because the Serengeti Command-Line Interface must be the same version as your Big Data Extensions deployment.

**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 Command-Line Interface 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 Command-Line Interface 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 Command-Line Interface Client download package.

9 Change to the CLI directory, and run the following command to open the Serengeti Command-Line Interface 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. See “Upgrade the CentOS 6.x Template,” on page 39.

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. See “Upgrade Big Data Extensions Virtual Machine Components,” on page 40.

### Upgrade the CentOS 6.x Template

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 can upgrade the previous CentOS 6.x template so that it is compatible with the Big Data Extensions upgrade.

You can upgrade a CentOS 6.x template from a previous version of Big Data Extensions so that the CentOS 6.x template is compatible with the Big Data Extensions upgrade. Alternatively, you can create a new CentOS 6.x template virtual machine that has CentOS 6.x and VMware Tools and that is compatible with the Big Data Extensions upgrade. See “Create a Hadoop Template Virtual Machine using RHEL Server 6.x and VMware Tools,” on page 57.

**Prerequisites**

There must be a previous version of the CentOS 6.x template on the CentOS Template virtual machine.

**Procedure**

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

2 Power on the CentOS Template virtual machine.

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

4 From the command shell on the Serengeti Management Server, copy the template upgrade scripts to the CentOS Template virtual machine.
   ```
   scp /opt/serengeti/www/nodeupgrade/serengeti-node-scripts.tar.gz serengeti@template_ip:/tmp
   ```

5 Open another command shell, such as Bash or PuTTY, and log in to the CentOS Template virtual machine.
6 From the command shell on the CentOS Template virtual machine, run the template upgrade scripts.

```bash
  cd /tmp
  sudo tar xzvf ./serengeti-node-scripts.tar.gz -C /
  sudo ln -sf /opt/serengeti/sbin/run_script_series /opt/serengeti/sbin/postupgrade
  sudo bash /opt/serengeti/sbin/postupgrade
  rm -rf /tmp/serengeti-node-scripts.tar.gz
  cp -f /opt/serengeti/sbin/setup-ip.py /opt/vmware/sbin/setup-ip.py
```

7 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 this file, virtual machines cloned from the template cannot get IP addresses. If you power on the Hadoop Template virtual machine to make changes, remove this file before shutting down this virtual machine.

8 From the vSphere Web Client, power off the CentOS Template virtual machine.

9 From the vSphere Web Client, delete any snapshots named `serengeti-snapshot` from the CentOS Template virtual machine.

10 Synchronize the Hadoop Template virtual machine's time 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` and select `Synchronize guest time with host`.

11 From the command shell on the Serengeti Management Server, restart the tomcat service.

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

**What to do next**

Upgrade the tools in your Hadoop and HBase clusters' virtual machines. See “Upgrade Big Data Extensions Virtual Machine Components,” on page 40.

---

**Upgrade Big Data Extensions Virtual Machine Components**

To enable the Serengeti Management Server to manage clusters 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.

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 98.

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 created using a previous version of Big Data Extensions.
Procedure

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 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 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.

What to do next

Stop and restart your Hadoop and HBase clusters.
Managing Hadoop Distributions

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

- **Hadoop Distribution Deployment Types** on page 43
  
  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 Apache Hadoop 1.2.1 and Bigtop 0.7.0 distributions are included in the OVA that you download and deploy.

- **Configure a Tarball-Deployed Hadoop Distribution** on page 44
  
  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.

- **Configuring Yum and Yum Repositories** on page 46
  
  You can deploy Cloudera CDH4, Intel, 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.

- **Create a Hadoop Template Virtual Machine using RHEL Server 6.x and VMware Tools** on page 57
  
  You can create a Hadoop Template virtual machine that has a customized version of the 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.

- **Maintain a Customized Hadoop Template Virtual Machine** on page 60
  
  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 Apache Hadoop 1.2.1 and Bigtop 0.7.0 distributions are included in the OVA that you download and deploy.

Depending on which Hadoop distribution that 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, and the distribution name, vendor abbreviation, and version number to use as input parameters when configuring the distribution for use with Big Data Extensions.
### Table 4-1. Hadoop Deployment Types

<table>
<thead>
<tr>
<th>Hadoop Distribution</th>
<th>Version Number</th>
<th>Vendor Abbreviation</th>
<th>Deployment Type</th>
<th>HVE Support?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache</td>
<td>1.2.1</td>
<td>Apache</td>
<td>Tarball</td>
<td>Yes</td>
</tr>
<tr>
<td>Bigtop</td>
<td>0.7.0 and later</td>
<td>BIGTOP</td>
<td>Yum</td>
<td>No</td>
</tr>
<tr>
<td>Pivotal HD</td>
<td>1.1-2.0</td>
<td>PHD</td>
<td>Yum</td>
<td>Yes</td>
</tr>
<tr>
<td>Hortonworks Data Platform</td>
<td>1.3.1</td>
<td>HDP</td>
<td>Tarball</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>2.0-2.1.1</td>
<td>HDP</td>
<td>Yum</td>
<td>No</td>
</tr>
<tr>
<td>Cloudera CDH4 and CDH5</td>
<td>4.3.0-5.0</td>
<td>CDH</td>
<td>Yum</td>
<td>No</td>
</tr>
<tr>
<td>MapR</td>
<td>3.0-3.1.0</td>
<td>MAPR</td>
<td>Yum</td>
<td>No</td>
</tr>
<tr>
<td>Intel</td>
<td>2.5-3.0.2</td>
<td>INTEL</td>
<td>Yum</td>
<td>No</td>
</tr>
</tbody>
</table>

### 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 creating 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 2.0 and later include Apache Bigtop, Cloudera CDH4 and CDH5, Hortonworks HDP 2.x, Intel 3.x, and Pivotal PHD 1.1 and later releases.

### Configure a Tarball-Deployed Hadoop Distribution

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

- 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). See “Hadoop Distribution Deployment Types,” on page 43.
- (Optional) Set the password for the Serengeti Management Server. See “Change the Password for the Serengeti Management Server,” on page 64.
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
```

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>--name</code></td>
<td>Name to identify the Hadoop distribution that you are downloading. For example, <code>hdp</code> for Hortonworks. This name can include alphanumeric characters ([a-z], [A-Z], [0-9]) and underscores (&quot;_&quot;).</td>
</tr>
<tr>
<td><code>--vendor</code></td>
<td>Vendor name whose Hadoop distribution you want to use. For example, <code>HDP</code> for Hortonworks.</td>
</tr>
<tr>
<td><code>--version</code></td>
<td>Version of the Hadoop distribution that you want to use. For example, <code>1.3</code>.</td>
</tr>
<tr>
<td><code>--hadoop</code></td>
<td>URL from which to download the Hadoop distribution tarball package from the Hadoop vendor’s Web site.</td>
</tr>
<tr>
<td><code>--pig</code></td>
<td>URL from which to download the Pig distribution tarball package from the Hadoop vendor’s Web site.</td>
</tr>
<tr>
<td><code>--hive</code></td>
<td>URL from which to download the Hive distribution tarball package from the Hadoop vendor’s Web site.</td>
</tr>
<tr>
<td><code>--hbase</code></td>
<td>(Optional) URL from which to download the HBase distribution tarball package from the Hadoop vendor’s Web site.</td>
</tr>
<tr>
<td><code>--zookeeper</code></td>
<td>(Optional) URL from which to download the ZooKeeper distribution tarball package from the Hadoop vendor’s Web site.</td>
</tr>
<tr>
<td>`--hve {true</td>
<td>false}`</td>
</tr>
<tr>
<td><code>--yes</code></td>
<td>(Optional) Specifies that all confirmation prompts from the config-distro.rb script are answered with a “yes” response.</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>name</strong></td>
<td>Directory that is named after the distribution. For example, apache.</td>
</tr>
<tr>
<td><strong>manifest</strong></td>
<td>The <code>manifest</code> file generated by <code>config-distro.rb</code> that is used to download the Hadoop distribution.</td>
</tr>
<tr>
<td><strong>manifest.example</strong></td>
<td>Example <code>manifest</code> file. This file is available before you perform the download. The manifest file is a JSON file with three sections: name, version, and packages.</td>
</tr>
</tbody>
</table>

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 will create. See Chapter 6, “Managing vSphere Resources for Hadoop and HBase Clusters,” on page 69.

You can create and deploy Hadoop or HBase clusters using your chosen Hadoop distribution. See “Create a Hadoop or HBase Cluster in the vSphere Web Client,” on page 78.

**Configuring Yum and Yum Repositories**

You can deploy Cloudera CDH4, Intel, 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 47
  You use the Yum repository configuration values in a file that you create to update the Yum repositories used to install or update Hadoop software on CentOS and other operating systems that use Red Hat Package Manager (RPM).

- **Create a Local Yum Repository for Apache Bigtop, Cloudera, Hortonworks, and MapR Hadoop Distributions** on page 50
  Although publically available Yum repositories exist for Apache Bigtop, Cloudera, Hortonworks, and MapR distributions, creating your own Yum repository can result in faster download times and greater control over the repository.

- **Create a Local Yum Repository for the Intel Hadoop Distribution** on page 52
  Intel does not provide a publically available Yum repository. Creating your own Yum repository for Intel provides you with better access and control over installing and updating your Intel Hadoop distribution software.
Yum Repository Configuration Values

You use the Yum repository configuration values in a file that you create to update the Yum repositories used to install or update Hadoop software on CentOS and other operating systems that use Red Hat Package Manager (RPM).

About Yum Repository Configuration Values

To create a local Yum repository, you create a configuration file that identifies a distribution's file and package names 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 table lists the values to use for the Apache Bigtop, Cloudera, Hortonworks, Intel, 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, "maprtech maprecosystem", and they must be combined into a single line when you use them.

### Apache Bigtop Yum Repository Configuration Values

**Table 4-2. Apache Bigtop Yum Repository Placeholder Values**

<table>
<thead>
<tr>
<th>Placeholder</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>repo_file_name</td>
<td>bigtop.repo</td>
</tr>
</tbody>
</table>
| 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 |
| 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 4-3. Cloudera Yum Repository Placeholder Values

<table>
<thead>
<tr>
<th>Placeholder</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>repo_file_name</td>
<td>cloudera-cdh.repo</td>
</tr>
<tr>
<td>package_info</td>
<td>If you use CDH4, use the values below.</td>
</tr>
<tr>
<td></td>
<td>[cloudera-cdh]</td>
</tr>
<tr>
<td></td>
<td>name=Cloudera’s Distribution for Hadoop</td>
</tr>
<tr>
<td></td>
<td>gpgcheck=1</td>
</tr>
<tr>
<td></td>
<td>If you use CDH5, use the values below.</td>
</tr>
<tr>
<td></td>
<td>[cloudera-cdh]</td>
</tr>
<tr>
<td></td>
<td>name=Cloudera’s Distribution for Hadoop</td>
</tr>
<tr>
<td></td>
<td>gpgcheck=1</td>
</tr>
<tr>
<td>mirror_cmds</td>
<td>reposync -r cloudera-cdh4</td>
</tr>
<tr>
<td>default_rpm_dir</td>
<td>cloudera-cdh/RPMS</td>
</tr>
<tr>
<td>target_rpm_dir</td>
<td>cdh/version_number</td>
</tr>
<tr>
<td>local_repo_info</td>
<td>[cloudera-cdh]</td>
</tr>
<tr>
<td></td>
<td>name=Cloudera’s Distribution for Hadoop</td>
</tr>
<tr>
<td></td>
<td>baseurl=http://ip_of_yum_repo_webserver/cdh/version_number/</td>
</tr>
<tr>
<td></td>
<td>enabled=1</td>
</tr>
<tr>
<td></td>
<td>gpgcheck=0</td>
</tr>
</tbody>
</table>

### Intel Yum Repository Configuration Values

### Table 4-4. Intel Yum Repository Placeholder Values

<table>
<thead>
<tr>
<th>Placeholder</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>repo_file_name</td>
<td>intel.repo</td>
</tr>
<tr>
<td>package_info</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>mirror_cmds</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>default_rpm_dir</td>
<td>intel</td>
</tr>
<tr>
<td>target_rpm_dir</td>
<td>intel/2</td>
</tr>
<tr>
<td>local_repo_info</td>
<td>[intel]</td>
</tr>
<tr>
<td></td>
<td>name=Intel Hadoop Distribution 2.x</td>
</tr>
<tr>
<td></td>
<td>baseurl=http://ip_of_yum_repo_webserver/intel/2/idh</td>
</tr>
<tr>
<td></td>
<td>enabled=1</td>
</tr>
<tr>
<td></td>
<td>gpgcheck=0</td>
</tr>
</tbody>
</table>
### Hortonworks Yum Repository Configuration Values

#### Table 4-5. Hortonworks Yum Repository Placeholder Values

<table>
<thead>
<tr>
<th>Placeholder</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>repo_file_name</td>
<td>hdp.repo</td>
</tr>
<tr>
<td>package_info</td>
<td>[hdp]</td>
</tr>
<tr>
<td></td>
<td>name=Hortonworks Data Platform Version - HDP-2.1.1.0</td>
</tr>
<tr>
<td></td>
<td>baseurl=<a href="http://public-repo-1.hortonworks.com/HDP/centos6/2.x/GA/2.1.1.0">http://public-repo-1.hortonworks.com/HDP/centos6/2.x/GA/2.1.1.0</a></td>
</tr>
<tr>
<td></td>
<td>gpgcheck=1</td>
</tr>
<tr>
<td></td>
<td>gpgkey=<a href="http://public-repo-1.hortonworks.com/HDP/centos6/2.x/GA/2.1.1.0/RPM-GPG-KEY/RPM-GPG-KEY-Jenkins">http://public-repo-1.hortonworks.com/HDP/centos6/2.x/GA/2.1.1.0/RPM-GPG-KEY/RPM-GPG-KEY-Jenkins</a></td>
</tr>
<tr>
<td></td>
<td>enabled=1</td>
</tr>
<tr>
<td></td>
<td>priority=1</td>
</tr>
<tr>
<td></td>
<td><strong>Note</strong> If you use a version other than HDP 2.1.1.0, use the exact version number of your Hortonworks distribution in the pathname.</td>
</tr>
<tr>
<td>mirror_cmds</td>
<td>reposync -r hdp</td>
</tr>
<tr>
<td>default rpm dir</td>
<td>hdp</td>
</tr>
<tr>
<td>target rpm dir</td>
<td>hdp/2</td>
</tr>
<tr>
<td>local repo info</td>
<td>[hdp]</td>
</tr>
<tr>
<td></td>
<td>name=Hortonworks Data Platform Version - HDP-2.1.1.0</td>
</tr>
<tr>
<td></td>
<td>baseurl=http://ip_of_yum_repo_webserver/hdp/2/</td>
</tr>
<tr>
<td></td>
<td>enabled=1</td>
</tr>
<tr>
<td></td>
<td>gpgcheck=0</td>
</tr>
</tbody>
</table>

### MapR Yum Repository Configuration Values

#### Table 4-6. MapR Yum Repository Placeholder Values

<table>
<thead>
<tr>
<th>Placeholder</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>repo_file_name</td>
<td>mapr.repo</td>
</tr>
<tr>
<td>package_info</td>
<td>[maprtech]</td>
</tr>
<tr>
<td></td>
<td>name=MapR Technologies</td>
</tr>
<tr>
<td></td>
<td>baseurl=<a href="http://package.mapr.com/releases/3.1.0/redhat/">http://package.mapr.com/releases/3.1.0/redhat/</a></td>
</tr>
<tr>
<td></td>
<td>enabled=1</td>
</tr>
<tr>
<td></td>
<td>gpgcheck=0</td>
</tr>
<tr>
<td></td>
<td>protect=1</td>
</tr>
<tr>
<td></td>
<td>[maprecosystem]</td>
</tr>
<tr>
<td></td>
<td>name=MapR Technologies</td>
</tr>
<tr>
<td></td>
<td>baseurl=<a href="http://package.mapr.com/releases/ecosystem/redhat">http://package.mapr.com/releases/ecosystem/redhat</a></td>
</tr>
<tr>
<td></td>
<td>enabled=1</td>
</tr>
<tr>
<td></td>
<td>gpgcheck=0</td>
</tr>
<tr>
<td></td>
<td>protect=1</td>
</tr>
<tr>
<td></td>
<td><strong>Note</strong> If you use a version other than 3.1.0, use the exact version number of your MapR distribution in the pathname.</td>
</tr>
<tr>
<td>mirror_cmds</td>
<td>reposync -r maprtech</td>
</tr>
<tr>
<td></td>
<td>reposync -r maprecosystem</td>
</tr>
<tr>
<td>default rpm dir</td>
<td>maprtech maprecosystem</td>
</tr>
</tbody>
</table>
Table 4-6. MapR Yum Repository Placeholder Values (Continued)

<table>
<thead>
<tr>
<th>Placeholder</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>target_rpm_dir</td>
<td>mapr/3</td>
</tr>
<tr>
<td>local_repo_info</td>
<td>[mapr]</td>
</tr>
<tr>
<td></td>
<td>name=MapR Version 3</td>
</tr>
<tr>
<td></td>
<td>baseurl=http://ip_of_yum_repo_webserver/mapr/3/</td>
</tr>
<tr>
<td></td>
<td>enabled=1</td>
</tr>
<tr>
<td></td>
<td>gpgcheck=0</td>
</tr>
<tr>
<td></td>
<td>protect=1</td>
</tr>
</tbody>
</table>

Pivotal Yum Repository Configuration Values

Table 4-7. Pivotal Yum Repository Placeholder Values

<table>
<thead>
<tr>
<th>Placeholder</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>repo_file_name</td>
<td>phd.repo</td>
</tr>
<tr>
<td>package_info</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>mirror_cmds</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>default_rpm_dir</td>
<td>pivotal</td>
</tr>
<tr>
<td>target_rpm_dir</td>
<td>phd/1</td>
</tr>
<tr>
<td>local_repo_info</td>
<td>[pivotalhd]</td>
</tr>
<tr>
<td></td>
<td>name=PHD Version 1.0</td>
</tr>
<tr>
<td></td>
<td>baseurl=http://ip_of_yum_repo_webserver/phd/1/</td>
</tr>
<tr>
<td></td>
<td>enabled=1</td>
</tr>
<tr>
<td></td>
<td>gpgcheck=0</td>
</tr>
</tbody>
</table>

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

Although publically available Yum repositories exist for Apache Bigtop, Cloudera, Hortonworks, and MapR 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.
- An HTTP server with which to create the Yum repository. For example, Apache Lighttpd.
- 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 47.
Procedure

1. If your Yum server requires an HTTP proxy server to connect to the Internet, open a command shell, such as Bash or PuTTY, log in to the Serengeti Management Server, and run the commands to export the http_proxy environment variable.

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

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>proxy_server</td>
<td>The IP address or domain name of the proxy server.</td>
</tr>
<tr>
<td>port</td>
<td>The network port number to use with the proxy server.</td>
</tr>
</tbody>
</table>

2. Install the Web server that you want to use as a Yum server.
   This example installs the Apache Web 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. If they are not installed, install the Yum utils and createrepo packages.
   The Yum utils package includes 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 Web Server document root.

   The default Apache document root is /var/www/html/. If you use the Serengeti Management Server as your Yum server machine, the document root is /opt/serengeti/www/.

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

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

   ```
   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 repo 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 55.

Create a Local Yum Repository for the Intel Hadoop Distribution

Intel does not provide a publicly available Yum repository. Creating your own Yum repository for Intel provides you with better access and control over installing and updating your Intel Hadoop distribution software.

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

Prerequisites
- High-speed Internet access.
- CentOS 6.x 64-bit operating system.

**Note** Because the Intel Hadoop distribution requires CentOS 6.1 or later 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 Web 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 server requires an HTTP proxy server, open a command shell, such as Bash or PuTTY, log in to the Serengeti Management Server, and run the commands to export the `http_proxy` environment variable.

```
# switch to root user
sudo su
export http_proxy=http://<proxy_server:port>
```

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>proxy_server</td>
<td>The IP address or domain name of the proxy server.</td>
</tr>
<tr>
<td>port</td>
<td>The network port number to use with the proxy server.</td>
</tr>
</tbody>
</table>
2 Install the Web server that you want to use with a Yum server.

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

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

3 If they are not installed, install the Yum utils and createrepo packages.

The Yum utils package includes the reposync command.

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

4 Download Intel Hadoop 2.5.1 from the Intel Web site.

5 Extract the tarball that you downloaded.

```bash
tar -xzf intelhadoop-2.5.1-20659-zh.el6.x86_64.tar.gz
```

6 Omit

7 Create and configure the local Yum repository.

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

The default Apache document root is `/var/www/html/`. If you use the Serengeti Management Server as your Yum server machine, the document root is `/opt/serengeti/www/`.

```bash
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 Intel Hadoop distribution.

```bash
mv intelhadoop/idh/ $doc_root/intel/2/
```

b (Optional) Create a file, `$doc_root/$target_rpm_dir/intelhadoop.repo`, and include the `local_repo_info`.

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:

```properties
# 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

- After you create a Yum repository for your Intel Hadoop software, configure your Intel Hadoop deployment for use with Big Data Extensions. See “Configure a Yum-Deployed Hadoop Distribution,” on page 55.
Create a Local Yum Repository for the Pivotal Hadoop Distribution

Pivotal does not provide a publically available Yum repository. Creating your own Yum repository for Pivotal 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 operating system.
- An HTTP server with which to create the Yum repository. For example, Apache Lighttpd.
- 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 server requires an HTTP proxy server, open a command shell, such as Bash or PuTTY, log in to the Serengeti Management Server, and run the commands to export the `http_proxy` environment variable.

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

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>proxy_server</td>
<td>The IP address or domain name of the proxy server.</td>
</tr>
<tr>
<td>port</td>
<td>The network port number to use with the proxy server.</td>
</tr>
</tbody>
</table>

2. Install the Web server that you want to use with a Yum server.

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

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

3. If they are not installed, install the Yum `utils` and `createrepo` packages.

   The Yum `utils` package includes the `reposync` command.

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

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

5. Extract the tarball that you downloaded.

   ```bash
   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 Web Server document root.

The default Apache document root is /var/www/html/. If you use the Serengeti Management Server as your Yum server machine, the document root is /opt/serengeti/www/.

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

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 mutli 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 distribution's official Yum repositories or your local Yum repositories.

Prerequisites

- High-speed Internet access from the Serengeti Management Server to download RPM packages from your chosen Hadoop distribution's official Yum repository. If you do not have adequate Internet access to download the RPM packages from your Hadoop vendor's Yum repository, you can create a local Yum repository for your Hadoop distribution.
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. See “Hadoop Distribution Deployment Types,” on page 43.

Create a local Yum repository for your Hadoop distribution. Creating your own repository can result in better access and more control over the repository. See “Configuring Yum and Yum Repositories,” on page 46.

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
```

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>--name</td>
<td>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 (“_”).</td>
</tr>
<tr>
<td>--vendor</td>
<td>Abbreviation of vendor name whose Hadoop distribution you want to use. For example, CDH.</td>
</tr>
<tr>
<td>--version</td>
<td>Version of the Hadoop distribution that you want to use. For example, 4.6.0.</td>
</tr>
<tr>
<td>--repos</td>
<td>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.</td>
</tr>
</tbody>
</table>

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
```

This 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 Intel Hadoop Distribution to Big Data Extensions.

```
config-distro.rb --name intel --vendor INTEL --version 2.5.1
--repos http://url_to_yum_repo/intel.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.

```bash
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. See Chapter 7, “Creating Hadoop and HBase Clusters,” on page 75.

### 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 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.

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.

**Prerequisites**

- Obtain the IP address of the Serengeti Management Server.
- Locate the VMware Tools version that corresponds to the ESXi version in your data center.

**Procedure**

1 Create a virtual machine template with a 20GB thin provisioned disk, and install RHEL 6.x.
   a Download the RHEL Server 6.x installation ISO from [www.redhat.com](http://www.redhat.com) to a datastore.
   b From vCenter Server, create a virtual machine template with a 20GB thin provision disk and select Red Hat Enterprise Linux 6 (64-bit) as the Guest OS.
   c Right-click the virtual machine, select **CD Device**, and select the datastore ISO file for the RHEL ISO file.
d Under Device Status, select connected and connect at power on, and click OK.

e From the console window, install the RHEL Server 6.x operating system using the default settings. See the Red Hat Enterprise Linux Installation Guide, available on the Red Hat website, for more information.

You can select the language and time zone you want the operating system to use, and you can specify that the swap partition use a smaller size to save disk space (for example, 500MB). The swap partition is not used by Big Data Extensions, so you can safely reduce the size.

From the Package Installation Defaults screen, select Minimal.

The new operating system installs into the virtual machine template.

2 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, skip to Step 4.
- If no IP address information appears, which is the case when DHCP is configured, continue with Step 3.

3 Configure the network.

a Using a text editor open the /etc/sysconfig/network-scripts/ifcfg-eth0 file.

b Locate the following parameters and specify the following configuration.

 DEVICE=eth0
 ONBOOT=yes
 BOOTPROTO=dhcp

c Save your changes and close the file.

d Restart the network service.

 sudo /sbin/service network restart

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

3 Configure the network.

a Using a text editor open the /etc/sysconfig/network-scripts/ifcfg-eth0 file.

b Locate the following parameters and specify the following configuration.

 DEVICE=eth0
 ONBOOT=yes
 BOOTPROTO=dhcp

c Save your changes and close the file.

d Restart the network service.

 sudo /sbin/service network restart

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

4 Install the latest JDK 7 RPM.

a From the Oracle® Java SE 7 Downloads page, download the latest JDK 7 Linux x64 RPM and copy it to the virtual machine template's root folder.

b Install the RPM.

 rpm -Uvh jdk-7uxx-linux-x64.rpm

c Delete the RPM file.

 rm -f jdk-7uxx-linux-x64.rpm

d Edit /etc/environment and add the following line: JAVA_HOME=/usr/java/default

5 Install VMware Tools for RHEL 6.x.

a Right-click the RHEL 6 virtual machine in Big Data Extensions, then select Guest > Install/Upgrade VMware Tools.

b 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
c Run the `tar xf` command to extract the VMware Tools package tar file.

```
tar xf VMwareTools-* .tar.gz
```

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

```
./vmware-install.pl
```

Press Enter to finish the installation.

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

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

6 (Optional) In the vSphere Web Client, right-click the virtual machine and select **Snapshot > Take Snapshot**.

Create a snapshot to use for recovery operations.

7 Deploy the Big Data Extension vApp.

8 Run the installation scripts to customize the virtual machine.

a 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.

b Download the scripts from `https://deployed_serengeti_server_IP/custos/custos.tar.gz`.

c 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
```

d 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.

9 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 this file, virtual machines cloned from the template cannot get IP addresses. If you power on the Hadoop Template virtual machine to make changes, remove this file before shutting down this virtual machine.

10 Shut down virtual machine.

11 If you created a snapshot as described in Step 6, delete it. In the vSphere Web Client, right-click the virtual machine, select **Snapshot > Snapshot Manager**, select the `serengeti-snapshot`, and click **Delete**.

12 Synchronize the Hadoop Template virtual machine’s time 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** and select **Synchronize guest time with host**.

13 In the vSphere Web Client, edit the template settings, and deselect (uncheck) all devices.
14 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.

15 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.

**What to do next**

To modify or update the Hadoop Template virtual machine operating system, remove the serengeti-snapshot that is created each time you shutdown and restart the virtual machine. See “Maintain a Customized Hadoop Template Virtual Machine,” on page 60.

**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**


**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 this file, virtual machines cloned from the template cannot get IP addresses. If you power on the Hadoop Template virtual machine to make changes, remove this file before shutting 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 Hadoop Template virtual machine’s time 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** and select **Synchronize guest time with host**.
Managing the Big Data Extensions Environment

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.

- **Add Specific User Names to Connect to the Serengeti Management Server** on page 64
  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 Command-Line Interface or the Big Data Extensions user interface for use with vSphere Web Client Web Client.

- **Change the Password for the Serengeti Management Server** on page 64
  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.

- **Configure vCenter Single Sign-On Settings for the Serengeti Management Server** on page 65
  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.

- **Create a User Name and Password for the Serengeti Command-Line Interface** on page 66
  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.

- **Stop and Start Serengeti Services** on page 66
  You can stop and start Serengeti services to make a reconfiguration take effect, or to recover from an operational anomaly.
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 Command-Line Interface or the Big Data Extensions user interface for use with vSphere Web Client.

**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. 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, 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 `Users.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 `Users.xml` file can be used to login to the Serengeti Management Server using the Serengeti Command-Line Interface or the Big Data Extensions user interface for use with vSphere Web Client.

**What to do next**

You can create a new user name and password for the Serengeti Command-Line Interface. See “Create a User Name and Password for the Serengeti Command-Line Interface,” on page 66.

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 any node’s virtual machine by using this procedure.
**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. 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.

   ```bash
   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. See “Create a User Name and Password for the Serengeti Command-Line Interface,” on page 66.

---

**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 to the URL of the Serengeti Management Server Administration Portal.

   ```plaintext
   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.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Update the certificate</td>
<td>Click Update Certificate.</td>
</tr>
<tr>
<td>Enable SSO for the first time</td>
<td>Type the Lookup Service URL, and click Enable SSO.</td>
</tr>
</tbody>
</table>

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 28.

**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.

**Prerequisites**


**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 64.

**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`
Managing vSphere Resources for Hadoop and HBase Clusters

Big Data Extensions lets you manage the resource pools, datastores, and networks that you use in the Hadoop and HBase clusters that you create.

- **Add a Resource Pool with the Serengeti Command-Line Interface** on page 70
  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.

- **Remove a Resource Pool with the Serengeti Command-Line Interface** on page 70
  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.

- **Add a Datastore in the vSphere Web Client** on page 70
  You can add datastores to Big Data Extensions to make them available to Hadoop and HBase clusters. Big Data Extensions supports both shared datastores and local datastores.

- **Remove a Datastore in the vSphere Web Client** on page 71
  You remove a datastore from Big Data Extensions when you no longer want the Hadoop clusters you create to use that datastore.

- **Add a Network in the vSphere Web Client** on page 72
  You add networks to Big Data Extensions to make the IP addresses contained by those networks available to Hadoop and HBase clusters.

- **Reconfigure a Static IP Network in the vSphere Web Client** on page 72
  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.

- **Remove a Network in the vSphere Web Client** on page 73
  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.
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.

**Prerequisites**
Deploy Big Data Extensions.

**Procedure**
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
   ```

**What to do next**
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.

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**
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 Hadoop and HBase clusters. Big Data Extensions supports both shared datastores and local datastores.

**Prerequisites**
Install Big Data Extensions.
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.
7 From the Type list, select the datastore type in vSphere.

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
</table>
| 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 Hadoop and HBase 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 Hadoop Cluster in the vSphere Web Client,” on page 88.

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 Hadoop and HBase clusters.

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, 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. Choose the type of addressing to use for the network: Use DHCP to obtain IP addresses or Use static IP addresses.
9. If you chose Use static IP addresses in Step 8, enter one or more IP address ranges.
10. Click OK to save your changes.

The IP addresses of the network are available to Hadoop and HBase clusters that you create within Big Data Extensions.

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, and select 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**


**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**, and select **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

Big Data Extensions lets you create and deploy Hadoop and HBase clusters. A Hadoop or HBase cluster is a special type of computational cluster designed specifically for storing and analyzing large amounts of unstructured data in a distributed computing environment.

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 through the Serengeti Command-Line Interface 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.

Additionally, 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 through the Serengeti Command-Line Interface, at least 550GB of storage and 55GB of memory are recommended. For other Hadoop distributions, at least 350GB of storage and 35GB of memory are recommended.

**CAUTION** When you create a cluster with Big Data Extensions, Big Data Extensions disables the cluster’s virtual machine automatic migration. 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.

- **About Hadoop and HBase Cluster Deployment Types** on page 76
  Big Data Extensions lets you deploy several types of Hadoop and HBase clusters. You need to know about the types of clusters that you can create.

- **About Cluster Topology** on page 77
  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.

- **About HBase Database Access** on page 77
  Serengeti supports several methods of HBase database access.

- **Create a Hadoop or HBase Cluster in the vSphere Web Client** on page 78
  After you complete deployment of the Hadoop distribution, you can create Hadoop and HBase 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.
To achieve a balanced workload or to improve performance and throughput, you can control how Hadoop virtual machines are placed by adding topology awareness to the Hadoop clusters. 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.

### About Hadoop and HBase Cluster Deployment Types

Big Data Extensions lets you deploy several types of Hadoop and HBase clusters. You need to know about the types of clusters that you can create.

You can create the following types of clusters.

- **Basic Hadoop Cluster**: You can create a simple Hadoop deployment for proof of concept projects and other small scale data processing tasks using the basic Hadoop cluster.

- **HBase Cluster**: You can create an HBase cluster. To run HBase MapReduce jobs, configure the HBase cluster to include JobTracker or TaskTracker nodes.

- **Data-Compute Separated Hadoop Cluster**: You can separate the data and compute nodes in a Hadoop cluster, and you can control how nodes are placed on your environment’s vSphere ESXi hosts.

- **Compute-Only Hadoop Cluster**: You can create a compute-only cluster to run MapReduce jobs. Compute-only clusters run only MapReduce services that read data from external HDFS clusters and that do not need to store data.

- **Customized Cluster**: You can use an existing cluster specification file to create clusters using the same configuration as your previously created clusters. You can also edit the file to customize the cluster configuration.

### Hadoop Distributions Supporting MapReduce v1 and MapReduce v2 (YARN)

If the Hadoop distribution you use supports both MapReduce v1 and MapReduce v2 (YARN), the default Hadoop cluster configuration creates a MapReduce v2 cluster.

In addition, if you are using two different versions of a vendor’s Hadoop distribution, and both versions support MapReduce v1 and MapReduce v2, the cluster you create using the latest version with the default Hadoop cluster will use MapReduce v2. Clusters you create with the earlier Hadoop version will use MapReduce v1. For example, if you have both Cloudera CDH 5 and CDH 4 installed within Big Data Extensions, clusters you create with CDH 5 will use MapReduce v2, and clusters you create with CDH 4 will use MapReduce v1.
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 Hadoop or HBase 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 cluster’s virtual machine automatic migration. 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 cluster’s placement policy, 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 HBase cluster’s client node Rest-ful Web Services, which listen on port 8080, by using the curl command.
  
  ```bash
  curl -I http://client_node_ip:8080/status/cluster
  ```

- Use the HBase cluster’s client node Thrift gateway, which listens on port 9090.

### Create a Hadoop or HBase Cluster in the vSphere Web Client

After you complete deployment of the Hadoop distribution, you can create Hadoop and HBase 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

- Deploy the Big Data Extensions vApp. See “Getting Started with Big Data Extensions,” on page 11.
- Install the Big Data Extensions plug-in. See “Install the Big Data Extensions Plug-In,” on page 27.
- Connect to a Serengeti Management Server. See “Connect to a Serengeti Management Server,” on page 28.
- Configure one or more Hadoop distributions. See “Configure a Tarball-Deployed Hadoop Distribution,” on page 44or “Configuring Yum and Yum Repositories,” on page 46.
- 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.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hadoop cluster name</strong></td>
<td>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 &lt; 80 characters.</td>
</tr>
<tr>
<td><strong>Hadoop distro</strong></td>
<td>Select the Hadoop distribution. The list contains the default Apache Hadoop distribution for Big Data Extensions and the distributions that you added to your Big Data Extensions environment. The distribution names match the <code>--name</code> parameter's value that was passed to the <code>config-distro.rb</code> script when the Hadoop distribution was configured. For example, <code>cdh4</code> and <code>mapr</code>. <strong>Note</strong>: To create an Apache Bigtop, Cloudera CDH4 and CDH5, or Pivotal PHD 1.1 cluster, you must configure a valid DNS and FQDN for the cluster's HDF5 and MapReduce traffic. Without valid DNS and FQDN settings, the cluster creation process might fail or the cluster is created but does not function.</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Deployment type</strong></td>
<td>Select the type of cluster you want to create.</td>
</tr>
<tr>
<td></td>
<td>- Basic Hadoop Cluster</td>
</tr>
<tr>
<td></td>
<td>- HBase Cluster</td>
</tr>
<tr>
<td></td>
<td>- Compute-only Hadoop Cluster</td>
</tr>
<tr>
<td></td>
<td>- Data/Compute Separation Hadoop Cluster</td>
</tr>
<tr>
<td></td>
<td>- Customize</td>
</tr>
<tr>
<td></td>
<td>The type of cluster you create determines the available node group selections.</td>
</tr>
<tr>
<td></td>
<td>If you select <strong>Customize</strong>, you can load an existing cluster specification file.</td>
</tr>
<tr>
<td><strong>DataMaster Node Group</strong></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Select a resource template from the drop-down menu, or select <strong>Customize</strong> to customize a resource template.</td>
</tr>
<tr>
<td></td>
<td>For the master node, use shared storage so that you protect this virtual machine with vSphere HA and vSphere FT.</td>
</tr>
<tr>
<td><strong>ComputeMaster Node Group</strong></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Select a resource template from the drop-down menu, or select <strong>Customize</strong> to customize a resource template.</td>
</tr>
<tr>
<td></td>
<td>For the master node, use shared storage so that you protect this virtual machine with vSphere HA and vSphere FT.</td>
</tr>
<tr>
<td><strong>HBaseMaster Node Group (HBase cluster only)</strong></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Select a resource template from the drop-down menu, or select <strong>Customize</strong> to customize a resource template.</td>
</tr>
<tr>
<td></td>
<td>For the master node, use shared storage so that you protect this virtual machine with vSphere HA and vSphere FT.</td>
</tr>
<tr>
<td><strong>Worker Node Group</strong></td>
<td>Worker nodes are virtual machines that run the Hadoop DataNode, TaskTracker, and HBase HRegionServer services. These nodes store HDFS data and execute tasks.</td>
</tr>
<tr>
<td></td>
<td>Select the number of nodes and the resource template from the drop-down menu, or select <strong>Customize</strong> to customize a resource template.</td>
</tr>
<tr>
<td></td>
<td>For worker nodes, use local storage.</td>
</tr>
<tr>
<td></td>
<td><strong>Note</strong>: You can add nodes to the worker node group by using <strong>Scale Out Cluster</strong>. You cannot reduce the number of nodes.</td>
</tr>
<tr>
<td><strong>Client Node Group</strong></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Select the number of nodes and a resource template from the drop-down menu, or select <strong>Customize</strong> to customize a resource template.</td>
</tr>
<tr>
<td></td>
<td><strong>Note</strong>: You can add nodes to the client node group by using <strong>Scale Out Cluster</strong>. You cannot reduce the number of nodes.</td>
</tr>
<tr>
<td><strong>Hadoop Topology</strong></td>
<td>Select the topology configuration that you want the cluster to use.</td>
</tr>
<tr>
<td></td>
<td>- RACK_AS_RACK</td>
</tr>
<tr>
<td></td>
<td>- HOST_AS_RACK</td>
</tr>
<tr>
<td></td>
<td>- HVE</td>
</tr>
<tr>
<td></td>
<td>- NONE</td>
</tr>
<tr>
<td></td>
<td>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 77</td>
</tr>
</tbody>
</table>
Option | Description
--- | ---
**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 cluster’s virtual machine nodes.  
- 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 are from 8 to 128 characters, and include only alphanumeric characters ([0-9, a-z, A-Z]) and 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.

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 a Cluster with Topology Awareness with the Serengeti Command-Line Interface

To achieve a balanced workload or to improve performance and throughput, you can control how Hadoop virtual machines are placed by adding topology awareness to the Hadoop clusters. 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.

#### 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 provided Apache Hadoop, add one or more Hadoop distributions. See the *VMware vSphere Big Data Extensions Administrator’s and User’s Guide*.

#### Procedure

1. Access the Serengeti Command-Line Interface.
2. (Optional) Run the `topology list` command to view the list of available topologies.
   ```
   topology list
   ```
(Optional) If you want the cluster to use HVE or RACK_AS_RACK topologies, create a topology rack-hosts mapping file and upload the file to the Serengeti Management Server.

```
topology upload --fileName name_of_rack_hosts_mapping_file
```

4 Run the `cluster create` command to create the cluster.

```
cluster create --name cluster-name ... --topology {HVE|RACK_AS_RACK|HOST_AS_RACK}
```

**NOTE** To create an Apache Bigtop, Cloudera CDH4 and CDH5, or Pivotal PHD 1.1 cluster, you must configure a valid DNS and FQDN for the cluster's HDFS and MapReduce traffic. Without valid DNS and FQDN settings, the cluster creation process might fail or the cluster is created but does not function.

This example creates an HVE topology.

```
cluster create --name cluster-name --topology HVE --distro name_of_HVE-supported_distro
```

5 View the allocated nodes on each rack.

```
cluster list --name cluster-name --detail
```
Managing Hadoop and HBase Clusters

You can use the vSphere Web Client to start and stop your Hadoop or HBase cluster and to modify 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.

- **Stop and Start a Hadoop Cluster in the vSphere Web Client** on page 84
  You can stop a running Hadoop cluster and start a stopped Hadoop cluster from the vSphere Web Client.

- **Scale Out a Hadoop Cluster in the vSphere Web Client** on page 84
  You specify the number of nodes to use when you create Hadoop clusters. You can scale out the cluster by increasing the number of worker nodes and client nodes.

- **Scale CPU and RAM in the vSphere Web Client** on page 85
  You can increase or decrease a cluster’s compute capacity to prevent CPU or memory resource contention among running jobs.

- **Reconfigure a Hadoop or HBase Cluster with the Serengeti Command-Line Interface** on page 86
  You can reconfigure any Hadoop or HBase cluster that you create with Big Data Extensions.

- **Delete a Hadoop Cluster in the vSphere Web Client** on page 88
  You can delete a cluster using the vSphere Web Client. Deleting a cluster removes it from both the inventory and datastore.

- **About Resource Usage and Elastic Scaling** on page 88
  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.

- **Use Disk I/O Shares to Prioritize Cluster Virtual Machines in the vSphere Web Client** on page 92
  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.

- **About vSphere High Availability and vSphere Fault Tolerance** on page 93
  The Serengeti Management Server leverages vSphere HA to protect the Hadoop master node virtual machine, which can be monitored by vSphere.
Recover from Disk Failure with the Serengeti Command-Line Interface Client on page 93

If there is a disk failure in a Hadoop 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.

Log in to Hadoop Nodes with the Serengeti Command-Line Interface Client on page 94

To perform troubleshooting or to run your management automation scripts, log in to Hadoop master, worker, and client nodes with password-less SSH from the Serengeti Management Server using SSH client tools such as SSH, PDSH, ClusterSSH, and Mussh.

Change the User Password on All of a Cluster's Nodes on page 94

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.

Stop and Start a Hadoop 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.

Scale Out a Hadoop Cluster in the vSphere Web Client

You specify the number of nodes to use when you create Hadoop clusters. You can scale out the cluster by increasing the number of worker nodes and client nodes.

You can scale the cluster by using the vSphere Web Client or the Serengeti Command-Line Interface Client. The command-line interface provides more configuration options than the vSphere Web Client. See the VMware vSphere Big Data Extensions Command-Line Interface Guide.

You cannot decrease the number of worker and client nodes from the vSphere Web Client.

**IMPORTANT** Even if you changed the user password on the cluster's nodes, the changed password is not used for the new nodes that are created when you scale out a cluster. 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

- Verify that the cluster is running. See “Stop and Start a Hadoop Cluster in the vSphere Web Client,” on page 84.
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 to scale out.
5. Click the **All Actions** icon, and select **Scale Out**.
6. From the **Node group** list, select the worker or client node group to scale out.
   - If a node group has 0 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**.
   - You cannot decrease the number of nodes. If you specify an instance number that is less than or equal to the current number of instances, a Scale Out Failed error occurs.

The cluster is updated to include the specified number of nodes.

**Scale CPU and RAM in the vSphere Web Client**

You can increase or decrease a cluster’s compute capacity 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 cluster’s CPU or RAM 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 a cluster’s master node CPU or RAM because Big Data Extensions powers down the virtual machine during the scaling process.

When you scale your cluster’s CPU and RAM, 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 Hadoop Cluster in the vSphere Web Client,” on page 84.

**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.
Reconfigure a Hadoop or HBase Cluster with the Serengeti Command-Line Interface

You can reconfigure any Hadoop or HBase 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 these 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

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster_name</td>
<td>Name of the cluster that you want to reconfigure.</td>
</tr>
<tr>
<td>file_path</td>
<td>The file system path at which to export the specification file.</td>
</tr>
<tr>
<td>cluster_spec_file_name</td>
<td>The name with which to label the exported cluster specification file.</td>
</tr>
</tbody>
</table>

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": {
            // note: any value (int, float, boolean, string) must be enclosed in double quotes
            // here is a sample:
            // "io.file.buffer.size": "4096"
        },
        "hdfs-site.xml": {
        },
        "mapred-site.xml": {
        },
        "hadoop-env.sh": {
            // "HADOOP_HEAPSIZE": "",
            // "HADOOP_NAMENODE_OPTS": "",
            // "HADOOP_DATANODE_OPTS": "",
            // "HADOOP_SECONDARYNAMENODE_OPTS": "",
            // "HADOOP_JOBTRACKER_OPTS": "",
            // "HADOOP_TASKTRACKER_OPTS": "",
```
3 (Optional) If your Hadoop distribution’s JAR files 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": {
      "text": "the full content of fair-scheduler.xml in one line"
    }
  }
}
```

4 Access the Serengeti Command-Line Interface.

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 cluster configuration file’s configuration section, or comment out the attribute using double back slashes (//).

   b  Re-run the cluster config command.
Delete a Hadoop Cluster in the vSphere Web Client

You can delete a cluster using the vSphere Web Client. Deleting a cluster removes it from both the inventory and datastore.

When you create a cluster, Big Data Extensions creates a folder and a resource pool for each cluster, as well as resource pools for each node group within 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 both the inventory and datastore.

You can delete a running cluster, a stopped cluster, or a cluster 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 Hadoop Cluster 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 cluster's resource pool according to the cluster's vSphere configuration. 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 cluster nodes' access priority for datastores 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, its scaling configuration is as follows.

- The cluster’s scaling mode is MANUAL, for manual scaling.

- The cluster’s 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 cluster’s 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 cluster’s 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 cluster's current state, 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

Verify that the cluster you want to optimize is data-compute separated. See “About Hadoop and HBase Cluster Deployment Types,” on page 76

**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.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elasticity mode</td>
<td>Select the type of elasticity mode you want to use. You can choose manual or automatic.</td>
</tr>
<tr>
<td>Target compute nodes</td>
<td>Specify the number of compute nodes the cluster should target for use. This option is applicable only to manual scaling (manual elasticity mode).</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td><strong>Note</strong> A value of “Unset” or “-1” means that the node setting has not been configured and is not applicable.</td>
</tr>
<tr>
<td>Min compute nodes</td>
<td>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).</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td>Max compute nodes</td>
<td>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).</td>
</tr>
</tbody>
</table>

**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 92.

**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.

   ```bash
   /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.

   ```bash
   # 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
   ```

**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**.
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.

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.

Recover from Disk Failure with the Serengeti Command-Line Interface Client

If there is a disk failure in a Hadoop 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 Just a Bunch of Disks). If several disks fail, the Hadoop data node might shutdown. Big Data Extensions lets 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 corresponding failed disk's storage type and placement policies.

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 cluster's nodes, the changed password is not used for the new nodes that are created by the 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.

**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]
   ```
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 password-less SSH from the Serengeti Management Server using SSH client tools such as SSH, PDSH, ClusterSSH, and Mussh.

You can use a user name and password authenticated login to connect to Hadoop cluster nodes over SSH. 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 Hadoop node's password by running the set-password -u command.

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

Change the User Password on All of a Cluster's Nodes

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.

You can change a user's password on all nodes within a given cluster.

**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

- Configure a Hadoop distribution to use with Big Data Extensions.
- Create a cluster. See Chapter 7, “Creating Hadoop and HBase Clusters,” on page 75.

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.
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 in the vSphere Web Client.

- **View Serengeti Management Server Initialization Status** on page 97
  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.

- **View Provisioned Clusters in the vSphere Web Client** on page 98
  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.

- **View Cluster Information in the vSphere Web Client** on page 99
  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.

- **Monitor the Hadoop Distributed File System Status in the vSphere Web Client** on page 100
  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.

- **Monitor MapReduce Status in the vSphere Web Client** on page 101
  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.

- **Monitor HBase Status in the vSphere Web Client** on page 101
  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.

### View Serengeti Management Server Initialization Status

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.

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 to 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.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>View Initialize Status</td>
<td>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.</td>
</tr>
<tr>
<td>View Chef Server Services</td>
<td>Click the Chef Server tree control to expand the list of Chef services.</td>
</tr>
<tr>
<td>Recover a Stopped or Failed Service</td>
<td>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.</td>
</tr>
<tr>
<td>Refresh</td>
<td>Click Refresh to update the information displayed in the Summary page.</td>
</tr>
</tbody>
</table>

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 12, “Troubleshooting,” on page 111.

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 Hadoop or Hbase clusters whose information you can view. See Chapter 7, “Creating Hadoop and HBase Clusters,” on page 75

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 9-1. Cluster Information

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Name of the cluster.</td>
</tr>
<tr>
<td>Status</td>
<td>Status of the cluster.</td>
</tr>
<tr>
<td>Distribution</td>
<td>Hadoop distribution in use by the cluster.</td>
</tr>
<tr>
<td>Elasticity Mode</td>
<td>The elasticity mode in use by the cluster.</td>
</tr>
<tr>
<td>Disk I/O Shares</td>
<td>The disk I/O shares in use by the cluster.</td>
</tr>
<tr>
<td>Resources</td>
<td>The resource pool or vCenter Server cluster in use by the Big Data cluster.</td>
</tr>
<tr>
<td>Information</td>
<td>Number and type of nodes in the cluster.</td>
</tr>
<tr>
<td>Progress</td>
<td>Status messages of actions being performed on the cluster.</td>
</tr>
</tbody>
</table>

### 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 9-2. Cluster Information

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node Group</td>
<td>Lists all nodes by type in the cluster.</td>
</tr>
<tr>
<td>VM Name</td>
<td>Name of the virtual machine on which a node is running.</td>
</tr>
<tr>
<td>Management Network</td>
<td>IP address of the virtual machine.</td>
</tr>
<tr>
<td>Host</td>
<td>Host name, IP address, or Fully Qualified Domain Name (FQDN) of the ESXi host on which the virtual machine is running.</td>
</tr>
<tr>
<td>Status</td>
<td>The virtual machine reports the following status types:</td>
</tr>
<tr>
<td></td>
<td>- Not Exist. Status before you create a virtual machine instance in vSphere.</td>
</tr>
<tr>
<td></td>
<td>- Powered On. The virtual machine is powered on after virtual disks and network are configured.</td>
</tr>
<tr>
<td></td>
<td>- VM Ready. A virtual machine is started and IP is ready.</td>
</tr>
<tr>
<td></td>
<td>- Service Ready. Services inside the virtual machine have been provisioned.</td>
</tr>
<tr>
<td></td>
<td>- Bootstrap Failed. A service inside the virtual machine failed to provision.</td>
</tr>
<tr>
<td></td>
<td>- Powered Off. The virtual machine is powered off.</td>
</tr>
<tr>
<td>Task</td>
<td>Status of in-progress Serengeti operations.</td>
</tr>
</tbody>
</table>
From the **Nodes** tab, select a node group. Information about the node group appears in the Node details panel of the **Nodes** tab.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node Group</td>
<td>Name of the selected node group.</td>
</tr>
<tr>
<td>VM Name</td>
<td>Name of the node group’s virtual machine.</td>
</tr>
<tr>
<td>Management network</td>
<td>Network used for management traffic.</td>
</tr>
<tr>
<td>HDFS Network</td>
<td>Network used for HDFS traffic.</td>
</tr>
<tr>
<td>MapReduce Network</td>
<td>Network used for MapReduce traffic.</td>
</tr>
<tr>
<td>Host</td>
<td>Host name, IP address, or Fully Qualified Domain Name (FQDN) of the ESXi host on which the virtual machine is running.</td>
</tr>
<tr>
<td>vCPU</td>
<td>Number of virtual CPUs assigned to the node.</td>
</tr>
<tr>
<td>RAM</td>
<td>Amount of RAM used by the node.</td>
</tr>
<tr>
<td></td>
<td><strong>Note</strong>: The RAM size that appears for each node shows the allocated RAM, not the RAM that is in use.</td>
</tr>
<tr>
<td>Storage</td>
<td>The amount of storage allocated for use by the virtual machine running the node.</td>
</tr>
<tr>
<td>Error</td>
<td>Indicates a node failure.</td>
</tr>
</tbody>
</table>

**Monitor the Hadoop Distributed File System 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. See Chapter 7, “Creating Hadoop and HBase Clusters,” on page 75.

**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.
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. See Chapter 7, “Creating Hadoop and HBase Clusters,” on page 75

**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.
Using Hadoop Clusters from the Serengeti Command-Line Interface

The Serengeti Command-Line Interface lets you perform Hadoop operations. You can run Hive and Pig scripts, HDFS commands, and MapReduce jobs.

The procedures in this section describe how to use the Serengeti Command-Line Interface, which is typically used by administrators. You can also access the client nodes that are provisioned by the Serengeti Management Server to perform standard Hadoop operations such as moving files and running jobs. And if you have other systems integrated with your Hadoop clusters, those systems can directly communicate with the clusters.

- **Run HDFS Commands with the Serengeti Command-Line Interface** on page 103
  You can run Hadoop Distributed File System (HDFS) commands from the Serengeti Command-Line Interface Client. The HDFS commands let you directly interact with the HDFS as well as other file systems that Hadoop supports.

- **Run MapReduce Jobs with the Serengeti Command-Line Interface** on page 104
  You can run MapReduce jobs on your Hadoop cluster.

- **Run Pig and PigLatin Scripts with the Serengeti Command-Line Interface** on page 104
  You can run Pig and PigLatin scripts from the Serengeti Command-Line Interface.

- **Run Hive and Hive Query Language Scripts with the Serengeti Command-Line Interface** on page 105
  You can run Hive and Hive Query Language (HQL) scripts from the Serengeti Command-Line Interface Client.

### Run HDFS Commands with the Serengeti Command-Line Interface

You can run Hadoop Distributed File System (HDFS) commands from the Serengeti Command-Line Interface Client. The HDFS commands let you directly interact with the HDFS as well as other file systems that Hadoop supports.

The File System (FS) shell provides various commands that let you interact with the HDFS, as well as other file systems that Hadoop supports such as Local FS, HFTP FS, S3 FS, and others. You invoke the FS shell using the `fs` command.

**Procedure**

1. Access the Serengeti CLI.
2. Run the `cluster list` command to list the available clusters.
3. Connect to the target cluster.
   ```
   cluster target --name cluster_name
   ```
Run HDFS commands.

This example uses the `fs put` command to move files from `/home/serengeti/data` to the HDFS path `/tmp`.

```
fs put --from /home/serengeti/data --to /tmp
```

This example uses the `fs get` command to download a file from a specific HDFS to your local filesystem.

```
fs get --from source_path_and_file --to dest_path_and_file
```

This example uses the `fs ls` command to display the contents of the dir1 directory in `/user/serengeti`.

```
fs ls /user/serengeti/dir1
```

This example uses the `fs mkdir` command to create the dir2 directory in the `/user/serengeti` directory.

```
fs mkdir /user/serengeti/dir2
```

Run MapReduce Jobs with the Serengeti Command-Line Interface

You can run MapReduce jobs on your Hadoop cluster.

**Procedure**

1. Access the Serengeti CLI, and connect to a Hadoop cluster.
2. Show available clusters.
   
   `cluster list`
3. Connect to the cluster where you want to run a MapReduce job.
   
   `cluster target --name cluster_name`
4. Run the `mr jar` command.
   
   `mr jar --jarfile path_and_jar_filename --mainclass class_with_main_method [--args double_quoted_arg_list]`

   This example runs the MapReduce job located in the `hadoop-examples-1.2.1.jar` JAR, which is located in the `/serengeti/cli/lib` directory. Two arguments pass to the MapReduce job: `/tmp/input` and `/tmp/output`.

   `mr jar --jarfile /opt/serengeti/cli/lib/hadoop-examples-1.2.1.jar --mainclass org.apache.hadoop.examples.WordCount --args "/tmp/input /tmp/output"

5. Show the output of the MapReduce job.
   
   `fs cat file_to_display_to_stdout`
6. Download the output of the MapReduce job from HDFS to the local file system.
   
   `fs get --from HDFS_file_path_and_name --to local_file_path_and_name`

Run Pig and PigLatin Scripts with the Serengeti Command-Line Interface

You can run Pig and PigLatin scripts from the Serengeti Command-Line Interface.

Pig lets you write PigLatin statements. PigLatin statements are converted by the Pig service into MapReduce jobs, which are executed across your Hadoop cluster.

**Prerequisites**

Create a PigLatin script to execute against your Hadoop cluster.
Procedure
1. Access the Serengeti CLI.
2. Show available clusters.
   
   `cluster list`
3. Connect to the cluster where you want to run a Pig script.
   
   `cluster target --name cluster_name`
4. Run the pig script command to run an existing PigLatin script.
   
   This example runs the PigLatin script `data.pig` located in the `/pig/scripts` directory.
   
   `pig script --location /pig/scripts/data.pig`

What to do next
If the PigLatin script stores its results in a file, you might want to copy that file to your local file system.

Run Hive and Hive Query Language Scripts with the Serengeti Command-Line Interface

You can run Hive and Hive Query Language (HQL) scripts from the Serengeti Command-Line Interface Client.

Hive lets you write HQL statements. HQL statements are converted by the Hive service into MapReduce jobs, which are executed across your Hadoop cluster.

Prerequisites
Create an HQL script to execute against your Hadoop cluster.

Procedure
1. Access the Serengeti CLI.
2. Show available clusters by running the `cluster list` command.
   
   `cluster list`
3. Connect to the cluster where you want to run a Hive script.
   
   `cluster target --name cluster_name`
4. Run the hive script command to run an existing Hive script.
   
   This example runs the Hive script `data.hive` located in the `/hive/scripts` directory.
   
   `hive script --location /hive/scripts/data.hive`
Accessing Hive Data with JDBC or ODBC

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.

- Configure Hive to Work with JDBC on page 107
  The Hive JDBC driver lets you access Hive from a Java program that you write, or a Business Intelligence or similar application that uses JDBC to communicate with database products.

- Configure Hive to Work with ODBC on page 109
  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.

Configure Hive to Work with JDBC

The Hive JDBC driver lets you access Hive from a Java program that you write, or 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.

   ```java
   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 ";
           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) + ":" + 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);
           while (res.next()) {
               System.out.println(res.getString(1) + "\t" + res.getString(2));
           }
           
           // select * query
           sql = "select * from " + tableName;
           System.out.println("Running: " + sql);
   }"
```java
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.

```bash
#!/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.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Source Name</td>
<td>Type a name by which to identify the DSN.</td>
</tr>
<tr>
<td>Host</td>
<td>Fully qualified hostname or IP address of the node running the Hive service.</td>
</tr>
<tr>
<td>Port</td>
<td>Port number for the Hive service. The default is 21000.</td>
</tr>
<tr>
<td>Hive Server Type</td>
<td>Set to HiveServer1 or HiveServer2.</td>
</tr>
<tr>
<td>Authentication</td>
<td>If you are using Hiveserver2, specify the following.</td>
</tr>
<tr>
<td></td>
<td>■ <strong>Mechanism</strong>. Set to User Name.</td>
</tr>
<tr>
<td></td>
<td>■ <strong>User Name</strong>. User name with which to run Hive queries.</td>
</tr>
</tbody>
</table>

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.
The troubleshooting topics provide solutions to problems that you might encounter when using Big Data Extensions.

- **Log Files for Troubleshooting** on page 112
  Big Data Extensions and Serengeti create log files that provide system and status information that you can use to troubleshoot deployment and operation problems.

- **Configure Serengeti Logging Levels** on page 113
  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.

- **Collect Log Files for Troubleshooting** on page 113
  You can collect log files from the Serengeti server or from a cluster's nodes to help you and the VMware support team with troubleshooting.

- **Big Data Extensions Virtual Appliance Upgrade Fails** on page 114
  The upgrade of the Big Data Extensions virtual appliance might fail. If the upgrade process fails, you can try the upgrade again.

- **Troubleshooting Cluster Creation Failures** on page 114
  The cluster creation process can fail for many reasons.

- **Cannot Restart or Reconfigure a Cluster After Changing Its Distribution** on page 121
  After you change a cluster's distribution vendor or distribution version, but not the distribution name, the cluster cannot be restarted or reconfigured.

- **Cannot Restart or Reconfigure a Cluster Whose Time Is Not Synchronized** on page 121
  When the time on the hosts and the Serengeti Management Server drifts apart, the cluster cannot be restarted or reconfigured.

- **Virtual Machine Cannot Get IP Address** on page 122
  When you run a Serengeti command, it fails.

- **vCenter Server Connections Fail to Log In** on page 122
  The Serengeti Management Server tries but fails to connect to vCenter Server.

- **SSL Certificate Error When Connecting to Non-Serengeti Server with the vSphere Console** on page 123
  From the vSphere Web Client, you cannot connect to a non-Serengeti server.
- **Serengeti Operations Fail After You Rename a Resource in vSphere** on page 123
  After you use vSphere to rename a resource, Serengeti commands fail for all Serengeti clusters that use that resource.

- **A New Plug-In Instance with the Same or Earlier Version Number as a Previous Plug-In Instance Does Not Load** on page 123
  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.

- **MapReduce Job Fails to Run and Does Not Appear In the Job History** on page 124
  A submitted MapReduce job fails to run and does not appear in the job history.

- **Cannot Submit MapReduce Jobs for Compute-Only Clusters with External Isilon HDFS** on page 125
  You cannot submit MapReduce Jobs for compute-only clusters that point to an external Isilon HDFS.

- **MapReduce Job Stops Responding on a PHD or CDH4 YARN Cluster** on page 125
  A MapReduce job stops responding on a PHD or CDH4 YARN cluster with one DataNode and one NodeManager agent, each with 378MB of memory.

- **Unable to Connect the Big Data Extensions Plug-In to the Serengeti Server** on page 125
  When you install Big Data Extensions on vSphere 5.1 or later, the connection to the Serengeti Management Server fails to authenticate.

- **Cannot Perform Serengeti Operations after Deploying Big Data Extensions** on page 126
  If the Big Data Extensions plug-in cannot communicate with vCenter Server during initialization, internal errors occur and you cannot perform Serengeti operations.

- **Host Name and FQDN Do Not Match for Serengeti Management Server** on page 127
  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.

- **Upgrade Cluster Error When Using Cluster Created in Earlier Version of Big Data Extensions** on page 128
  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.

## 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 12-1. Log Files**

<table>
<thead>
<tr>
<th>Category</th>
<th>File Name</th>
<th>Information</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serengeti vApp boot-up log</td>
<td>serengeti-boot.log</td>
<td>Deployment time messages, which you can use to troubleshoot an unsuccessful deployment.</td>
<td>/opt/serengeti/logs</td>
</tr>
<tr>
<td>Serengeti server service log</td>
<td>serengeti.log</td>
<td>Web service component logs.</td>
<td>/opt/serengeti/logs</td>
</tr>
<tr>
<td>Serengeti server installation and configuration log</td>
<td>ironfan.log</td>
<td>Software installation and configuration information.</td>
<td>/opt/serengeti/logs</td>
</tr>
<tr>
<td>Serengeti server elastic scaling log</td>
<td>vhm.log</td>
<td>Elastic scaling logs.</td>
<td>/opt/serengeti/logs</td>
</tr>
</tbody>
</table>
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.

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 server or from a cluster's nodes 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.
   ```bash
   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.

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.

Troubleshooting Cluster Creation Failures

The cluster creation process can fail for many reasons.

If cluster creation fails, try to resume the 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** on page 115
  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`.
- **Cannot Create a Cluster Whose Time Is Not Synchronized** on page 116
  When you run the `cluster create` or `cluster create ... --resume` command, the command can fail if there are time discrepancies in the environment.
- **Cannot Create a Cluster with the hdfs-hbase-template-spec.json File** on page 116
  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.
If sufficient storage resources are not available when you run the `cluster create` or `cluster create ... --resume` command, cluster creation fails.

If the Hadoop distribution's server is down when you run the `cluster create` or `cluster create ... --resume` command, cluster creation fails.

The IP address of the Serengeti Management Server changes unexpectedly.

If you disconnect a host from vCenter Server after a failed cluster creation attempt, you cannot successfully resume the cluster creation.

When you create, configure, or resume creating or configuring a cluster, the process stops responding.

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.

### 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/serengeti/logs/ironfan.log`, an error message appears:

```
```

#### 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` or `cluster create ... --resume` command to complete the cluster provisioning process.
Cannot Create a Cluster Whose 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.

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 Hadoop distribution’s server is down when you run the `cluster create` or `cluster create ... --resume` command, cluster creation fails.

Problem

If the Hadoop distribution's server is down when you run the `cluster create` or `cluster create ... --resume` command, cluster creation fails. 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.

<table>
<thead>
<tr>
<th>Distribution Type</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>tarball-deployed</td>
<td>Ensure that the httpd service on the Serengeti Management Server is running.</td>
</tr>
<tr>
<td>Yum-deployed</td>
<td>Ensure that the Yum repository file URLs are correctly configured in the manifest file.</td>
</tr>
</tbody>
</table>

2. Ensure that you can download the necessary file from the failed node.

<table>
<thead>
<tr>
<th>Distribution Type</th>
<th>Necessary File</th>
</tr>
</thead>
<tbody>
<tr>
<td>tarball-deployed</td>
<td>tarball</td>
</tr>
<tr>
<td>Yum-deployed</td>
<td>Yum repository file</td>
</tr>
</tbody>
</table>

**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 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.
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 or 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.

      ```bash
      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.

       ```bash
       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.
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.

Cannot Restart or Reconfigure a Cluster Whose 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.
Virtual Machine Cannot Get IP Address

When you run a Serengeti command, it fails.

Problem
A Serengeti command fails, and the Command-Line Interface 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.

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.
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.

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.

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.

**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.
        ```bash
        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.

### 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


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


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.

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
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 98.

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.
Numerics
401 unauthorized error 115
70-persistent-net.rules file 39

A
accessing
HBase databases 77
Hive data 107, 109
adding
datastores 70
networks 72
resource pools 70
topology 80
Administration Portal, Serengeti Management Server 65, 97
Apache Hadoop distribution
Big Data Extensions feature support 14
configuration values for Big Data Extensions 43
Hadoop feature support 16
Apache log4j logs 113
architecture 14
authenticating, troubleshooting 126

B
balancing workloads 80
baseline, upgrade 36
basic Hadoop clusters 76
Big Data Extensions
support for Hadoop features 14
system requirements 20
troubleshooting 126
upgrading 33
Big Data Extensions plug-in
cannot connect to the Serengeti Server 125
changing versions 123
connecting to Serengeti Management Server 28
installing 27
loading 123
registering 27
upgrading 35, 38
Big Data Extensions vApp
deploying 23
upgrading 35
bootstrap failed 115

C
cannot restart or reconfigure a cluster whose time is not synchronized 121
CentOS 6.x template, upgrading 39
certificates, SSO 65
changing, cluster node passwords 94
CLI command-line interface, upgrading 38
CLI console 29
client nodes for Hadoop 76, 78
clock synchronization, and cluster creation failure 115
Cloudera distribution
administrative commands with the Serengeti CLI 29, 30
Big Data Extensions feature support 14
configuration values for Big Data Extensions 43
DNS and FQDN for cluster traffic 78, 80
Hadoop feature support 16
local Yum repository 50, 55
Service Level vSphere HA 14, 16
Yum repository for deployment 46
cluster nodes, passwords 94
cluster target command 104
cluster config command 86
cluster creation failure
401 unauthorized error 115
Bootstrap Failed 115
distribution download failure 117
insufficient storage space 117
Serengeti server IP address changes 118
template-cluster-spec.json file 116
cluster export command 86
cluster fix command 93
cluster reconfiguration failure 121
cluster restart failure 121
cluster settings requirements 20
cluster specification files
Hadoop distribution JARs 86
reconfiguring clusters 86
resource pool symbolic link 70
samples 116
cluster upgrade, cluster upgrade 40
cluster upgrade command 40
clusters
  basic Hadoop 76
  compute-only 76
  creating. See creating clusters
  customized 76
  data-compute separated 76
  deleting 88
  deploying under different resource pools 70
  failover 93
  HBase 76
  HBase status 101
  managing 83
  provisioning hangs 119
  reconfiguring 86
  running MapReduce jobs on PHCD or CDH4 125
  scaling CPU and RAM 85
  stopping and starting 84
  topology 77, 80
  using 103
  viewing 98, 99
  Command-Line Interface, accessing 30
  compliance settings 37
  configuration files, converting Hadoop XML to Serengeti JSON 86
  configuring
    Hive 107, 109
    local Yum repository 55
    logging levels 113
    Yum and Yum repositories 46
    Yum repositories 47
    Yum repository for Cloudera distribution 55
    Yum repository for MapR distribution 55
  connect command 30
  connecting
    Serengeti service 30
    to Serengeti Management Server 28
  connections
    fail 122, 125
    to vCenter Server 122
  convert-hadoop-conf.rb conversion tool 86
  converting Hadoop XML to Serengeti JSON 86
  CPU and RAM, increasing and decreasing 85
  creating
    clusters, See creating clusters
    local Yum repository 50, 52, 54
    Serengeti CLI users 66
    Yum repository for Cloudera distribution 50
    Yum repository for Intel distribution 52
    Yum repository for MapR distribution 50
    Yum repository for Pivotal distribution 54
  creating clusters
    Hadoop or HBase 78
topology-aware 80
troubleshooting 114–120
D
data-compute separated clusters 76
datastores
  adding 70
  removing 71
  deleting clusters 88
  See also removing deploying Big Data Extensions vApp 23
  OVA 23
deployment log files 112
Direct Attached Storage requirements 20
disconnecting, hosts 119
disk failure, recovering from 93
disk I/O shares 92
distribution download failure 117
E
elastic scaling, log file 112
EULA (license agreement) 36
F
feature support by Hadoop distribution 16
feature support for Hadoop distributions 14
fixed elastic scaling 90, 91
fs get command 103
fs ls command 103
fs mkdir command 103
fs put command 103
G
glossary 7
Greenplum distribution
  Big Data Extensions feature support 14
  configuration values for Big Data Extensions 43
  Hadoop feature support 16
H
Hadoop clusters
  creating 78
  See also clusters
Hadoop distributions
  JAR files 86
  managing 43
  Hadoop Distributed File System, monitoring 100
  Hadoop distribution deployment types 43
  Hadoop feature support by distribution 16
Hadoop nodes
logging in 94
passwords 94
Hadoop Template virtual machines 60
Hadoop Virtualization Extensions (HVE) 43, 77, 80
hadoop-template virtual machine 34
hardware requirements 20
HBase clusters
creating 78
See also clusters
HBase database access 77
HDFS status, See Hadoop Distributed File System
HDFS commands, running 103
Hive, configuring 107, 109
Hive data, accessing 107
Hive scripts, running 105
Hive Query Language 105
hive script command 105
Hortonworks distribution
Big Data Extensions feature support 14
configuration values for Big Data Extensions 43
Hadoop feature support 16
HOST_AS_RACK 77
I
I/O shares 92
initialization status 97
installing, Big Data Extensions plug-in 27
installing remote CLI Client 29
insufficient storage space 117
Intel distribution
Big Data Extensions feature support 14
configuration values for Big Data Extensions 43
Hadoop feature support 16
local Yum repository 52
Yum repository for deployment 46
intended audience 7
IP addresses
adding segments to networks 72
monitoring 99
IP address of virtual machine, cannot get 122
J
Java Database Connectivity, See JDBC
Java Runtime Environment (JRE) 30
JBOD, See Direct Attached Storage
JDBC 107
Just A Bunch Of Disks, See Direct Attached Storage
L
license agreement 36
loading, Big Data Extensions plug-in 123
log files 112
log4j.properties file 86, 113
logging levels, configuring 113
logging in to Hadoop nodes 94
M
Management Server virtual machine, See Serengeti Management Server
managing
clusters 83
vSphere resources 69
manual scaling 88
MapR distribution
administrative commands with the Serengeti CLI 29, 30
Big Data Extensions feature support 14
configuration values for Big Data Extensions 43
Hadoop feature support 16
local Yum repository 50, 55
Yum repository for deployment 46
MapReduce jobs
and compute-only clusters 125
cannot submit 125
fail to run 124
hanging 125
monitoring 101
running 104
master nodes for Hadoop 76, 78
monitoring
Big Data Extensions environment 97
Hadoop Distributed File System 100
HBase clusters status 100
MapReduce jobs 101
nodes 99
resource allocation 99
mr jar command 104
N
network settings requirements 20
networks
adding 72
adding IP addresses 72
removing 73
nodes
cannot get IP address 122
monitoring 99
O
ODBC 107, 109
Open Database Connectivity, See ODBC operations fail 123
optimizing resource usage 90

P
passwords
changing Serengeti Management Server 64
cluster nodes 94
Hadoop nodes 94
pig script command 104
Pig scripts, running 104
Pivotal distribution
administrative commands with the Serengeti CLI 29, 30
Big Data Extensions feature support 14
configuration values for Big Data Extensions 43
DNS and FQDN for cluster traffic 78, 80
Hadoop feature support 16
local Yum repository 54
Yum repository for deployment 46
plug-in, upgrading 38
plug-ins, See Big Data Extensions plug-in
Project Serengeti 13
provisioning, hangs 119
proxy settings, configuring 35

R
rack topologies 80
RACK_AS_RACK 77
reconfiguring, networks 72
recovering from disk failure 93
registering, Big Data Extensions plug-in 27
remote CLI Client, installing 29
Remote Command-Line Interface, upgrading 38
removing
datastores 71
networks 73
resource pools 70
serengeti-snapshot 60
See also deleting
renaming, vSphere resources 123
resource pools
adding 70
removing 70
resource requirements
the Hadoop Cluster 20
vSphere Management Server and templates 20
Resource Requirements for vSphere Management server 20
resource usage 88
resourcepool add command 70
resourcepool delete command 70
resourcepool list command 70
resources, renaming in vSphere 123
REST API transport layer error 126
RHEL 6.x template
customized virtual machines 60
customizing 57
running
HDFS commands 103
Hive scripts 105
MapReduce jobs 104
Pig scripts 104
S
Scale Out Failed error message 84
scaling
CPU and RAM 85
elastic 90
manual 88
scaling out cluster 84
scheduling, fixed elastic scaling 91
scripts
serengeti-start-services.sh 66
serengeti-stop-services.sh 66
set_hostname.sh 127
Serengeti servers
fail to connect to vCenter Server 122
IP address changes 118
service log file 112
Serengeti service 30
Serengeti services, stopping and starting 66
Serengeti operations, troubleshooting 126
Serengeti Command-Line Interface upgrading 35
user names and passwords 66
Serengeti Management Server Administration Portal 65, 97
changing passwords 64
connecting to 28, 35
FQDN 127
host name 127
IP address 118
Single Sign-On (SSO) 65
Serengeti operation failure 123
serengeti-snapshot
created during upgrades 60
creating 57
removing 39, 57, 60
serengeti-snapshot.sh script 66
serengeti-stop-services.sh script 66
Service Level vSphere HA for Cloudera CDH4 14
set_hostname.sh script 127
settings, proxy 35
<table>
<thead>
<tr>
<th>Page Numbers</th>
<th>Index Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>30, 65, 125</td>
<td>Single Sign-On (SSO)</td>
</tr>
<tr>
<td>123</td>
<td>SSL certificates, errors</td>
</tr>
<tr>
<td>84</td>
<td>starting clusters</td>
</tr>
<tr>
<td>66</td>
<td>Serengeti services</td>
</tr>
<tr>
<td>84</td>
<td>stopping clusters</td>
</tr>
<tr>
<td>66</td>
<td>Serengeti services</td>
</tr>
<tr>
<td>125</td>
<td>submitting, problems with MapReduce jobs</td>
</tr>
<tr>
<td>116</td>
<td>synchronize time with Serengeti Management Server</td>
</tr>
<tr>
<td>116</td>
<td>synchronize time on hosts with Serengeti Management Server</td>
</tr>
<tr>
<td>20</td>
<td>system requirements, Big Data Extensions</td>
</tr>
<tr>
<td>43, 44</td>
<td>tarball-deployed Hadoop distributions</td>
</tr>
<tr>
<td>9</td>
<td>updated information</td>
</tr>
<tr>
<td>36</td>
<td>upgrade baseline</td>
</tr>
<tr>
<td>114</td>
<td>upgrade failure</td>
</tr>
<tr>
<td>114</td>
<td>upgrade process, troubleshooting</td>
</tr>
<tr>
<td>37</td>
<td>upgrade compliance settings</td>
</tr>
<tr>
<td>37</td>
<td>upgrade process, running</td>
</tr>
<tr>
<td>37</td>
<td>upgrade remediation task, configuring</td>
</tr>
<tr>
<td>36</td>
<td>upgrade source, downloading</td>
</tr>
<tr>
<td>35</td>
<td>upgrading Big Data Extensions</td>
</tr>
<tr>
<td>35</td>
<td>Big Data Extensions plug-in</td>
</tr>
<tr>
<td>35</td>
<td>Big Data Extensions vApp</td>
</tr>
<tr>
<td>39</td>
<td>CentOS 6.x template</td>
</tr>
<tr>
<td>40</td>
<td>virtual machine components</td>
</tr>
<tr>
<td>34</td>
<td>hadoop-template virtual machine</td>
</tr>
<tr>
<td>34</td>
<td>preparing to upgrade</td>
</tr>
<tr>
<td>35</td>
<td>Serengeti Command-Line Interface</td>
</tr>
<tr>
<td>34</td>
<td>vSphere Update Manager</td>
</tr>
<tr>
<td>80</td>
<td>uploading, topology</td>
</tr>
<tr>
<td>66</td>
<td>user passwords</td>
</tr>
<tr>
<td>66</td>
<td>user names</td>
</tr>
<tr>
<td>64</td>
<td>user names, add</td>
</tr>
<tr>
<td>90</td>
<td>variable elastic scaling</td>
</tr>
<tr>
<td>122</td>
<td>vCenter Server, connections fail to log in</td>
</tr>
<tr>
<td>123</td>
<td>versions, Big Data Extensions plug-in</td>
</tr>
<tr>
<td>99</td>
<td>cluster information</td>
</tr>
<tr>
<td>98</td>
<td>provisioned clusters</td>
</tr>
<tr>
<td>92</td>
<td>virtual machines</td>
</tr>
<tr>
<td>92</td>
<td>disk shares</td>
</tr>
<tr>
<td>57</td>
<td>Hadoop Template</td>
</tr>
<tr>
<td>92</td>
<td>prioritizing</td>
</tr>
<tr>
<td>114</td>
<td>virtual appliance, upgrade failure</td>
</tr>
<tr>
<td>40</td>
<td>virtual machine components, upgrading</td>
</tr>
<tr>
<td>122</td>
<td>virtual machine, cannot get IP address</td>
</tr>
<tr>
<td>93</td>
<td>vSphere Fault Tolerance (FT)</td>
</tr>
<tr>
<td>69</td>
<td>vSphere resources</td>
</tr>
<tr>
<td>69</td>
<td>managing</td>
</tr>
<tr>
<td>70</td>
<td>resource pools</td>
</tr>
<tr>
<td>93</td>
<td>vSphere High Availability (HA)</td>
</tr>
<tr>
<td>34</td>
<td>vSphere Update Manager</td>
</tr>
<tr>
<td>76, 78</td>
<td>worker nodes for Hadoop</td>
</tr>
<tr>
<td>80</td>
<td>workloads, balancing</td>
</tr>
<tr>
<td>46</td>
<td>Yum and Yum repositories, configuring</td>
</tr>
<tr>
<td>47</td>
<td>Yum repository configuration values</td>
</tr>
<tr>
<td>55</td>
<td>Yum repository, local</td>
</tr>
<tr>
<td>50, 52, 54</td>
<td>creating</td>
</tr>
<tr>
<td>43</td>
<td>Yum-deployed Hadoop distributions</td>
</tr>
</tbody>
</table>