Hortonworks Data Platform

Installing HDP on Windows

(Sep 30, 2015)

docs.hortonworks.com

Hortonworks Data Platform: Installing HDP on Windows

The Hortonworks Data Platform, powered by Apache Hadoop, is a massively scalable and 100% open source platform for storing, processing and analyzing large volumes of data. It is designed to deal with data from many sources and formats in a very quick, easy and cost-effective manner. The Hortonworks Data Platform consists of the essential set of Apache Hadoop projects including MapReduce, Hadoop Distributed File System (HDFS), HCatalog, Pig, Hive, HBase, Zookeeper and Ambari. Hortonworks is the major contributor of code and patches to many of these projects. These projects have been integrated and tested as part of the Hortonworks Data Platform release process and installation and configuration tools have also been included.

Unlike other providers of platforms built using Apache Hadoop, Hortonworks contributes 100% of our code back to the Apache Software Foundation. The Hortonworks Data Platform is Apache-licensed and completely open source. We sell only expert technical support, training and partner-enablement services. All of our technology is, and will remain free and open source.

Please visit the Hortonworks Data Platform page for more information on Hortonworks technology. For more information on Hortonworks services, please visit either the Support or Training page. Feel free to Contact Us directly to discuss your specific needs.

Table of Contents

List of Tables

1. Before You Begin

1.1. HDP Components

The Hortonworks Data Platform consists of three layers:

- **Core Hadoop 2**: The basic components of Apache Hadoop version 2.x.
  - **Hadoop Distributed File System (HDFS)**: A special purpose file system designed to provide high-throughput access to data in a highly distributed environment.
  - **YARN**: A resource negotiator for managing high volume distributed data processing. Previously part of the first version of MapReduce.
  - **MapReduce 2 (MR2)**: A set of client libraries for computation using the MapReduce programming paradigm and a History Server for logging job and task information. Previously part of the first version of MapReduce.
- **Essential Hadoop**: A set of Apache components designed to ease working with Core Hadoop.
  - **Apache Pig**: A platform for creating higher level data flow programs that can be compiled into sequences of MapReduce programs, using Pig Latin, the platform's native language.
  - **Apache Hive**: A tool for creating higher level SQL-like queries using HiveQL, the tool's native language, that can be compiled into sequences of MapReduce programs.
  - **Apache HCatalog**: A metadata abstraction layer that insulates users and scripts from how and where data is physically stored.
  - **WebHCat (Templeton)**: A component that provides a set of REST-like APIs for HCatalog and related Hadoop components.
  - **Apache HBase**: A distributed, column-oriented database that provides the ability to access and manipulate data randomly in the context of the large blocks that make up HDFS.
  - **Apache ZooKeeper**: A centralized tool for providing services to highly distributed systems. ZooKeeper is necessary for HBase installations.
- **Supporting Components**: A set of components that allow you to monitor your Hadoop installation and to connect Hadoop with your larger compute environment.
  - **Apache Oozie**: A server based workflow engine optimized for running workflows that execute Hadoop jobs.

- **Apache Sqoop**: A component that provides a mechanism for moving data between HDFS and external structured datastores. Can be integrated with Oozie workflows.
- **Apache Flume**: A log aggregator. This component must be installed manually.
- **Apache Mahout**: A scalable machine learning library that implements several different approaches to machine learning.
- **Apache Knox**: A REST API gateway for interacting with Apache Hadoop clusters. The gateway provides a single access point for REST interactions with Hadoop clusters.
- **Apache Storm:** A distributed, real-time computation system for processing large volumes of data.
- **Apache Spark:** An in-memory data processing engine with access to development APIs to enable rapid execution of streaming, machine learning or SQL workloads requiring iterative access to datasets.
- **Apache Phoenix:** A relational database layer on top of Apache HBase.
- **Apache Tez:** An extensible framework for building high performance batch and interactive data processing applications, coordinated by YARN in Apache Hadoop. For additional information, see the Hortonworks website.
- **Apache Falcon:** A framework for simplifying and orchestrating data management and pipeline processing in Apache Hadoop. For additional information, see the Hortonworks website.
- **Apache Ranger:** The Hadoop cluster security component. Range provides centralized security policy administration for authorization, auditing, and data protection requirements.
- **Apache DataFu:** A library for user defined functions for common data analysis task.
- **Apache Slider:** A YARN-based framework to deploy and manage long running or always-on data access applications.

1.2. Minimum System Requirements

To run the Hortonworks Data Platform, your system must meet minimum requirements.

1.2.1. Hardware Recommendations

When installing HDP, 5 GB of free space is required on the system drive.

1.2.2. Operating System Requirements

The following operating systems are supported:

- Windows Server 2008 R2 (64-bit)
- Windows Server 2012 (64-bit)

- Windows Server 2012 R1 (64-bit)
- Windows Server 2012 R2 (64-bit)

1.2.3. Software Requirements

The following table provides installation information for each software prerequisite. You can also use Microsoft SQL Server for Hive and Oozie metastores. If you plan to install Ranger, MySQL is required. For more information, see the following two subsections.

Table 1.1. Required software installation information

Software	Version	Environment Variable	Description	Installation Notes
Python	2.7.X	PATH	Add the directory where Python is installed, following the instructions in this guide. The path is c:\python.	Spaces in the path to the executable are not allowed. Do not install Python in the default location (Program Files). For more information, see Installing Required Software.
Java JDK	JDK 1.7.0_51	PATH	Add the directory where the Java application is installed; for example, c:\java\jdk1.7.0\bin	Spaces in the path to the executable are not allowed. Do not install Java in the default location (Program Files). For more information, see Installing Required Software.
		JAVA_Home	Create a new system variable for JAVA_HOME that points to the directory where the JDK is installed; for example, c:\java]jdk1.7.0.	
Microsoft Visual C++	2010	PATH	Default location added automatically.	Install with default parameters. For more information, see Installing Required Software.
Microsoft .NET Framework	4.0	PATH	Default location added automatically.	Install with default parameters. For more information, see Installing Required Software.

1.2.4. (Optional) Microsoft SQL Server for Hive and Oozie Database Instances

By default, Hive and Oozie use an embedded Derby database for its metastore. However you can also use Microsoft SQL Server. (For details on installing and configuring Microsoft SQL Server, see TechNet instructions, such as SQL Server 2012.)

To use an external database for Hive and Oozie metastores, ensure that Microsoft SQL Server is deployed and available in your environment, and that your database administrator creates the following databases and users. You will need the following information when you configure HDP:

- For Hive, create a SQL database instance:

  1. Create a Hive database instance in SQL and record its name, such as `hive_dbname`.
  2. Create Hive user on SQL, add it to the `sysadmin` role within SQL, and record the name and password, such as `hive_dbuser/hive_dbpasswd`.
  3. Set the security policy for SQL to use both SQL and Windows authentication. The default setting is Windows authentication only.

- For Oozie, create a SQL database instance:

  1. Create an Oozie database instance and record its name, such as `oozie_dbname`.
  2. Create Oozie user on SQL, add it to the `sysadmin` role within SQL, and record the user name and password, such as `oozie_dbuser/oozie_dbpasswd`.
  3. Set the security policy for SQL to use both SQL and Windows authentication, the default setting is Windows authentication only.

- The following steps are required after installing SQL Server; refer to SQL Server documentation for more information:

  1. Ensure that `TCP/IP` is enabled under **SQL Server Network Configuration**. This might require restarting SQL Server.
  2. Add firewall exceptions for SQL Server and SQL Browser services.
  3. Before using SQL server for Hive or Oozie metastores, set up the Microsoft SQL Server JDBC Driver. For instructions, see (Optional) Install Microsoft SQL Server JDBC Driver.

1.2.5. Requirements for Installing Ranger

Ranger offers a centralized security framework, including security policy administration: authorization, accounting, and data protection. It is an optional component. Before you install Ranger, you must meet several additional requirements.

Table 1.2. Ranger Installation Requirements

Requirement	Installation instructions
Active Directory or LDAP Server	For information on setting up Active Directory or an LDAP Server, see the Microsoft documentation.
MySQL	1. Install MySQL Client. 2. Add the path to `mysql.exe` to your system PATH variable. 3. Set the `JAVA_HOME` variable to the installed JDK version; for example, `$ENV:JAVA_HOME="c:\Java\jdk1.7.0_67"` 4. Download the MySQL Connector Jar file. 5. Copy the jar file into your installation folder: • If you plan to install Ranger using the MSI Setup GUI, copy the jar file into the folder containing the MSI.

Requirement	Installation instructions
	• If you plan to install Ranger using the command-line interface, copy the jar file into the folder containing your cluster properties file.

1.3. Preparing for Hadoop Installation

To deploy HDP across a cluster, you need to prepare your multi-node cluster deploy environment. Follow the steps in this section to ensure each cluster node is prepared to be an HDP cluster node.

1.3.1. Gather Hadoop Cluster Information

Before deploying your HDP installation, collect the host name or IPv4 address of each the following cluster components:

- **Required components:**
  - NameNode and optional Secondary NameNode
  - ResourceManager
  - Hive Server
  - SlaveNode
  - WebHCat
  - Client Host
- **Optional components:**
  - Zookeeper
  - HBase Master
  - Flume
  - Knox Gateway
  - Ranger (requires MySQL client)
  - Microsoft SQL Server configured with a Hive and Oozie database instance
  - system account names and passwords

Note

The installer fails if it cannot resolve the host name of each cluster node. To determine the host name for a particular cluster node, open the command line interface on that system. Execute `hostname` and then `nslookup hostname` to verify that the name resolves to the correct IP address.

1.3.2. Configure the Network Time Server (NTS)

The clocks of all nodes in your cluster must be able to synchronize with each other.

To configure the Network Time Server (NTS) for Windows Server, use the instructions provided in this Microsoft KB article.

You can also download and configure Network Time Protocol.

1.3.3. Set Interfaces to IPv4 Preferred

Configure all the Windows Server nodes in your cluster to use IPv4 addresses only. You can either disable IPv6 (see How to disable IP version 6 or its specific components in Windows) or set the preference to IPv4.

Ensure that the host's fully-qualified domain name ("FQDN") resolves to an IPv4 address as follows:

1. To verify that IPv4 is set to `preferred`, enter:

```
ipconfig /all
```

   The system should display:

```
Connection-specific DNS Suffix . . :
Description . . . . . . . . . . . .:
Intel(R) PRO/1000 MT Network
Connection Physical Address. . . : XX-XX-XX-XX-XX
DHCP Enabled. . . .. . . . .    .: No
Autoconfiguration Enabled .. . . : Yes
IPv4 Address. . . . . . . . . . . : 10.0.0.2(Preferred)
Subnet Mask . . . . . . . . . . . .: 255.255.255.0
Default Gateway . . . . . . . . . .: 10.0.0.100
DNS Servers . . . . . . . . . . . .: 10.10.0.101
NetBIOS over Tcpip. . . . . . . . .: Enabled
```

2. To flush the DNS cache, enter:

```
ipconfig /flushdns
```

3. To verify that the host name of the system resolves to the correct IP address, enter:

```
ping -a 10.0.0.2
```

   The system should display:

```
Pinging win08r2-node1.HWXsupport.com 10.0.0.2 with 32 bytes of data:
Reply from 10.0.0.2: bytes=32 time<1ms TTL=128
Reply from 10.0.0.2: bytes=32 time<1ms TTL=128
Reply from 10.0.0.2: bytes=32 time<1ms TTL=128
```

1.3.4. (Optional) Create Hadoop user

HDP services run under the ownership of a Windows user account. The HDP installer establishes a `hadoop` user as follows:

- If the `hadoop` user account does not exist, HDP installer automatically creates a local user.
- If the `hadoop` user account already exists, HDP installer uses the current password.

You can specify the Hadoop user password in the MSI command line. The administrator can change the password later, but it must be changed both in the user configuration and in the service objects installed on each machine via Service Manager.

1.3.5. Enable Remote PowerShell Script Execution

The MSI installation scripts and many HDP utility scripts require remote execution of PowerShell scripts on all nodes in the Hadoop cluster. For example, the scripts for starting and stopping the entire cluster with a single command (provided with HDP) require remote scripting. Therefore, we strongly recommend that you complete the following steps at every host in your cluster.

1.3.6. Enable Remote PowerShell Execution for Nodes in a Workgroup

You can set these in `Active Directory` via `Group Policies` (for a Group including all hosts in your Hadoop cluster), or you can execute the given PowerShell commands on every host in your cluster.

Important

Ensure that the Administrator account on the Windows Server node has a password. The following instructions will not work if the Administrator account has an empty password.

- To enable remote scripting using PowerShell commands:

  1. At each host in the cluster, open a PowerShell window with "Run as Administrator" privileges, and enter:

     ```
     Set-ExecutionPolicy "RemoteSigned"
     Enable-PSRemoting
     Set-item WSMan:\localhost\Client\allowunencrypted $true
     Set-item wsman:localhost\client\trustedhosts -value "host1,host2"
     ```

     where `host1,host2` is a list of comma-separated host names in your cluster. For example,`"HadoopHost1, HadoopHost2, HadoopHost3"`.

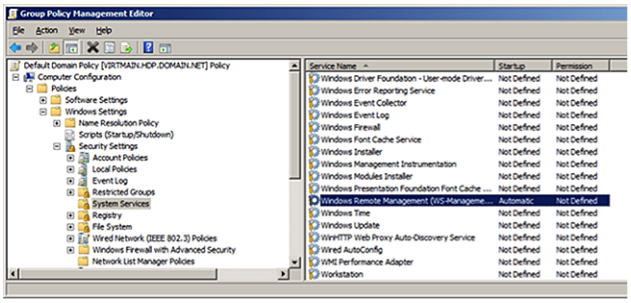
1.3.7. (Optional) Enable Networking Configurations for Active Directory Domains

If your environment is using Active Directory, you must enable remote scripting and configure domain policies for Windows Remote Management, complete the following instructions on a domain controller machine.

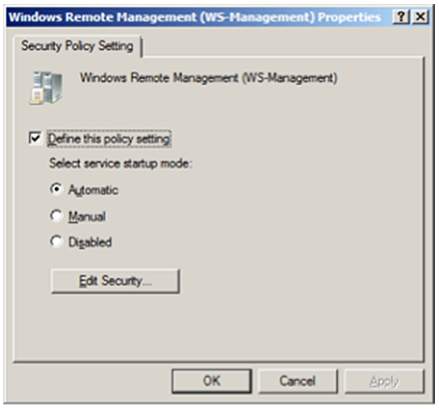
1. Open the Group Policy Management Editor by clicking Default Domain Policy from `Group Policy Management > Domains > <domain name> > Default Domain Policy`, and then click Edit.

2. Set the WinRM service to `autostart`.

   a. From the Group Policy Management Editor, go to `Computer Configuration > Policies > Windows Settings > Security Settings > Windows Remote Management (WS-Management).`



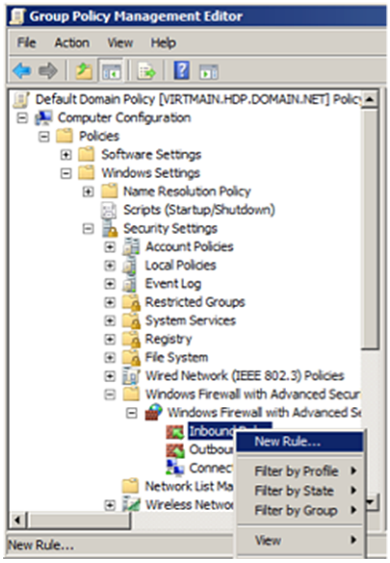
   b. Set `Startup Mode` to `Automatic`.



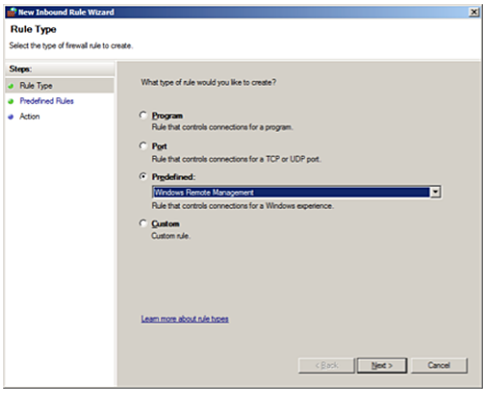
3. Add firewall exceptions to allow the service to communicate.

   a. Go to `Computer Configuration > Policies > Windows Settings > Security Settings > Windows Firewall with Advanced Security`.

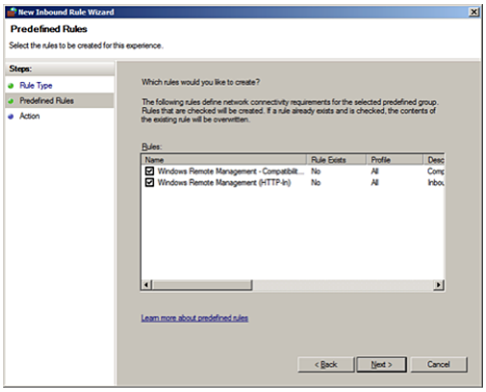
   b. To create a new Inbound Rule, right-click `Windows Firewall with Advanced Security`.



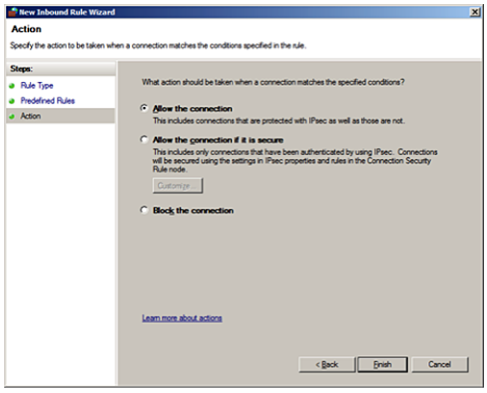
c. Specify the rule type as `Predefined, Windows Remote Management`.



The Predefined rule automatically creates two rules:



d. Configure `Action` as `Allow the connection`.



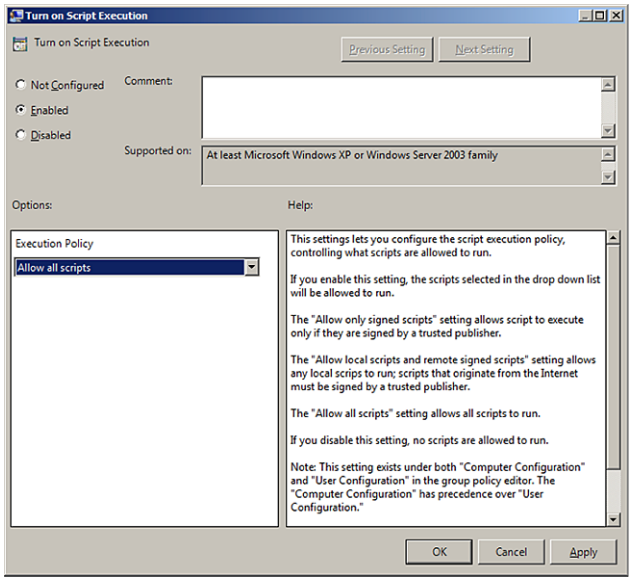
e. Click `Finish`.

4. Set script execution policy.

   a. Go to `Computer Configuration > Policies > Administrative Templates > Windows Components > Windows PowerShell`.

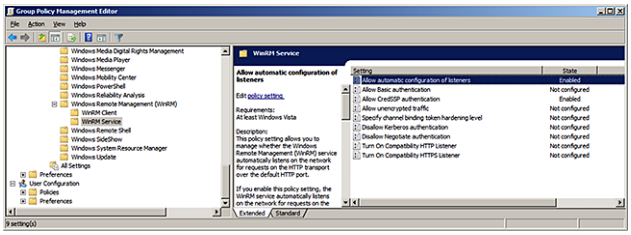
   b. At `Setting`, select `Turn on Script Execution`.

   c. Set `Execution Policy` to `Allow all scripts`.



5. Set up the WinRM service.

   a. Go to `Computer Configuration > Policies > Administrative Templates > Windows Components > Windows Remote Management (WinRM) > WinRM Service`.

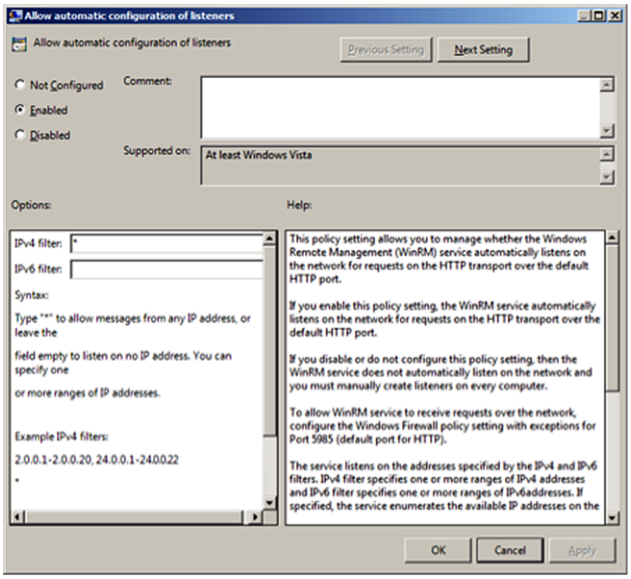


Note

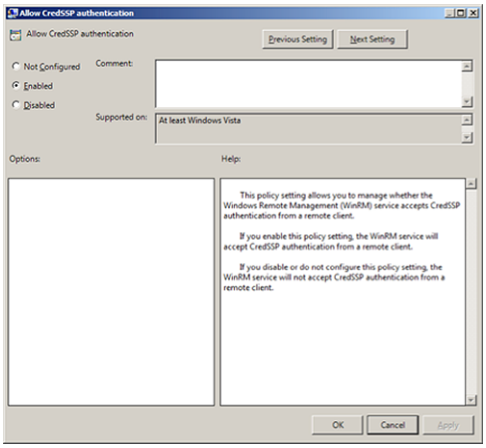
In Windows Server 2012, the "`Allow automatic configuration of listeners`" option has changed to "`Allow remote server management through WinRM`".

b. Create a WinRM listener.

i. To allow automatic configuration of listeners, select `Enabled`, and then set `IPv4 filter` to `*` (all addresses) or specify a range:



ii. Allow `CredSSP` authentication and click `OK`.

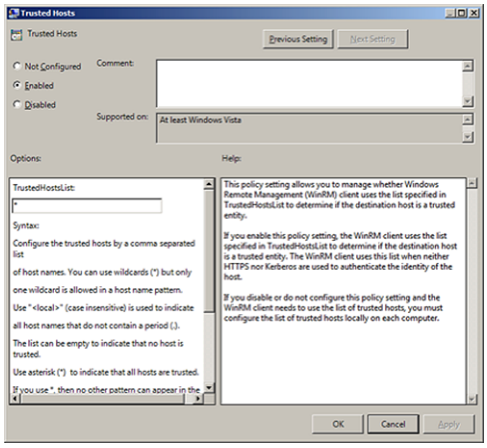


6. Set up the WinRM client.

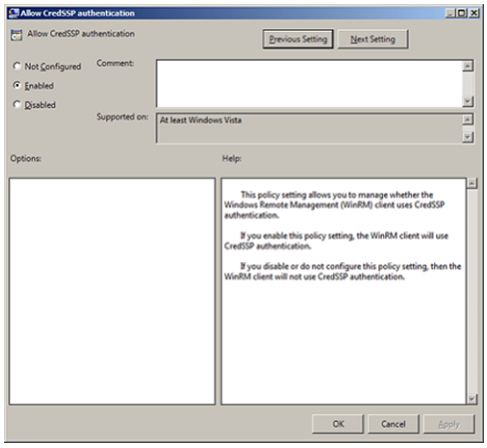
   a. Go to `Computer Configuration > Policies > Administrative Templates > Windows Components > Windows Remote Management (WinRM) > WinRM Client`.

   b. Configure the trusted host list (the IP addresses of the computers that can initiate connections to the WinRM service).

      Set `TrustedHostsList` to * (all addresses) or specify a range.



   c. Set `Allow CredSSP authentication` to `Enabled`, and click `OK`.

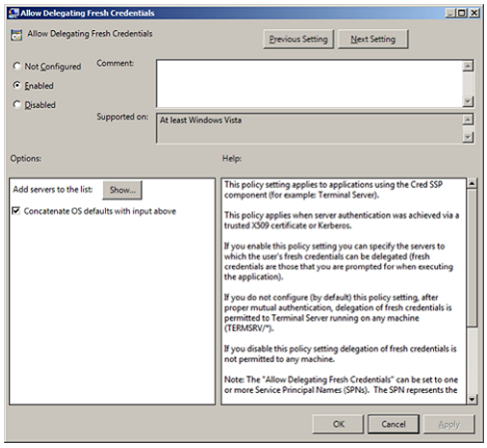


7. Enable credentials delegation.

   a. Go to `Computer Configuration > Policies > Administrative Templates > System > Credentials Delegation`.

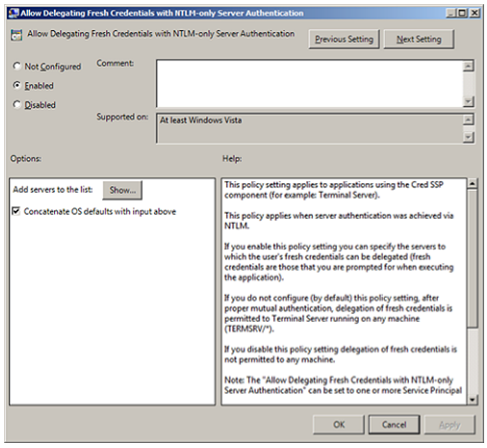
   b. To allow delegation of fresh credentials, select `Enabled`.

   c. Under `Options`, select Show. Set `WSMAN` to * (all addresses) or specify a range. Click `Next Setting`.



   d. Select `Enabled` to allow delegation of fresh credentials with NTLM-only server authentication.

   e. Under `Options` click Show. Set `WSMAN` to * (all addresses), or specify a range. Click `Finish`.



8. Enable the creation of WSMAN SPN.

   a. Go to `Start > Run`. In the dialog box, enter `ADSIEdit.msc`. Click `Enter`.

   b. Expand the `OU=Domain Controllers` menu item and select `CN=domain controller hostname`.

   c. Go to `Properties > Security > Advanced > Add`.

   d. Enter `NETWORK SERVICE`, click `Check Names`, then click `OK`.

   e. In the `Permission` field, select `Validated write to service principal name`.

   f. Click `Allow`.

   g. To save your changes, click OK.

9. Restart the WinRM service and update policies.

   a. At the domain controller machine, open a PowerShell window and enter:

   ```
   Restart-Service WinRM
   ```

   b. At each of the other hosts in domain, enter:

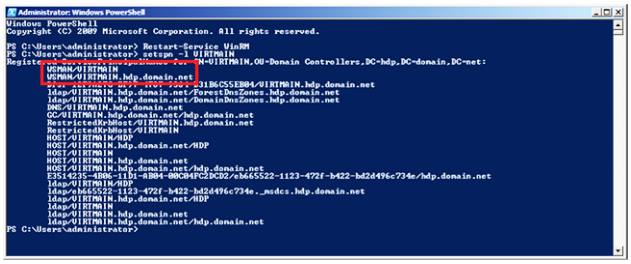
   ```
   gpupdate /force
   ```

   c. Ensure that SPN-s WSMAN is created for your environment.

   At your domain controller machine, enter:

   ```
   setspn -l Domain_Controller_Hostname
   ```

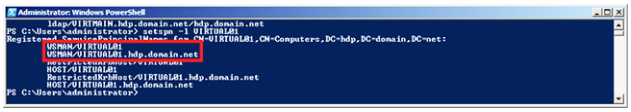
   You should see output similar to the following:



10. Check the WSMAN SPN on other hosts in the domain. Run the following command on any one of your host machines:

```
setspn -l Domain_Controller_Hostname
```

You should see output similar to the following:



1.3.8. Configure Ports

HDP uses multiple ports for communication with clients and between service components. For example, the Hive Metastore port represents the port that the metastore database can use to connect and communicate. To enable HDP communication, open the specific ports that HDP uses.

To open specific ports, you can set the access rules in Windows. For example, the following command will open up port 80 in the active Windows Firewall:

```
netsh advfirewall firewall add rule name=AllowRPCCommunication dir=in action=
allow protocol=TCP localport=80
```

The following command will open up ports 49152-65535 in the active Windows Firewall:

```
netsh advfirewall firewall add rule name=AllowRPCCommunication dir=in action=
allow protocol=TCP localport=49152-65535
```

The following tables specify which ports must be opened for specific ecosystem components to communicate with each other. Open the appropriate ports before you install HDP.

Table 1.3. HDFS Ports

Service	Servers	Default Ports Used	Protocol	Description	Needs End-user Access?	Configuration Parameters
NameNode WebUI	MasterNodes (NameNode and any backup NameNodes)	50070	http	WebUI to look at current status	Yes (typically admins, Development/ Support teams)	self-addressed
NameNode metadata service	Master Nodes (NameNode and any	8020/9000	IPC	File system metadata operations	Yes (all clients who need to interact	Embedded in URI specified by `defaulters`

Service	Servers	Default Ports Used	Protocol	Description	Needs End-user Access?	Configuration Parameters
	backup NameNodes)				directly with the HDFS)	
DataNode	All Slave Nodes	50076	http	DataNode WebUI to access the status, logs, etc.	Yes (typically admins, Development/ Support teams)	dfs.datanode. http.address
DataNode	All Slave Nodes	50010		Data transfer		dfs.datanode. address
DataNode	All Slave Nodes	50020	IPC	Metadata operations	No	dfs.datanode. address
Secondary NameNode	Secondary NameNode and any backup Secondary NameNodes	50090	http	Checkpoint for NameNode metadata	No	dfs.secondary. http.address

Table 1.4. YARN Ports

Service	Servers	Default Ports Used	Protocol	>Description	Needs End-user Access?	Configuration Parameters
Resource Manager WebUI	Master Nodes (Resource Manager and any back-up Resource Manager node)	8088	http	WebUI for Resource Manager	Yes	yarn.resource-manager.webapp. address
Resource Manager WebUI	Master Nodes (Resource Manager and any back-up Resource Manager node)	8090	https	WebUI for Resource Manager	Yes	yarn.resource-manager.webapp. https.address
Resource Manager Admin Interface	Master Nodes (Resource Manager and any back-up Resource Manager node)	8032	IPC	For application submissions	Yes (All clients who need to submit the YARN applications including Hive, Hive server, Pig)	yarn.resource-manager.admin. address
Resource Manager Scheduler	Master Nodes (Resource Manager and any back-up Resource Manager node)	8033		Administrative interface	Yes (Typically admins and support teams)	yarn.resource-manager. scheduler.address
NodeManager Web UI	All Slave Nodes	8031	http	Resource Manager interface	Yes (Typically admins, Development/ Support teams)	yarn. nodemanager. webapp.address

Table 1.5. Hive Ports

Service	Servers	Default Ports Used	Protocol	Description	Needs End-user Access?	Configuration Parameters
HiveServer2	HiveServer 2 machine (usually a utility machine)	10001	thrift	Service for programmatically (Thrift/JDBC) connecting to Hive	Yes	ENV Variable HIVE_PORT
HiveServer	HiveServer machine (usually a utility machine)	10000	thrift	Service for programmatically (Thrift/JDBC) connecting to Hive	Yes (Clients who need to connect to Hive either programmatically or through UI SQL tools that use JDBC)	ENV Variable HIVE_PORT
Hive Metastore		9083	thrift	Service for programmatically (Thrift/JDBC) connecting to Hive metadata	Yes (Clients that run Hive, Pig and potentially M/R jobs that use HCatalog)	hive.metastore.uris

Table 1.6. WebHCat Ports

Service	Servers	Default Ports Used	Protocol	Description	Need End User Access?	Configuration Parameters
WebHCat Server	Any utility machine	50111	http	Web API on top of HCatalog and other Hadoop services	Yes	templeton.port

Table 1.7. HBase Ports

Service	Servers	Default Ports Used	Protocol	Description	Need End User Access?	Configuration Parameters
HMaster	Master Nodes (HBase Master Node and any back- up HBase Master node)	60000			Yes	hbase.master.port
HMaster Info WebUI	Master Nodes (HBase master Node and backup HBase Master node if any)	60010	http	The port for the HBaseMaster WebUI. Set to -1 if you do not want the info server to run.	Yes	hbase.master.info.port
Region Server	All Slave Nodes	60020			Yes (typically admins, development/support teams)	hbase.regionserver.port
ZooKeeper	All ZooKeeper Nodes	2888		Port used by ZooKeeper peers to talk to each other. For further information, see the Apache website.	No	hbase.zookeeper.peerport

Service	Servers	Default Ports Used	Protocol	Description	Need End User Access?	Configuration Parameters
ZooKeeper	All ZooKeeper Nodes	3888		Port used by ZooKeeper for leader election. For further information, see the Apache website.		hbase.zookeeper.leaderport
		2181		Property from ZooKeeper's configuration file, `zoo.cfg`. The port at which the clients connect.		hbase.zookeeper.property.clientPort

1.3.9. Install Required Software

Install the following software on each node in the cluster:

- Python v2.7 or higher
- Java Development Kit version: 1.7.0_51
- Microsoft Visual C++, 2010 only
- Microsoft .NET Framework v4.0

See the following subsections for more information.

1.3.9.1. Installing Required Software using PowerShell CLI

Identify a workspace directory that will have all the software installation files.

In the PowerShell instructions in this section, the `WORKSPACE` environment variable refers to the full path of the workspace directory where the installer is located; for example:

```
setx WORKSPACE "c:\workspace" /m
```

After setting the environment variable from the command prompt using `setx`, restart PowerShell. Alternately, if you are using a script you might want to set `WORKSPACE` as a standard PowerShell variable to avoid having to restart PowerShell.

Ensure that you install the following software on every host machine in your cluster:

- Python 2.7.X

  To manually install Python in your local environment:

  1. Download Python from here into the workspace directory.
  2. Install Python and update the `PATH` environment variable.

     From the PowerShell window, as the Administrator, enter:

```
$key = "HKLM:\SYSTEM\CurrentControlSet\Control\Session
Manager\Environment” $currentPath = (Get-ItemProperty -Path $key -name
Path).Path + ';' $pythonDir = "c:\Python\" msiexec /qn
/norestart /l* $env:WORKSPACE\python_install.log /i
$env:WORKSPACE\python-2_7_5_amd64.msi TARGETDIR=$pythonDir ALLUSERS=1
setx PATH "$currentPath$pythonDir" /m
```

Note

If the downloaded Python MSI name is different from `python-2_7_5_amd64.msi`, substitute the correct MSI file name.

- Microsoft Visual C++ 2010 Redistributable Package (64-bit)

  1. Use the instructions provided here to download Microsoft Visual C++ 2010 Redistributable Package (64-bit) to the workspace directory.

  2. From PowerShell, as Administrator, enter:

```
& "$env:WORKSPACE\vcredist_x64.exe" /q /norestart /log "$env:WORKSPACE\
 C_2010_install.log"
```

- Microsoft .NET framework 4.0

  1. Use the instructions provided here to download Microsoft .NET framework 4.0 to the workspace directory.

  2. From PowerShell, as Administrator, enter:

```
 & "$env:WORKSPACE\NDP451-KB2858728-x86-x64-AllOS-ENU.exe" /q /norestart /
log "$env:WORKSPACE\NET-install_log.htm"
```

- JDK version 7

  1. Check which version of Java is currently installed. From a command shell or PowerShell window, enter:

     ```
     java -version
     ```

     If the JDK version is less than v1.6 update 31, uninstall the Java package.

  2. Go to the Oracle Java SE Downloads page and download the JDK installer to the workspace directory.

  3. From PowerShell, as Administrator, enter:

```
$key = "HKLM:\SYSTEM\CurrentControlSet\Control\Session Manager\
Environment”
$currentPath =(Get-ItemProperty -Path $key -name Path).Path +
';' $javaDir = "c:\java\jdk1.7.0_51\" &
"$env:WORKSPACE\jdk-7u51-windows-x64.exe" /qn /norestart /log
"$env:WORKSPACE\jdk-install.log"INSTALLDIR="c:\java
"ALLUSERS=1 setx JAVA_HOME "$javaDir" /m setx PATH
"$currentPath$javaDir\bin" /m
```

where WORKSPACE is an environment variable for the directory path where the installer is located and `c:\java\jdk1.7.0_51\` is the path where java will be installed. Ensure that no white space characters are present in the installation directory's path. For example, `c:\Program`[space]`Files` is not allowed.

4. Verify that Java installed correctly and that the Java application directory is in your PATH environment variable.

   From a command shell or PowerShell window, enter:

   ```
   java -version
   ```

   The system should display:

   ```
   java version "1.7.0_51"
   Java(TM) SE Runtime Environment (build 1.7.0_51-b18)
   Java HotSpot(TM) 64-Bit Server VM (build 24.51-b03, mixed mode)
   ```

1.3.9.2. Installing Required Software Manually

This section explains how to install the following software:

- Python
- Microsoft Visual C++ 2010 Redistributable Package (64 bit)
- Microsoft .NET framework 4.0
- Java JDK

Python

1. Download Python from here and install to a directory that contains no white space in the path, such as `c:\Python`.
2. Update the PATH environment variable using Administrator privileges:
   a. Open the `Control Panel -> System` pane and click on the `Advanced system settings` link.
   b. Click on the `Advanced` tab.
   c. Click the `Environment Variables` button.
   d. Under `System Variables`, find PATH and click `Edit`.
   e. After the last entry in the PATH value, enter a semi-colon and add the installation path to the Python installation directory, such as `;c:\Python27`.
   f. Click OK twice to close the Environment Variables dialog box.
   g. To validate your settings from a command shell or PowerShell window, type:

   ```
   python -V Python 2.7.6
   ```

Microsoft Visual C++ 2010 Redistributable Package (64-bit)

Download and install using the defaults.

Microsoft .NET Framework 4.0

Download and install using the defaults.

Oracle Java JDK

1. Download the Oracle JDK and install to a directory that contains no white space in the path, such as `c:\Java`.
2. Go to `Control Panel > System` and click `Advanced system settings`.
3. Click `Advanced`.
4. Click `Environment Variables`.
5. Add a new system environment variable, `JAVA_HOME`. The value for this variable should be the installation path for the Java Development Kit; for example, `c:\Java\jdk1.7.0_51`.
6. Click OK.
7. To validate the environment variable you just added, enter the following command at a command-line prompt:

   ```
   echo %JAVA_HOME%
   ```

   You should see the path you specified when you created `JAVA_HOME`:

   ```
   c:\Java\jdk1.7.0_45\
   ```

8. As Administrator, update the PATH variable:
   a. Under `System Variables`, find PATH. Click `Edit`.
   b. After the last entry in the Path value, enter a semi-colon and add the installation path to the JDK. For example:

      ```
      ...;c:\Java\jdk1.7.0_51\bin
      ```

   c. Click OK.
   d. To validate the change you just made, open a DOS command line and enter:

      ```
      java -version
      ```

      DOS should return the expected Java version and details; for example, `java version "1.7.0"`.
9. Click OK to close the `Environment Variables` dialog box.

2. Defining Cluster Properties

The Hortonworks Data Platform consists of multiple components that are installed across the cluster. The cluster properties file specified directory locations and node host names for each of the components. When you run the installer, it checks the host name against the properties file to determine which services to install.

After downloading the HDP installer, use one of the following methods to modify the cluster properties file:

- **Option 1:** Use the HDP Setup Interface to generate a cluster properties file for GUI to use, or to export a generated `clusterproperties.txt` file for a CLI installation. (Recommended for first-time users and single-node installations.)
- **Option 2:** Manually define cluster properties in a file. (Recommended for users who are familiar with their systems and with HDP requirements.)

2.1. Download the HDP Installer

Download the HDP Installation zip file and extract the files. The zip contains the following files:

- HDP MSI installer
- Sample `clusterproperties.txt` file
- Compression files:
  - `hadoop-lzo-0.4.19.2.2.0.0-2060`
  - `gplcompression.dll`
  - `lzo2.dll`

2.2. Using the HDP Setup Interface

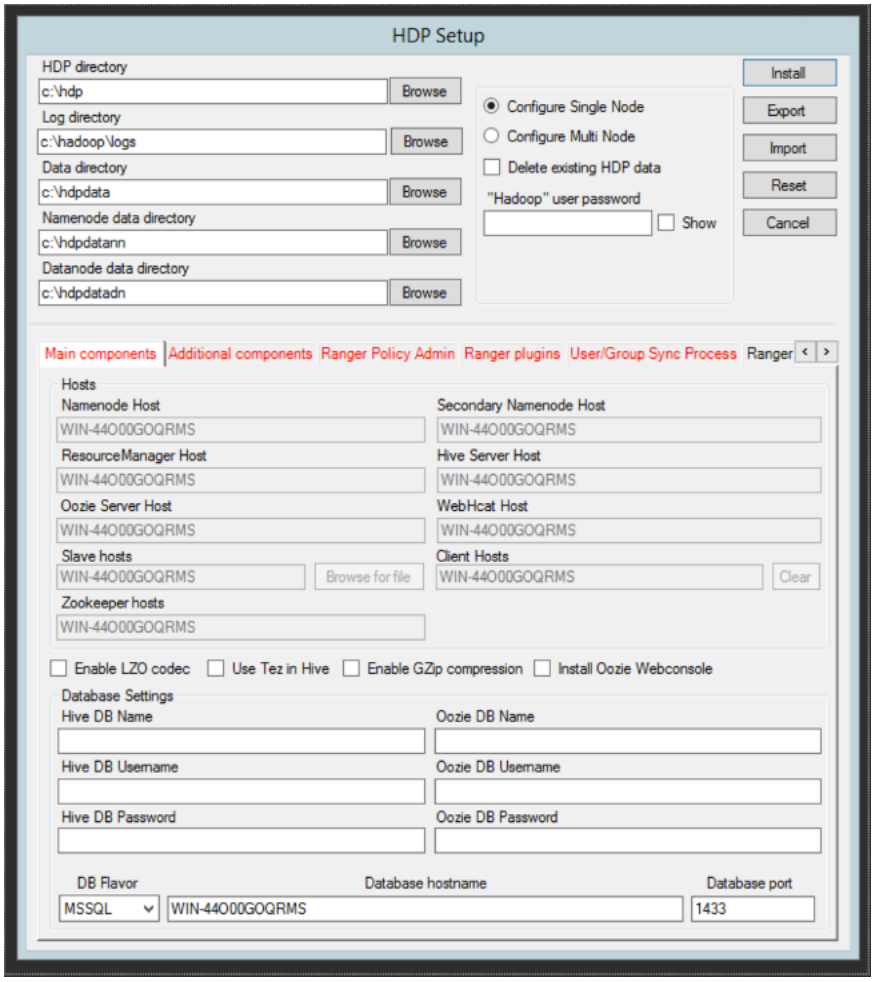
You can define cluster properties with the HDP Setup Interface or define them manually (described in the next subsection). If you use the HDP Setup Interface you can either export the configuration and use it to deploy HDP from the command line (or within a script), or you can start deployment from the Setup Interface itself.

To start the Setup Interface, enter the following command at the command prompt:

```
runas /user:administrator "msiexec /i hdp-2.3.0.0.winpkg.msi
 MSIUSEREALADMINDETECTION=1"
```

The HDP Setup form displays.

(The following image shows the form with the Main components tab selected.)



2.3. Set HDP Properties

The top part of the form, which includes HDP directory, Log directory, Data directory, Name node data directory and Data node data directory, is filled in with default values. Customize these entries as needed, and note whether you are configuring a single- or multi-node installation.

1. Complete the fields at the top of the HDP Setup form:

Table 2.1. Main component screen values

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
HDP directory	HDP installation directory	d:\hdp	Mandatory
Log directory	HDP's operational logs are written to this directory on each cluster host. Ensure that you have sufficient disk space for storing these log files.	d:\hadoop\logs	Mandatory
Data directory	HDP data will be stored in this directory on each cluster node. You can add	d:\hdp\data	Mandatory

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
	multiple comma-separated data locations for multiple data directories.		
Name node data directory	Determines where on the local file system the HDFS name node should store the name table (fsimage). You can add multiple comma-separated data locations for multiple data directories.	d:\hdpdata\hdfs	Mandatory
Data node data directory	Determines where on the local file system an HDFS data node should store its blocks. You can add multiple comma-separated data locations for multiple data directories.	d:\hdpdata\hdfs	Mandatory

2. To choose single- or multi-node deployment, select one of the following:

   - **Configure Single Node** – installs all cluster nodes on the current host; the host name fields are pre-populated with the name of the current computer. For information on installing on a single-node cluster, see the Quick Start Guide for Installing HDP for Windows on a Single-Node Cluster.
   - **Configure Multi Node** – creates a property file, which you can use for cluster deployment or to manually install a node (or subset of nodes) on the current computer.

3. Specify whether or not you want to delete existing HDP data.

   If you want to delete existing HDP data, select `Delete existing HDP data` and supply the `hadoop` user password in the field immediately below. (You can either shield the password while entering it or select `Show` to show it.)

4. **Mandatory**: Enter the password for the `hadoop` super user (the administrative user). This password enables you to log in as the administrative user and perform administrative actions. Password requirements are controlled by Windows, and typically require that the password include a combination of uppercase and lowercase letters, digits, and special characters.

5. Specify component-related values:

Table 2.2. Component configuration property information

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
NameNode Host	The FQDN for the cluster node that will run the NameNode master service.	NAMENODE_MASTER. acme.com	Mandatory
Secondary NameNode Host	The FQDN for the cluster node that will run the Secondary NameNode master service. (Not applicable for HA.)	SECONDARY_NN_MASTER. acme.com	Mandatory/NA

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
ResourceManager Host	The FQDN for the cluster node that will run the YARN Resource Manager master service.	RESOURCE_MANAGER. acme.com	Mandatory
Hive Server Host	The FQDN for the cluster node that will run the Hive Server master service.	HIVE_SERVER_MASTER. acme.com	Mandatory
Oozie Server Host	The FQDN for the cluster node that will run the Oozie Server master service.	OOZIE_SERVER_MASTER. acme.com	Mandatory
WebHCat Host	The FQDN for the cluster node that will run the WebHCat master service.	WEBHCAT_MASTER. acme.com	Mandatory
Slave hosts	A comma-separated list of FQDN for those cluster nodes that will run the DataNode and TaskTracker services.	slave1.acme.com, slave2.acme.com, slave3.acme.com	Mandatory
Client hosts	A comma-separated list of FQDN for those cluster nodes that will store JARs and other job-related files.	client.acme.com, client1.acme.com, client2.acme.com	Optional
Zookeeper Hosts	A comma-separated list of FQDN for those cluster nodes that will run the ZooKeeper hosts.	ZOOKEEPER-HOST.acme.com	Mandatory
Enable LZO codec	Use LZO compression for HDP.	Selected	Optional
Use Tez in Hive	Install Tez on the Hive host.	Selected	Optional
Enable GZip compression	Enable gzip file compression.	Selected	Optional
Install the Oozie Web console	Install web-based console for Oozie. The Oozie Web console requires the ext-2.2.zip file.	Selected	Optional

6. Enter database information for Hive and Oozie at the bottom of the form:

Table 2.3. Hive and Oozie configuration property information

Configuration Property Name	Description	Example Value	Mandatory/Optional
Hive DB Name	The name of the database used for Hive.	hivedb	Mandatory
Hive DB User name	User account credentials for Hive metastore database instance. Ensure that this user account has appropriate permissions.	hive_user	Mandatory
Hive DB Password	User account credentials for Hive metastore database instance. Ensure that this user account has appropriate permissions.	hive_pass	Mandatory

Configuration Property Name	Description	Example Value	Mandatory/Optional
Oozie DB Name	Database for Oozie metastore. If using SQL Server, ensure that you create the database on the SQL Server instance.	ooziedb	Mandatory
Oozie DB User name	User account credentials for Oozie metastore database instance. Ensure that this user account has appropriate permissions.	oozie_user	Mandatory
Oozie DB Password	User account credentials for Oozie metastore database instance. Ensure that this user account has appropriate permissions.	oozie_pass	Mandatory
DB Flavor	Database type for Hive and Oozie metastores (allowed databases are SQL Server and Derby). To use default embedded Derby instance, set the value of this property to derby. To use an existing SQL Server instance as the metastore DB, set the value as mssql.	msql or derby	Mandatory
Database host name	FQDN for the node where the metastore database service is installed. If using SQL Server, set the value to your SQL Server host name. If using Derby for Hive metastore, set the value to HIVE_SERVER_HOST.	sqlserver1.acme.com	Mandatory
Database port	This is an optional property required only if you are using SQL Server for Hive and Oozie metastores. By default, the database port is set to 1433.	1433	Optional

7. To install HBase, Falcon Knox, Storm, Flume, Spark, Phoenix, Slider, Ranger, DataFu or HiveDR, click the `Additional componentstab`, and complete the fields as shown in the table below:

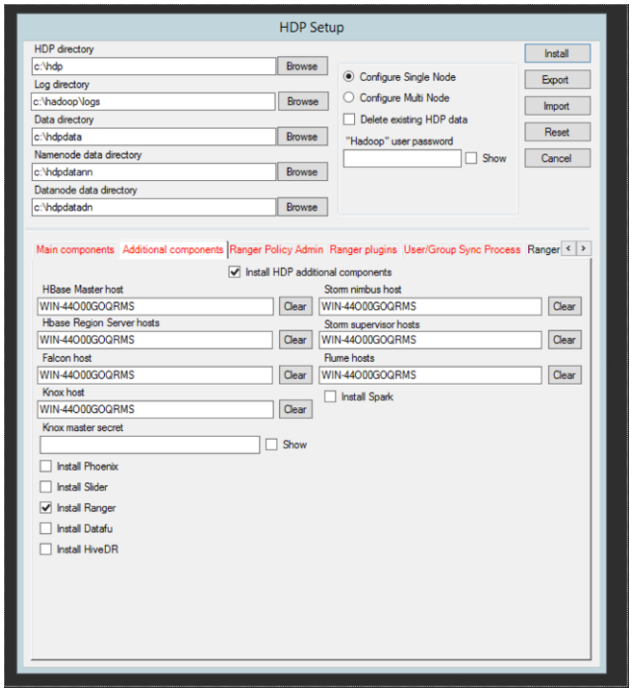


Table 2.4. Additional components screen values

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
HBase Master host	The FQDN for the cluster node that runs the HBase master	HBASE-MASTER.acme.com	Mandatory
Storm nimbus host	The FQDN for the cluster node that runs the Storm Nimbus master service	storm-host.acme.com	Optional
HBase region Server hosts	A comma-separated list of FQDN for cluster nodes that run the HBase Region Server services	slave1.acme.com, slave2.acme.com, slave3.acme.com	Mandatory
Storm supervisor hosts	A comma-separated list of FQDN for those cluster nodes that run the Storm Supervisors.	storm-sup=host.acme.com	Optional
Falcon host	The FQDN for the cluster node that runs Falcon	falcon-host.acme.com	
Flume hosts	A comma-separated list of FQDN for cluster nodes that run Flume	flume-host.acme.com	Optional
Install Spark	Indicates that you want to install Spark	Check box selected	Optional
Spark job history server	Specifies the Spark job history server	spark-host.acme.com	Optional
Spark hive metastore	Specifies the Spark hive metastore value	metastore	Optional
Knox host	The FQDN for the cluster node that runs Knox	knox-host.acme.com	Mandatory

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
Knox Master secret	Password for starting and stopping the gateway	knox-secret	Mandatory
Install Phoenix	Install Phoenix on the HBase server	Selected	Optional
Install Slider	Install Slider platform services for the YARN environment	Selected	Optional
Install Ranger	Installs Ranger security	Selected	Optional
Install DataFu	Install DataFu user-defined functions for data analysis	Selected	Optional
Install HiveDR	Installs HiveDR	Check box selected	Optional

2.4. Set High Availability Properties

To ensure that a multi-node cluster remains available, configure and enable High Availability. The configuration process for High Availability includes defining locations and names of hosts in a cluster that are available to act as journal nodes, and a standby name node in the event that the primary name node fails. To configure name node High Availability, select the HA `components` tab. Define the locations and names of hosts in a cluster that are available to act as JournalNodes and the Resource Manager. Specify a standby name node in case the primary name node fails.

To enable High Availability, you must also run several commands while starting cluster services.

Table 2.5. High Availability configuration property information

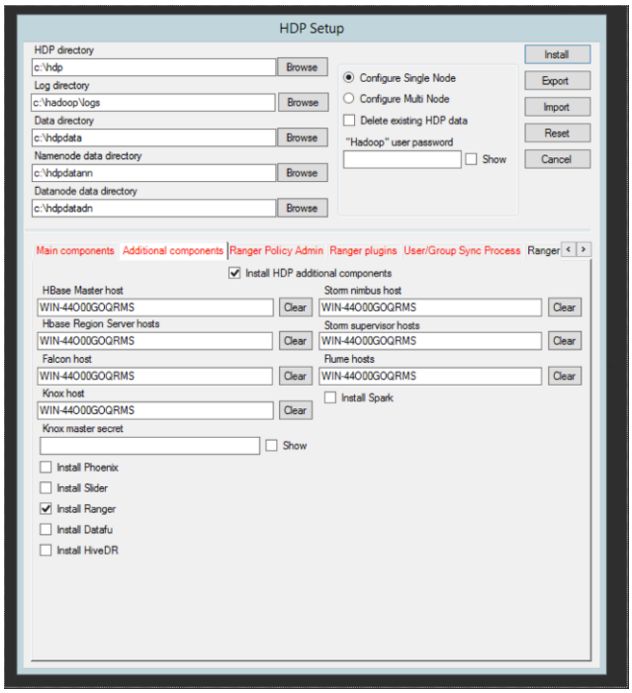
Configuration Property Name	Description	Example Value	Mandatory/Optional
Enable HA	Whether to deploy a highly available NameNode or not.	yes or no	Optional
NN Journal Node Hosts	A comma-separated list of FQDN for cluster nodes that will run the JournalNode processes.	journalnode1.acme.com, journalnode2.acme.com, journalnode3.acme.com	Optional
NN HA Cluster Name	This name is used for both configuration and authority component of absolute HDFS paths in the cluster.	hdp2-ha	Optional
NN Journal Node Edits Directory	This is the absolute path on the JournalNode machines where the edits and other local state used by the JournalNodes (JNs) are stored. You can only use a single path for this configuration.	d:\hadoop\journal	Optional
NN Standby Host	The host for the standby NameNode.	STANDBY_NAMENODE. acme.com	Optional
RM HA Cluster Name	A logical name for the Resource Manager cluster.	HA Resource Manager	Optional
RM Standby Host	The FQDN of the standby resource manager host.	rm-standby-host.acme.com	Optional

Configuration Property Name	Description	Example Value	Mandatory/Optional
Enable Knox HA	Enable Knox High Availability mode	Selected	Optional
Enable Oozie HA	Enable Oozie High Availability mode	Selected	Optional
Oozie server hosts	A comma-separated list of FQDN for cluster nodes that will run Oozie servers.	ooziehost1.acme.com, ooziehost2.acme.com	Optional
Oozie load balancer URL	URL for Oozie Load Balancer	http://oozie.lb.com:11000/oozie	Optional

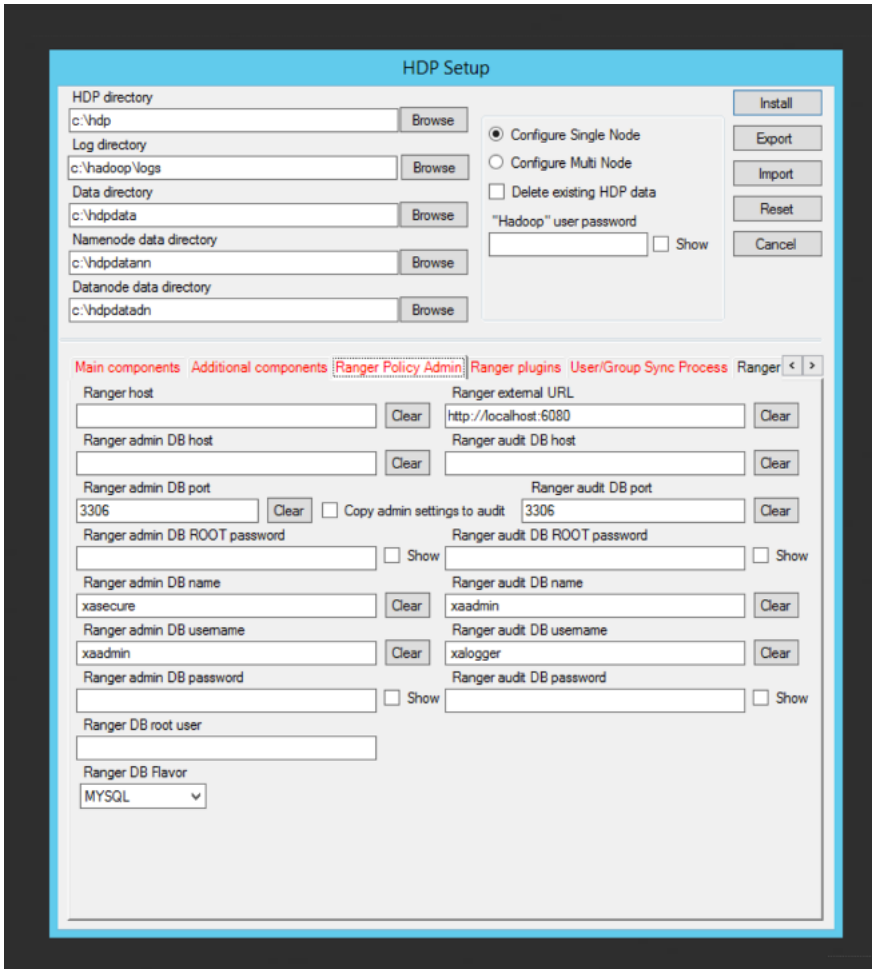
2.5. Set Ranger Properties

(**Optional**) To configure Ranger using the Setup GUI, complete the following steps.

1. Enable Ranger from the `Additional components` tab.



2. Click the `Ranger Policy Admin` tab in the middle of the HDP Setup Form.



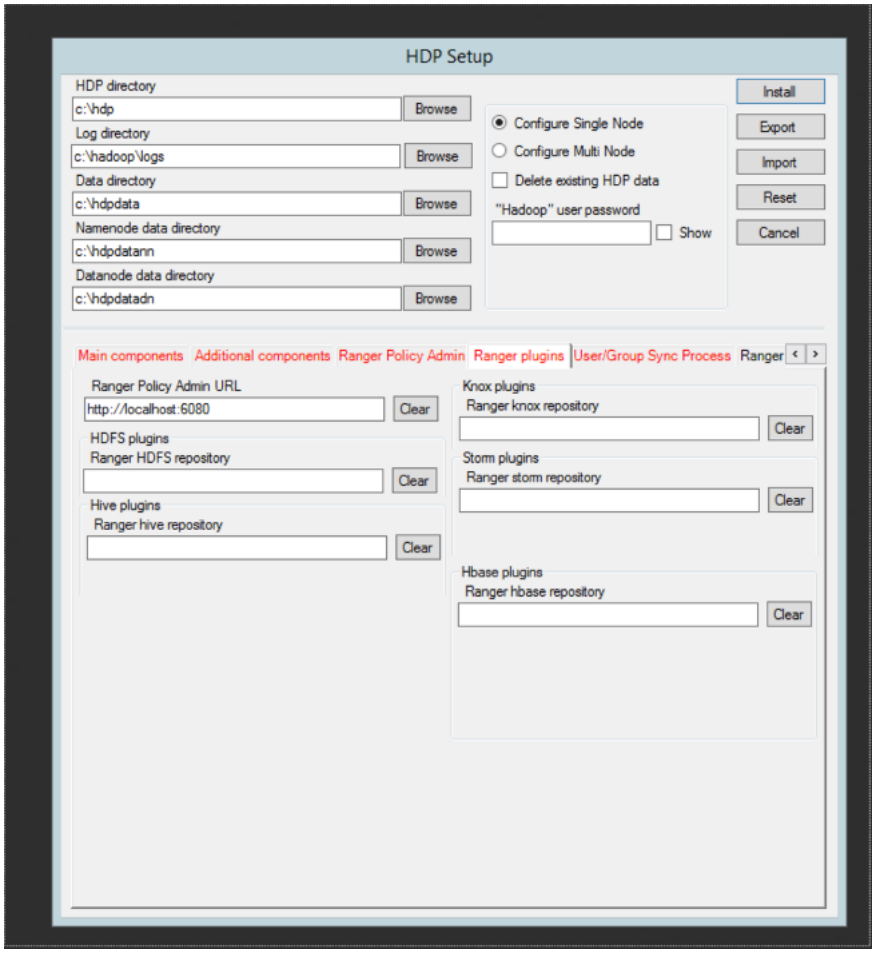
3. Enter host information, credentials for database saving policies, Admin user credentials, and Audit user credentials.

Table 2.6. Ranger Policy Admin screen values

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
Ranger host	Host name of the host where Ranger-Admin and Ranger-UserSync services will be installed	WIN-Q0EOPEACTR1	Mandatory
Ranger external URL	URL used for Ranger	http://localhost:6080	Mandatory
Ranger admin DB host	MySQL server instance for use by the Ranger Admin database host. (MySQL should be up and running at installation time.)	localhost	Mandatory
Ranger admin DB port	Port number for Ranger-Admin database server	3306	Mandatory
Ranger admin DB ROOT password	Database password for the Ranger admin DB user name	RangerAdminPassW0rd	Mandatory

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
Ranger admin DB name	Ranger-Admin policy database name	ranger (default)	Mandatory
Ranger admin DB user name	Ranger-Admin policy database user name	rangeradmin (default)	Mandatory
Ranger admin DB password	Password for the Ranger admin DB user	RangerAdminPassW0Rd	Mandatory
Copy admin settings to audit	Use admin settings for audit database	Selected	
Ranger audit DB host	Host for Ranger Audit database. (MySQL should be up and running at installation time). This can be the same as the Ranger host, or you can specify a different server.	localhost	Mandatory
Ranger audit DB name	Ranger audit database name. This can be a different database in the same database server mentioned above.	ranger_audit (default)	Mandatory
Ranger audit DB port	Port number where Ranger-Admin runs audit service	3306	Mandatory
Ranger audit DB ROOT password	Database password for the Ranger audit DB user name (required for audit database creation)	RangerAuditPassW0Rd	Mandatory
Ranger audit DB user name	Database user that performs all audit logging operations from Ranger plugins	rangerlogger (default)	Mandatory
Ranger audit DB password	Database password for the Ranger audit DB user name	RangerAuditPassW0Rd	Mandatory

4. Click the `Ranger Plugins` tab in the middle of the HDP Setup Form.



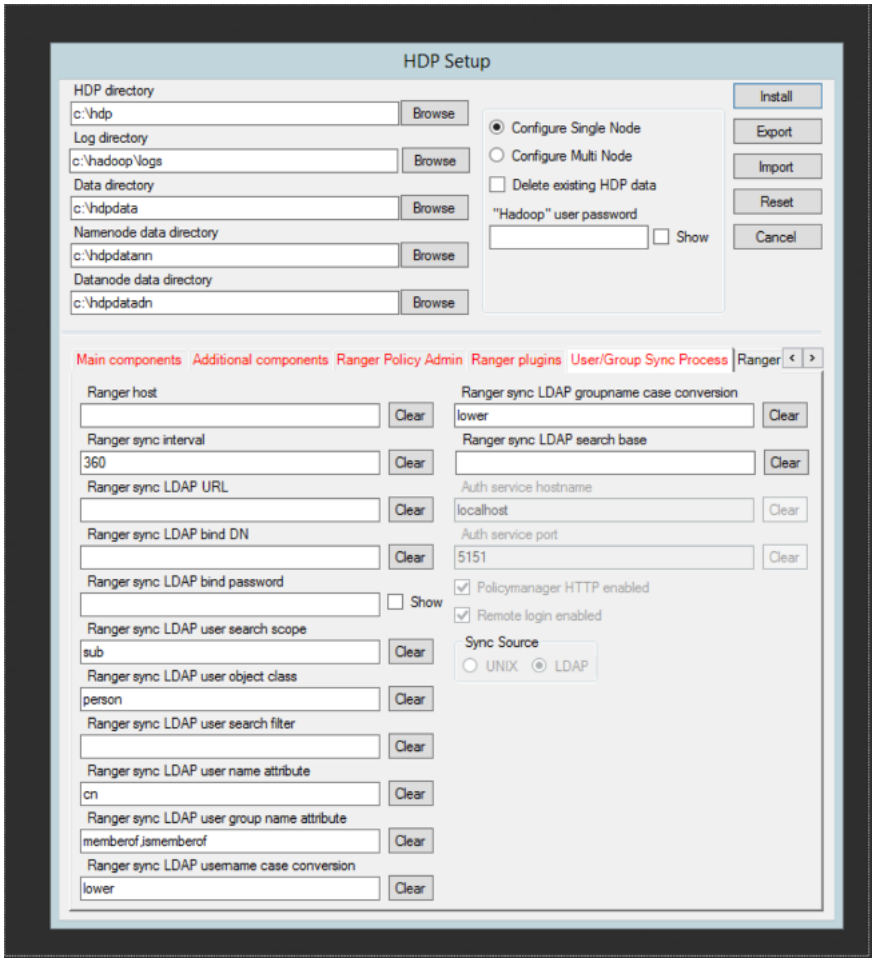
5. Complete the following fields. These allow communication between Ranger-Admin and each plugin.

Table 2.7. Ranger Plugins screen values

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
Ranger Policy Admin URL	URL used within policy admin tool when a link to its own page is generated in the policy admin tool website	http://localhost:6080	Mandatory
Knox agents: Ranger Knox repository	The repository name used in Policy Admin Tool for defining policies for Knox	knoxdev	Mandatory if using Ranger on Knox
HDFS agents: Ranger HDFS repository	The repository name used in Policy Admin Tool for defining policies for HDFS	hadoopdev	Mandatory if using Ranger on HDFS
Storm agents: Ranger storm repository	The repository name used in Policy Admin Tool for defining policies for Storm	stormdev	Mandatory if using Ranger on Storm
Hive agents: Ranger hive repository	The repository name used in Policy Admin Tool for defining policies for Hive	hivedev	Mandatory if using Ranger on Hive

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
HBase agents: Ranger hbase repository	The repository name used in Policy Admin Tool for defining policies for HBase	hbasedev	Mandatory if using Ranger on HBase

6. Click the `User/Group Sync Process` tab in the middle of the HDP Setup Form.



7. Complete the following fields.

   a. Add the Ranger-Admin host URL to Ranger User/Group Sync; this enables communication between Ranger-Admin and the User-Sync service.

   b. Set appropriate values for the other parameters based on sync source:

      - If users will be synchronized from an LDAP server, supply LDAP server credentials and all properties associated with synchronizing users and groups from the LDAP server.

      - If users will be synchronized with an Active Directory, supply Active Directory credentials and all properties associated with synchronizing users and groups via Active Directory.

Table 2.8. User/Group Sync Process screen field values

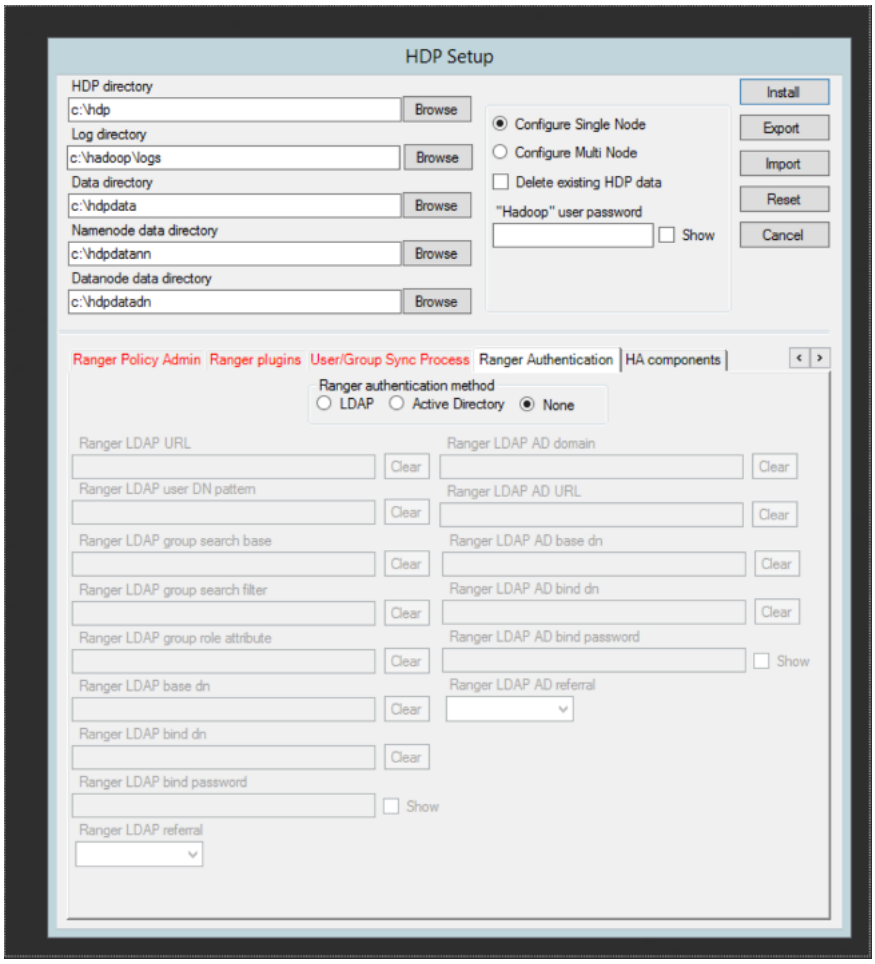
Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
Ranger host	host name of the host where Ranger-Admin and Ranger-UserSync services will be installed	WIN-Q0EOPEACTR1	Mandatory
Ranger sync interval	Specifies the interval (in minutes) between synchronization cycles. Note: the second sync cycle will NOT start until the first sync cycle is complete.	5	Mandatory
Ranger sync LDAP search base	Search base for users	ou=users, dc=hadoop, dc=apache, dc=org	Mandatory
Ranger sync LDAP URL	LDAP URL for synchronizing users	ldap:// ldap.example.com:389	Mandatory
Ranger sync LDAP bind DN	LDAP bind DN used to connect to LDAP and query for users and group. This must be a user with admin privileges to search the directory for users/groups.	cn=admin,ou=users, dc=hadoop,dc=apache, dc-org	Mandatory
Ranger sync LDAP bind password	Password for the LDAP bind DN	LdapAdminPassW0Rd	Mandatory
Ranger sync LDAP user search scope	Scope for user search	base, one, and sub are supported values	Mandatory
Ranger sync LDAP user object class	Object class to identify user entries	person (default)	Mandatory
Ranger sync LDAP user search filter	Additional filter constraining the users selected for syncing	[objectcategory=person]	Optional
Ranger sync LDAP user name attribute	Attribute from user entry that will be treated as user name	cm (default)	Mandatory
Ranger sync LDAP user group name attribute	Attribute from user entry whose values will be treated as group values to be pushed into the Policy Manager database.	One or more attribute names separated by commas, such as: member of,ismemberof	Mandatory
Ranger sync LDAP user name case conversion	Convert all user names to lowercase or uppercase	none: no conversion; keep as-is in SYNC_SOURCE. lower: (default) convert to lowercase when saving user names to the Ranger database. upper: convert to uppercase when saving user names to the Ranger db.	Mandatory
Ranger sync LDAP group name case conversion	Convert all group names to lowercase or uppercase	(same as user name case conversion)	Mandatory

8. After specifying Ranger-UserSync properties, make sure that the following properties are defined on other tabs:

   - On the Additional Components tab, set the Ranger authentication method to LDAP, Active Directory, or None, based on your synchronization source.

- On the Ranger Policy Admin tab, make sure that you have specified Authentication Properties.

9. Click the `Ranger Authentication` tab in the middle of the HDP Setup Form.



10. Specify whether you want to use LDAP or Active Directory Ranger authentication and complete the fields pertaining to your choice.

Table 2.9. Ranger Authentication screen field values for LDAP authentication

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
Ranger LDAP URL	Specifies the LDAP Server URL	ldap://10.129.86.185:10389	Mandatory
Ranger LDAP user DN pattern	The user distinguished name (DN) pattern is expanded when a user is logging in. For example, if the user `ldapadmin` attempts to log in, the LDAP Server attempts to bind against the DN `uid=ldapadmin,ou=users,dc=example,dc=com`,	cn=(0),ou=users,o=example	Mandatory

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
	and uses the password user `ldapadmin` provides.		
Ranger LDAP group search base	Defines the part of the directory under which you want group searches to be performed.	o=example	Mandatory
Ranger LDAP group search filter	Defines the filter you want to use to search for group membership. The default is `uniqueMember={0}`, corresponding to the `groupOfUniqueNames` LDAP class. For Ranger authentication, the substituted parameter is the full, distinguished name of the user. You can use parameter `{0}` if you want to filter on the login name.	(member=cn=(0),ou=users,o=example)	Mandatory
Ranger LDAP group role attribute	Specifies the attribute that contains the name of the authority defined by the group entry.	cn	Mandatory
Ranger LDAP base dn	Specifies the DN of the starting point for your directory server searches.	o=example	Mandatory
Ranger LDAP bind dn	Specifies the full DN, including the common name (CN), of the LDAP user account that has privileges to search for users.	cn=admin,ou=users,o=freestone	Mandatory
Ranger LDAP bind password	Specifies the password for the account that can search for users.	RangerLDAPBindPassW0rd	Mandatory
Ranger LDAP referral	Defines search result processing behavior. Possible values are follow, ignore, and throw.	follow	Mandatory

Table 2.10. Ranger Authentication screen field values for Active Directory authentication

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
Ranger LDAP AD domain	LDAP Server domain name using a `prefix.suffix` format.	example.com	Mandatory
Ranger LDAP AD URL	Specifies the LDAP Server URL.	ldap://10.129.86.200:389	Mandatory
Ranger LDAP AD base dn	Specifies the distinguished name (DN) of the starting point for directory server searches.	dc=example,dc=local	Mandatory
Ranger LDAP AD bind dn	Specifies the full DN, including common name (CN), of an Active Directory user account that has	cn=Administrator,cn=Users,dc=example,dc=local	Mandatory

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
	privileges to search for users. This user account must have at least domain user privileges.		
Ranger LDAP AD bind password	Specifies the password for the account that can search for users.	Ranger_LDAP_AD_Bind_Password	Mandatory
Ranger LDAP AD referral	Defines search result processing behavior. Possible values are follow, ignore, and throw.	follow	Mandatory

2.6. Complete the GUI Installation Process

To continue with the GUI installation process, select `Install`. To clear all fields and start over again, select `Reset`. To export your HDP Setup configuration as a `clusterproperties.txt` file and switch to the CLI installation process, select `Export`. Export stops the GUI installation process and produces a `clusterproperties.txt` file based on your GUI fields. Before exporting, verify that all information is accurate.

2.7. Manually Creating a Cluster Properties File

Use the following instructions to manually configure the cluster properties file for deploying HDP from the command-line interface or in a script.

1. Create a file for the cluster properties, or use the sample `clusterproperties.txt` file extracted from the HDP Installation zip file. You'll pass the name of the cluster properties file to the `msiexec` call when you install HDP. The following examples use the file name `clusterproperties.txt`.
2. Add the properties to the `clusterproperties.txt` file as described in the table below. As you add properties, keep in mind the following:
   - All properties in the cluster properties file must be separated by a newline character.
   - Directory paths cannot contain white space characters. (For example, `c:\Program Files\Hadoop` is an invalid directory path for HDP.)
   - Use Fully Qualified Domain Names (FQDN) to specify the network host name for each cluster host.

     The FQDN is a DNS name that uniquely identifies the computer on the network. By default, it is a concatenation of the host name, the primary DNS suffix, and a period.
   - When specifying the host lists in the cluster properties file, if the hosts are multi-homed or have multiple NIC cards, make sure that each name or IP address is the preferred name or IP address by which the hosts can communicate among themselves. In other words, these should be the addresses used internal to the cluster, not those used for addressing cluster nodes from outside the cluster.
   - To Enable NameNode HA, you must include the HA properties and exclude the SECONDARY_NAMENODE_HOST definition.

Table 2.11. Configuration Values for Deploying HDP

Configuration Property Name	Description	Example Value	Mandatory/Optional
HDP_LOG_DIR	HDP's operational logs are written to this directory on each cluster host. Ensure that you have sufficient disk space for storing these log files.	d:\hadoop\logs	Mandatory
HDP_DATA_DIR	HDP data will be stored in this directory on each cluster node. You can add multiple comma-separated data locations for multiple data directories.	d:\hdp\data	Mandatory
HDFS_NAMENODE_ DATA_DIR	Determines where on the local file system the HDFS name node should store the name table (fsimage). You can add multiple comma-separated data locations for multiple data directories.	d:\hadoop\data\hdfs\nn,c:\hdpdata,d:\hdpdatann	Mandatory
HDFS_DATANODE_ DATA_DIR	Determines where on the local file system an HDFS data node should store its blocks. You can add multiple comma-separated data locations for multiple data directories.	d:\hadoop\data\hdfs\dn,c:\hdpdata,d:\hdpdatadn	Mandatory
NAMENODE_HOST	The FQDN for the cluster node that will run the NameNode master service.	NAMENODE-MASTER.acme.com	Mandatory
SECONDARY_NAMENODE_ HOST	The FQDN for the cluster node that will run the Secondary NameNode master service.	SECONDARY-NN-MASTER.acme.com	Mandatory when no HA
RESOURCEMANAGER_HOST	The FQDN for the cluster node that will run the YARN Resource Manager master service.	RESOURCE-MANAGER.acme.com	Mandatory
HIVE_SERVER_HOST	The FQDN for the cluster node that will run the Hive Server master service.	HIVE-SERVER-MASTER.acme.com	Mandatory
OOZIE_SERVER_HOST	The FQDN for the cluster node that will run the Oozie Server master service.	OOZIE-SERVER-MASTER.acme.com	Mandatory
WEBHCAT_HOST	The FQDN for the cluster node that will run the WebHCat master service.	WEBHCAT-MASTER.acme.com	Mandatory
FLUME_HOSTS	A comma-separated list of FQDN for those cluster nodes that will run the Flume service.	FLUME-SERVICE1.acme.com, FLUME-SERVICE2.acme.com, FLUME-SERVICE3.acme.com	Mandatory
HBASE_MASTER	The FQDN for the cluster node that will run the HBase master.	HBASE-MASTER.acme.com	Mandatory
HBASE_REGIONSERVERS	A comma-separated list of FQDN for those cluster nodes that will run the	slave1.acme.com, slave2.acme.com, slave3.acme.com	Mandatory

Configuration Property Name	Description	Example Value	Mandatory/Optional
	HBase Region Server services.		
SLAVE_HOSTS	A comma-separated list of FQDN for those cluster nodes that will run the DataNode and TaskTracker services.	slave1.acme.com, slave2.acme.com, slave3.acme.com	Mandatory
ZOOKEEPER_HOSTS	A comma-separated list of FQDN for those cluster nodes that will run the ZooKeeper hosts.	ZOOKEEEPER-HOST.acme.com	Optional
FALCON_HOST	A comma-separated list of FQDN for those cluster nodes that will run the Falcon hosts.	falcon.acme.com, falcon1.acme.com, falcon2.acme.com	Optional
KNOX_HOST	The FQDN of the Knox Gateway host.	KNOX-HOST.acme.com	Optional
STORM_SUPERVISORS	A comma-separated list of FQDN for those cluster nodes that will run the Storm Supervisor hosts.	supervisor.acme.com, supervisor1.acme.com, supervisor2.acme.com	Optional
STORM_NIMBUS	The FQDN of the Storm Nimbus Server.	STORM-HOST.acme.com	Optional
DB_FLAVOR	Database type for Hive and Oozie metastores (allowed databases are SQL Server and Derby). To use default embedded Derby instance, set the value of this property to derby. To use an existing SQL Server instance as the metastore DB, set the value as mssql.	mssql or derby	Mandatory
DB_PORT	Port address, required only if you are using SQL Server for Hive and Oozie metastores.	1433 (default)	Optional
DB_HOSTNAME	FQDN for the node where the metastore database service is installed. If using SQL Server, set the value to your SQL Server host name. If using Derby for Hive metastore, set the value to HIVE_SERVER_HOST.	sqlserver1.acme.com	Mandatory
HIVE_DB_NAME	Database for Hive metastore. If using SQL Server, ensure that you create the database on the SQL Server instance.	hivedb	Mandatory
HIVE_DB_USERNAME	User account credentials for Hive metastore database instance. Ensure that this user account has appropriate permissions.	hive_user	Mandatory
HIVE_DB_PASSWORD	User account credentials for Hive metastore database instance. Ensure that	hive_pass	Mandatory

Configuration Property Name	Description	Example Value	Mandatory/Optional
	this user account has appropriate permissions.		
OOZIE_DB_NAME	Database for Oozie metastore. If using SQL Server, ensure that you create the database on the SQL Server instance.	ooziedb	Mandatory
OOZIE_DB_USERNAME	User account credentials for Oozie metastore database instance. Ensure that this user account has appropriate permissions.	oozie_user	Mandatory
OOZIE_DB_PASSWORD	User account credentials for Oozie metastore database instance. Ensure that this user account has appropriate permissions.	oozie_pass	Mandatory
DEFAULT_FS	Default file system.	HDFS	
RESOURCEMANAGER_HOST	Host used for Resource Manager		
IS_TEZ	Installs the Tez component on Hive host.	YES or NO	Optional
ENABLE_LZO	Enables the LZO codec for compression in HBase cells.	YES or NO	Optional
IS_PHOENIX	Installs Phoenix on the HBase hosts.	YES or NO	Optional
IS_HDFS_HA	Specify whether to enable High Availability for HDFS	YES or NO	Mandatory
SPARK_JOB_SERVER	Specifies the Spark job history server	onprem-ranger1	Optional
SPARK_HIVE_METASTORE	Specifies the Hive metastore for Spark	metastore	Optional
HIVE_DR	Indicates whether you want to install HiveDR	YES or NO	Optional

Configuration Values: High Availability

To ensure that a multi-node cluster remains available, configure and enable High Availability. Configuring High Availability includes defining locations and names of hosts in a cluster that are available to act as journal nodes and a standby name node in the event that the primary name node fails. To configure High Availability, add the following properties to your cluster properties file, and set their values as follows:

Note

To enable High Availability, you must also run several HA-specific commands when you start cluster services.

Table 2.12. High Availability configuration information

Configuration Property Name	Description	Example Value	Mandatory/Optional
HA	Whether to deploy a highly available NameNode or not.	yes or no	Optional

Configuration Property Name	Description	Example Value	Mandatory/Optional
NN_HA_JOURNALNODE_HOSTS	A comma-separated list of FQDN for those cluster nodes that will run the JournalNode processes.	journalnode1.acme.com, journalnode2.acme.com, journalnode3.acme.com	Optional
NN_HA_CLUSTER_NAME	This name is used for both configuration and authority component of absolute HDFS paths in the cluster.	hdp2-ha	Optional
NN_HA_JOURNALNODE_EDITS_DIR	This is the absolute path on the JournalNode machines where the edits and other local state used by the JournalNodes (JNs) are stored. You can only use a single path for this configuration.	d:\hadoop\journal	Optional
NN_HA_STANDBY_NAMENODE_HOST	The host for the standby NameNode.	STANDBY_NAMENODE.acme.com	Optional
RM_HA_CLUSTER_NAME	A logical name for the Resource Manager cluster.	HA Resource Manager	Optional
RM_HA_STANDBY_RESOURCEMANAGER_HOST	The FQDN of the standby resource manager host.	rm-standby-host.acme.com	Optional

Configuration Values: Ranger

Note

"Mandatory" means that the property must be specified if Ranger is enabled.

Table 2.13. Ranger configuration information

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
RANGER_HOST	Host name of the host where Ranger-Admin and Ranger-UserSync services will be installed	WIN-Q0E0PEACTR	Mandatory
RANGER_ADMIN_DB_HOST	MySQL server instance for use by the Ranger Admin database host. (MySQL should be up and running at installation time.)	localhost	Mandatory
RANGER_ADMIN_DB_PORT	Port number for Ranger-Admin database server	3306	Mandatory
RANGER_ADMIN_DB_ROOT_PASSWORD	Database root password (required for policy/audit database creation)	adm2	Mandatory
RANGER_ADMIN_DB_DBNAME	Ranger-Admin policy database name	ranger (default)	Mandatory
RANGER_ADMIN_DB_USERNAME	Ranger-Admin policy database user name	rangeradmin (default)	Mandatory
RANGER_ADMIN_DB_PASSWORD	Password for the RANGER_ADMIN_DB_USERNAME user	RangerAdminPassW0Rd	Mandatory
RANGER_AUDIT_DB_HOST	Host for Ranger Audit database. (MySQL should	localhost	Mandatory

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
	be up and running at installation time). This can be the same as RANGER_ADMIN_DB_HOST or you can specify a different server.		
RANGER_AUDIT_DB_PORT	Port number where Ranger-Admin runs audit service	3306	Mandatory
RANGER_AUDIT_DB_ROOT_PASSWORD	Database password for the RANGER_AUDIT_DB_USERNAME (required for audit database creation)	RangerAuditPassW0Rd	Mandatory
RANGER_EXTERNAL_URL	URL used for Ranger	localhost:8080	Optional
RANGER_AUDIT_DB_ DBNAME	Ranger audit database name. This can be a different database in the same database server mentioned above.	ranger_audit (default)	Mandatory
RANGER_AUDIT_DB_ USERNAME	Database user that performs all audit logging operations from Ranger plugins	rangerlogger (default)	Mandatory
RANGER_AUDIT_DB_ PASSWORD	Database password for the RANGER_AUDIT_DB_USERNAME user	RangerAuditPassW0Rd	Mandatory
RANGER_AUTHENTICA-TION_METHOD	Authentication Method used to login into the Policy Admin Tool.	None: allows only users created within Policy Admin Tool (default) LDAP: allows users to be authenticated using Corporate LDAP. AD: allows users to be authenticated using a Active Directory.	Mandatory
RANGER_LDAP_URL	URL for the LDAP service	ldap://71.127.43.33:386	Mandatory if authentication method is LDAP
RANGER_LDAP_ USERDNPATTERN	LDAP DN pattern used to locate the login user (uniquely)	uid={0},ou=users,dc=ranger2, dc=net	Mandatory if authentication method is LDAP
RANGER_LDAP_ GROUPSEARCHBASE	Defines the part of the LDAP directory tree under which group searches should be performed	ou=groups,dc=ranger2, dc=net	Mandatory if authentication method is LDAP
RANGER_LDAP_ GROUPSEARCHFILTER	LDAP search filter used to retrieve groups for the login user	(member=uid={0},ou=users, dc=ranger2,dc=net)	Mandatory if authentication method is LDAP
RANGER_LDAP_ GROUPROLEATTRIBUTE	Contains the name of the authority defined by the group entry, used to retrieve the group names from the group search filters	cn	Mandatory if authentication method is LDAP
RANGER_LDAP_AD_ DOMAIN	Active Directory Domain Name used for AD login	rangerad.net	Mandatory if authentication method is Active Directory
RANGER_LDAP_AD_URL	Active Directory LDAP URL for authentication of user	ldap://ad.rangerad.net:389	Mandatory if authentication method is Active Directory
RANGER_POLICY_ADMIN _URL	URL used within policy admin tool when a link to its own page is generated	localhost:6080	Optional

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
	in the policy admin tool website		
RANGER_HDFS_REPO	The repository name used in Policy Admin Tool for defining policies for HDFS	hadoopdev	Mandatory if using Ranger on HDFS
RANGER_HIVE_REPO	The repository name used in Policy Admin Tool for defining policies for Hive	hivedev	Mandatory if using Ranger on Hive
RANGER_HBASE_REPO	The repository name used in Policy Admin Tool for defining policies for HBase	hbasedev	Mandatory if using Ranger on HBase
RANGER_KNOX_REPO	The repository name used in Policy Admin Tool for defining policies for Knox	knoxdev	Mandatory if using Ranger on Knox
RANGER_STORM_REPO	The repository name used in Policy Admin Tool for defining policies for Storm	stormdev	Mandatory if using Ranger on Storm
RANGER_SYNC_INTERVAL	Specifies the interval (in minutes) between synchronization cycles. Note: the second sync cycle will NOT start until the first sync cycle is complete.	5	Mandatory
RANGER_SYNC_LDAP_URL	LDAP URL for synchronizing users	ldap:// ldap.example.com:389	Mandatory
RANGER_SYNC_LDAP_ BIND_DN	LDAP bind DN used to connect to LDAP and query for users and group. This must be a user with admin privileges to search the directory for users/groups.	cn=admin,ou=users, dc=hadoop,dc=apache, dc-org	Mandatory
RANGER_SYNC_LDAP_ BIND_PASSWORD	Password for the LDAP bind DN	LdapAdminPassW0Rd	Mandatory
RANGER_SYNC_LDAP_ USER_SEARCH_SCOPE	Scope for user search	base, one and sub are supported values	Mandatory
RANGER_SYNC_LDAP_ USER_OBJECT_CLASS	Object class to identify user entries	person (default)	Mandatory
RANGER_SYNC_LDAP_ USER_NAME_ATTRIBUTE	Attribute from user entry that will be treated as user name	cn (default)	Mandatory
RANGER_SYNC_LDAP_ USER_GROUP_NAME _ATTRIBUTE	Attribute from user entry whose values will be treated as group values to be pushed into the Policy Manager database.	One or more attribute names separated by commas, such as: memberof,ismemberof	Mandatory
RANGER_SYNC_LDAP_ USERNAME_CASE _CONVERSION	Convert all user names to lowercase or uppercase	none: no conversion; keep as-is in SYNC_SOURCE. lower: (default) convert to lowercase when saving user names to the Ranger database. upper: convert to uppercase when saving user names to the Ranger db.	Mandatory
RANGER_SYNC_LDAP_ GROUPNAME_CASE _CONVERSION	Convert all group names to lowercase or uppercase	(same as user name case conversion property)	Mandatory

Configuration Property Name	Description	Example Value	Mandatory/Optional/ Conditional
RANGER_SYNC_LDAP_ USER_SEARCH_BASE	Search base for users	ou=users,dc=hadoop, dc=apache,dc=org	Mandatory
AUTHSERVICEHOSTNAME	Server Name (or IP address) where Ranger-Usersync module is running (along with Unix Authentication Service)	localhost (default)	Mandatory
AUTHSERVICEPORT	Port Number where Ranger-Usersync module is running the Unix Authentication Service	5151 (default)	Mandatory
POLICYMGR_HTTP_ENABLED	Flag to enable/disable HTTP protocol for downloading policies by Ranger plugin modules	true (default)	Mandatory
REMOTELOGINENABLED	Flag to enable/disable remote Login via Unix Authentication Mode	true (default)	Mandatory
SYNCSOURCE	Specifies where the user/ group information is extracted to be put into ranger database.	LDAP	

Sample Cluster Properties File

The following snapshot illustrates a sample cluster properties file:

```
A Typical Hadoop Cluster.
 #Log directory
 HDP_LOG_DIR=d:\hadoop\logs

 #Data directory
 HDP_DATA_DIR=d:\hadoop\data
 HDFS_NAMENODE_DATA_DIR=d:\hadoop\data\hdfs\nn,c:\hdpdata,d:\hdpdatann
 HDFS_DATANODE_DATA_DIR=d:\hadoop\data\hdfs\dn,c:\hdpdata,d:\hdpdatadn

 #Hosts
 NAMENODE_HOST=onprem-ranger1
 SECONDARY_NAMENODE_HOST=onprem-ranger1
 HIVE_SERVER_HOST=onprem-ranger1
 OOZIE_SERVER_HOST=onprem-ranger1
 WEBHCAT_HOST=onprem-ranger1
 FLUME_HOSTS=onprem-ranger1
 HBASE_MASTER=onprem-ranger1
 HBASE_REGIONSERVERS=onprem-ranger2
 SLAVE_HOSTS=onprem-ranger2
 ZOOKEEPER_HOSTS=onprem-ranger1
 KNOX_HOST=onprem-ranger2
 STORM_SUPERVISORS=onprem-ranger2
 STORM_NIMBUS=onprem-ranger1
 SPARK_JOB_SERVER=onprem-ranger1
 SPARK_HIVE_METASTORE=metastore
 IS_SLIDER=

 #Database host
 DB_FLAVOR=mssql
 DB_PORT=9433
 DB_HOSTNAME=singlehcatms7.cloudapp.net
```

```
#Hive properties
HIVE_DB_NAME=onpremranger1hive
HIVE_DB_USERNAME=hive
HIVE_DB_PASSWORD=hive
HIVE_DR=YES

#Oozie properties
OOZIE_DB_NAME=onpremranger1oozie
OOZIE_DB_USERNAME=oozie
OOZIE_DB_PASSWORD=oozie

#ASV/HDFS properties
DEFAULT_FS=HDFS
RESOURCEMANAGER_HOST=onprem-ranger1
IS_TEZ=yes
ENABLE_LZO=yes
RANGER_HOST=onprem-ranger1
RANGER_ADMIN_DB_HOST=localhost
RANGER_ADMIN_DB_PORT=3306
RANGER_ADMIN_DB_ROOT_PASSWORD=hcattest
RANGER_ADMIN_DB_DBNAME= xasecure
RANGER_ADMIN_DB_USERNAME= xaadmin
RANGER_ADMIN_DB_PASSWORD=admin
RANGER_AUDIT_DB_HOST=localhost
RANGER_AUDIT_DB_PORT=3306
RANGER_AUDIT_DB_ROOT_PASSWORD=hcattest
RANGER_EXTERNAL_URL=http://localhost:6080
RANGER_AUDIT_DB_DBNAME= xasecure
RANGER_AUDIT_DB_USERNAME= xalogger
RANGER_AUDIT_DB_PASSWORD=xalogger
RANGER_AUTHENTICATION_METHOD=LDAP
RANGER_LDAP_URL=ldap://71.127.43.33:389
RANGER_LDAP_USERDNPATTERN=uid={0},ou=users,dc=xasecure,dc=net
RANGER_LDAP_GROUPSEARCHBASE=ou=groups,dc=xasecure,dc=net
RANGER_LDAP_GROUPSEARCHFILTER=(member=uid={0},ou=users,dc=xasecure,dc=net)
RANGER_LDAP_GROUPROLEATTRIBUTE=cn
RANGER_POLICY_ADMIN_URL=http://localhost:6080
RANGER_HDFS_REPO=hadoopdev
RANGER_HIVE_REPO=hivedev
RANGER_HBASE_REPO=hbasedev
RANGER_KNOX_REPO=knoxdev
RANGER_STORM_REPO=stormdev
RANGER_SYNC_INTERVAL=360
RANGER_SYNC_LDAP_URL=ldap://10.0.0.4:389
RANGER_SYNC_LDAP_BIND_DN=cn=Administrator,cn=users,dc=hwqe,dc=net
RANGER_SYNC_LDAP_BIND_PASSWORD=Horton!#%works
RANGER_SYNC_LDAP_USER_SEARCH_SCOPE=sub
RANGER_SYNC_LDAP_USER_OBJECT_CLASS=person
RANGER_SYNC_LDAP_USER_NAME_ATTRIBUTE=cn
RANGER_SYNC_LDAP_USER_GROUP_NAME_ATTRIBUTE=memberof,ismemberof
RANGER_SYNC_LDAP_USERNAME_CASE_CONVERSION=lower
RANGER_SYNC_LDAP_GROUPNAME_CASE_CONVERSION=lower
AUTHSERVICEHOSTNAME=localhost
AUTHSERVICEPORT=5151
RANGER_SYNC_LDAP_USER_SEARCH_BASE=cn=users,dc=hwqe,dc=net
POLICYMGR_HTTP_ENABLED=true
REMOTELOGINENABLED=true
SYNCSOURCE=LDAP
```

3. Deploying a Multi-node HDP Cluster

This section describes the HDP MSI Installer, and explains three different options for deploying a multi-node Hadoop cluster from the command line or from a script. When installing HDP from the command line, the Hadoop setup script parses the cluster properties file and determines which services to install for each host.

3.1. About the HDP MSI Installer and HDP Public Properties

This section describes the HDP MSI installer command line options and explains which HDP public properties to use when installing a multi-node Hadoop Cluster. The installation process runs in the background.

HDP MSI Installer command format

The HDP MSI Installer command includes the `msiexec` command, a set of standard installer options, and HDP public properties. For example:

```
msiexec /qn /lv log_file /i msi_file MSIUSEREALADMINDETECTION=1
HDP_DIR=install_dir
HDP_LAYOUT=cluster_properties_file
HDP_USER_PASSWORD=password
DESTROY_DATA=YES_OR_NO
HDP=YES_OR_NO
FLUME=YES_or_NO
HBASE=YES_or_NO
KNOX=YES_or_NO
KNOX_MASTER_SECRET=secret
FALCON=YES_or_NO
STORM=YES_or_NO
RANGER=YES_or_NO
SPARK=YES_or_NO
```

where:

`msiexec /qn /lv log_file /i msi_file MSIUSEREALADMINDETECTION=1` is the set of standard installer options recommended by Hortonworks.

Everything following `/i msi_file MSIUSEREALADMINDETECTION=1` is a public property.

3.1.1. Standard Installer Options

- `/qn` (quiet, no UI) suppresses the HDP Setup Window. Use /qb (quiet basic) to suppress the HDP Setup and show a progress bar.
- `/lv log_file` (log verbose) creates a verbose installation log with the name you specified. If only a file name is provided, the installation log file is created in the directory where `msiexec` was launched.
- `/i msi_file` points to the HDP Installer file. We recommend specifying the absolute path.

- `MSIUSEREALADMINDETECTION=1` ensures that the user running the installer has true administrator permissions.

For more information about standard `msiexec` options, enter `msiexec /?` in a command prompt.

3.1.2. HDP Public Properties

You can set the following properties when you run `msiexec`:

Table 3.1. Property value information

Property	Mandatory?	Value	Associated Value(s) in Cluster Properties file	Description
DESTROY_DATA	Y	Yes or No	none	Specify `No` to keep existing data from previous installation. `No` does not format the NameNode. Specify `Yes` to *remove all existing or previous HDP data* and format the NameNode, creating an installation with a clean data slate.
HDP_USER_PASSWORD	Y	*password*	none	Password defined when creating the Hadoop user. Note that if the password does not meet your password policy standards, the installation will fail.
HDP_LAYOUT	Y	*clusterproperites_ full_path*	none	Absolute path to the Cluster Properties file. Note that relative paths are not supported and the path may not contain spaces. For example, c:\MSI_Install \clusterproperties.txt.
HDP_DIR	N	install_dir	none	Absolute path to the Hadoop root directory where HDP components are installed.
HDP	N	Yes or No		Setting this to `Yes` instructs the MSI to install optional HDP components such as Flume, HBase, Knox, Falcon and Storm. When enabled, you must specify the components on the command line; for example: `HDP="YES" KNOX="YES" KNOX_SECRET="secret" FALCON="NO"`

Property	Mandatory?	Value	Associated Value(s) in Cluster Properties file	Description
				HBASE="YES" FLUME="NO" STORM="NO". Excluding the optional components from the command line causes the installation to fail.
FLUME	N	Yes or No	FLUME_HOSTS	Includes the installation of Flume components on the hosts matching the name defined in the cluster properties file.
HBASE	N	Yes or No	HBASE_MASTER HBASE_REGIONSERVERS	Includes the installation of HBase components on the hosts matching the name defined in the cluster properties file.
KNOX	N	Yes or No	KNOX_HOST	Includes the installation of Knox gateway on the host matching the name defined in the cluster properties file. When yes, the KNOX_SECRET must also be specified as a parameter.
KNOX_MASTER_SECRET	N	*secret*	none	Specified only when KNOX="YES". The master secret to protect Knox security components, such as SSL certificates.
FALCON	N	Yes or No	FALCON_HOSTS	Includes the installation of the Falcon components on the host matching the name defined in the cluster properties file.
STORM	N	Yes or No	STORM_NUMBER STORM_SUPERVISORS	Includes the installation of the Storm components on the host matching the name defined in the cluster properties file.
RANGER	N	Yes or No	RANGER_HOST	Includes the installation of the Ranger Admin and User Sync components on the host matching the name defined in the cluster properties file.
SPARK	N	Yes or No	SPARK_JOB_SERVER SPARK_HIVE_METASTORE	Includes the installation of the Spark components on the host matching the name defined in the cluster properties file.

For optional HDP Components such as Knox and Falcon, include `HDP=yes` and specify "yes" or "no" for the components you would like to install or not, respectively. For example: `FLUME=no HBASE=yes KNOX=no FALCON=no STORM=no`.

This command needs to run in the command-line interface of each node in your cluster. If you are not installing any optional components, specify `HDP=no`.

Components are only installed if the host name matches a value in the cluster properties file. For examples of msiexec commands, see Option 3: Installing HDP from the Command Line.

3.2. Option 1: Central Push Install Using a Deployment Service

Many Windows data centers have standard corporate procedures for performing centralized push-install of software packages to hundreds or thousands of computers at the same time. In general, these same procedures also allow a centralized push-install of HDP to a Hadoop cluster.

If your Data Center already has such procedures in place, then follow this checklist:

1. Identify and configure the hosts for the Hadoop cluster nodes.
2. On the host nodes, complete all the prerequisites described in Before You Begin. Make sure you set an environment variable for `JAVA_HOME`. (Remember, Java cannot be installed in a location where the pathname includes spaces.)

   Be especially careful to identify:

   - Supported operating system
   - Dependent software and environment variable settings
   - Enable PowerShell remote scripting and set cluster nodes as trusted hosts
   - Resolvable host names and static IPv4 addresses
   - Open ports required for HDP operation
3. Download the HDP Windows Installation package. This package includes a sample cluster properties file called `clusterproperties.txt`.
4. Create (or edit) the cluster properties file using your host information. See Defining Cluster Properties.

Important

Nodes in the cluster communicate with each other using the host name or IP address defined in the cluster properties file. For multi-homed systems (systems that can be accessed internally and externally) and systems with more than one NIC, ensure that the preferred name or IP address is specified in the Cluster Properties file.

5. Use your standard procedures to push both the HDP Installer MSI and the custom cluster properties file to each node in the cluster.

6. Continue to use your standard procedures to execute the installation remotely, using the parameters documented in About the HDP MSI Installer and HDP Public Properties. For examples of `msiexec` commands, see Option 3: Installing HDP from the Command Line.

Note

The HDP Installer unpacks the MSI contents to `SystemDrive\HadoopInstallFiles`. A detailed installation log is located at `SystemDrive\HadoopInstallFiles\HadoopSetupTools\hdp-2.2.0.0.winpkg.install`. Do not remove this folder; it is required for uninstalling HDP.

7. Examine the results and/or logs from your standard procedures to ensure that all nodes were successfully installed.

   After the installation completes, configure and start the Hadoop services.

3.3. Option 2: Central HDP Install Using the Push Install HDP Script

Hortonworks provides a PowerShell script called `push_install_hdp.ps1`, which is included in the resources directory of the installer zip. The script installs HDP one system at a time on all hosts defined in the cluster properties file. Use this script to deploy HDP to a small test cluster. The script does not require shared storage, it copies the installation files to the target using the Windows Administrative Share.

Before running the script, ensure that the Admin Share is enabled on all cluster hosts, and that the Administrator account executing the script has the privileges to write to the cluster hosts.

To use the Push Install HDP script:

1. On the host nodes, complete all the prerequisites described in Before You Begin. Make sure you set an environment variable for `JAVA_HOME`. (Remember, Java cannot be installed in a location where the pathname includes spaces.)

   Be especially careful to identify:

   - Supported operating system
   - Dependent software and environment variable settings
   - Enable PowerShell remote scripting and set cluster nodes as trusted hosts
   - Resolvable host names and static IPv4 addresses
   - Open ports required for HDP operation

2. Additionally, on each host:

- Enable the Administrative Share:

```
netsh firewall set service type remoteadmin enabled
```

- Create the a target directory to which the installer can copy the files used for the installation:

```
mkdir D:\MSI_Install
```

3. Download the HDP Windows Installation package. The package includes a sample cluster properties file called `clusterproperties.txt`.

4. Define your cluster properties and save them in a file; see Manually Creating a Cluster Properties File.

Important

Nodes in the cluster communicate with each other using the host name or IP address defined in the cluster properties file. For multi-homed systems (systems that can be accessed internally and externally) and systems with more than one NIC, ensure that the preferred name or IP address is specified in the Cluster Properties file.

5. Copy the HDP MSI Installer, your custom cluster properties file, and `push_install_hdp.ps1` to the source directory on the master install node (the host from which you are running the push install).

6. Determine the MSI command line parameters. For information about parameters, see About the HDP MSI Installer and HDP Public Properties. For examples of `msiexec` commands, see Installing HDP from the Command Line.

7. On the master install node, open a command prompt with `run as Administrator`, and enter:

```
cd source_pathpowershell
-File push_install_hdp.ps1
source_path
destination_path
clusterproperties_file
files_list
skip
msiexec_command
-parallel
```

where:

- `source_path` is the absolute path to the installation files. This directory must contain the HDP MSI and the cluster properties file, as well as any other files the installer will push to the cluster nodes; for example, `D:\MSI_Install`.

- `destination_path` is the absolute path to an existing directory on the target cluster nodes. All nodes must have this directory. The installer copies files_list from `source_path` to `destination_path`. Specify destination path as a local path on the target host; for example, `D:\MSI_Install`.

- `clusterproperties_file` is the name of your custom cluster properties file; for example, `clusterproperties.txt`. (Do NOT include the path to the file.)
- `files_list` is a comma-delimited list of file names that the installer copies from `source_path` to all cluster hosts.

  The list must contain both the cluster property and HDP Installer file names; for example, `hdp-2.2.0.0.winpkg.msi,cluster.properties`. The list cannot contain spaces. Ensure that all the listed files are in the `source_path`.

  Tip

  When deploying HDP with the LZO compression enabled, put the following three files (from the Windows Installation zip) into the directory that contains the HDP for Windows Installer, and the cluster.properties file, and include them in the file list:

  - `hadoop-lzo-0.4.19.2.2.0.0-2060`
  - `gplcompression.dll`
  - `lzo2.dll`
- `skip` forces the installation on all nodes of a cluster.
- `msiexec_command` is the complete installation command that the script executes on the target nodes.
- `-parallel` allows parallel installation on all hosts, rather than installing one at a time.

  Note

  Parallel results are not always correctly deployed, and you must manually validate any parallel installations.

The installer script returns error messages or successful results to the Install Master host. These messages are displayed when the script finishes. Examine these results to ensure that all nodes were successfully installed.

On each node, the HDP Installer unpacks the MSI contents to `SystemDrive\HadoopInstallFiles\HadoopSetupTools\hdp-2.2.0.0.winpkg.install`. This folder is required to uninstall HDP; do not remove it.

3.4. Option 3: Installing HDP from the Command Line

Use the following instructions to install a single Hadoop cluster node from the command line using a cluster properties file:

1. On the host nodes, complete all the prerequisites described in Before You Begin. Be especially careful to identify:

   - Supported operating system
   - Dependent software and environment variable settings
   - Enable PowerShell remote scripting and set cluster nodes as trusted hosts
   - Resolvable host names and static IPv4 addresses
   - Open ports required for HDP operation

2. Download the HDP Windows Installation package. This package includes a sample cluster properties file, called `clusterproperties.txt`.

3. Create a cluster properties file using your host information; see Defining Cluster Properties.

   Note

   Nodes in the cluster communicate with each other using the host name or IP address defined in the cluster properties file. For multi-homed systems (systems that can be accessed internally and externally) and systems with more than one NIC, ensure that the preferred name or IP address is specified in the Cluster Properties file.

4. Place the MSI and custom cluster properties file in a local subdirectory on the host. Only the Hadoop Services that match the system's host name in the cluster properties file will be installed.

5. Use the same cluster properties file on every node in the cluster.

6. **(Optional)** When installing HDP with HDFS compression enabled, put the following three files (from the HDP for Windows Installation zip) into the directory that contains the HDP for Windows Installer and the cluster properties file:

   - `hadoop-lzo-0.4.19.2.2.0.0-2060`
   - `gplcompression.dll`
   - `lzo2.dll`

7. The following two examples show `msiexec` commands with HDP parameters.

   Warning

   These examples assume that you would like to destroy any existing HDP data. If you have existing data that you wish to keep, set DESTROY_DATA to no.

   Open a command prompt with the `run as Administrator` option, and enter the following:

```
msiexec /qn /i c:\MSI_Download\hdp-2.3.0.0\hdp-2.3.0.0\hdp-2.3.0.0.winpkg.
msi
/lv c:\MSI_Download\hdp.log
HDP_LAYOUT=c:\MSI_Download\Cluster.properties
HDP_DIR=c:\hdp
HDP=yes
DESTROY_DATA=yes
USEROOT=yes
HDP_USER_PASSWORD=TestUser123
KNOX=no
FALCON=no
STORM=no
RANGER="YesorNo"
KNOX_MASTER_SECRET=password
SPARK=no
```

To install a basic cluster with HBase, use the following command on every node:

```
msiexec /qn /i D:\MSI_Install\hdp-2.3.0.0.winpkg.msi
/lv D:\MSI_Install\hdp.log
MSIUSEREALADMINDETECTION=1
HDP_LAYOUT=D:\MSI_Install\cluster.properties
HDP_DIR=D:\hdp
DESTROY_DATA=yes
HDP_USER_PASSWORD=password
HDP=yes
KNOX=no
FALCON=no
HBase=yes
STORM=no
FLUME=no
RANGER=No
KNOX_MASTER_SECRET=password
SPARK=no
```

For a description of command line options, see Standard Installer Options.

8. The HDP Installer unpacks the MSI contents to `SystemDrive\HadoopInstallFiles`. A detailed installation log is located at `SystemDrive\HadoopInstallFiles \HadoopSetupTools\hdp-2.3.0.0.winpkg.install`. This folder is required to uninstall HDP; do not remove it.

Warning

If you are reinstalling HDP and wish to delete existing data but you did not specify DESTROY_DATA=YES, you need to format the HDFS file system. *Do not format the file system if you are upgrading an existing cluster and wish to preserve your data. Formatting will delete your existing HDP data.*

To format the HDFS file system, open the Hadoop Command Line shortcut on the Windows desktop, and then enter:

```
runas /user:hadoop "cmd /K HADOOP_HOME\bin\hadoop namenode -format"
```

3.5. Installing HDP Client Libraries on a Remote Host

HDP client libraries are Java libraries that facilitate communication from a remote host. An HDP client library has all the HDP JAR files on it for communicating with Hive, HDFS, etc. Note that you will not find any HDP service running on the client host machine.

Use the following instructions to install HDP client libraries on a remote host:

1. Copy existing clusterproperties.txt file from any host machine in your cluster.

2. Run the HDP installer from the client host. Execute the following command on your client host machine:

```
msiexec
/i "<$MSI_PATH>"
/lv "<$PATH_to_Installer_Log_File>"
HDP_LAYOUT="<$PATH_to_clusterproperties.txt_File>"
HDP_DIR="<$PATH_to_HDP_Install_Dir>"
DESTROY_DATA="<Yes_OR_No>"
```

   where:

   - **HDP_LAYOUT**: Mandatory parameter. Provide location of the copied clusterproperties.txt file on your client host machine (For example, `d:\config\clusterproperties.txt`). The path to the `clusterproperties.txt` file must be absolute. Relative paths do not work.

   - **HDP_DIR**: Optional parameter. Install directory for HDP (For example, d:\hdp). The default value is <$Default_Drive>/hdp.

   - **DESTROY_DATA**: Optional parameter. Specifies whether to preserve or delete existing data in target data directories (allowed values are undefined(default), yes, and no).

4. Configuring HDP Components and Services

This section describes component settings that must be updated, or additional software that must be installed, after installing HDP components.

Use one of the methods in Defining Cluster Properties to modify the cluster properties file. When you are finishing modifying cluster properties, start HDP services.

4.1. Configuring Hadoop Client Memory

You can use the Hadoop client to submit jobs to the Hadoop cluster. You might need to optimize the Hadoop client memory allocated by to the client machine. The Hadoop client memory is the amount of RAM utilized by the Hadoop client process and is defined in the `%HADOOP_HOME%\etc\hadoop\hadoop-env.cmd` configuration file. To set the Hadoop client memory configuration, run the following command:

```
set HADOOP_CLIENT_OPTS=-XmxMemory_amount
```

where `Memory_amount is the new RAM specification.`

4.2. Enable HDP Services

By default the following HDP services are disabled:

- Apache Falcon
- Apache Flume agent
- Apache Knox REST API
- Apache Hadoop thrift or Apache Hadoop thrift2

To enable these services to start and stop using the Start Local or Remote HDP script, first enable them **in the following order.** Note that the `sc config` command requires a space between the `start` option and its value.

1. Enable Thrift or Thrift2 on a cluster node.

Note

Thrift and Thrift2 use the same port, so they cannot run at the same time.

```
sc config thrift start= demand
```

OR

```
sc config thrift2start= demand
```

2. Enable Apache Falcon:

```
sc config falcon start= demand
```

3. Enable the Flume agent:

```
sc config flumeagent start= demand
```

4. **(Optional)** To allow access to the cluster through the Knox Gateway, enable REST on a cluster node:

```
sc config rest start= demand
```

4.3. (Microsoft SQL Server Only:) Configure Hive when Metastore DB is in a Named Instance

If your site uses Microsoft SQL Server for the Hive metadata store and the Hive database is not in the default instance (that is, in a named instance), you must configure the connection string after the installation completes:

1. On the Hive host, open the `hive-site.xml` file in a text editor.

2. Add the instance name to the property of the connection URL:

```
<property>
<name>javax.jdo.option.ConnectionURL</name>
<value>jdbc:sqlserver://sql-host/instance-name:port/hive_db;create=true</
value>
<description>JDBC connect string for a JDBC metastore</description>
</property>
```

   where:

   - `sql-host` is the SQL Server host name
   - `instance-name` is the name of the instance that the Hive database is in
   - `hive_db` is the name of the Hive database

3. Save the changes to `hive-site.xml`.

4. Finish configuring Hive as described later in this chapter, before restarting the Apache Hadoop Hive service.

4.4. Configure MapReduce on HDFS

To use MapReduce, create the MapReduce history folder, `tmp` directory, application logs, and a YARN folder in HDFS. Then set folder permissions:

```
%HADOOP_HOME%\bin\hdfs dfs -mkdir -p /mapred/history/done/mapred/history/
done_intermediate
%HADOOP_HOME%\bin\hdfs dfs -chmod -R 1777 /mapred/history/done_intermediate
```

```
%HADOOP_HOME%\bin\hdfs dfs -chmod 770/mapred/history/done
%HADOOP_HOME%\bin\hdfs dfs -chown -R hadoop:hadoopUsers /mapred
%HADOOP_HOME%\bin\hdfs dfs -chmod 755 /mapred /mapred/history
%HADOOP_HOME%\bin\hdfs dfs -mkdir /tmp
%HADOOP_HOME%\bin\hdfs dfs -chmod 777 /tmp
%HADOOP_HOME%\bin\hdfs dfs -mkdir /app-logs
%HADOOP_HOME%\bin\hdfs dfs -chown hadoop:hadoopUsers /app-logs
%HADOOP_HOME%\bin\hdfs dfs -chmod 1777 /app-logs
%HADOOP_HOME%\bin\hdfs dfs -mkdir -p /yarn /yarn/generic-history/
%HADOOP_HOME%\bin\hdfs dfs -chmod -R 700 /yarn
%HADOOP_HOME%\bin\hdfs dfs -chown -R hadoop:hadoop /yarn
```

4.5. Configure HBase on HDFS

To use HBase, create the HBase application data folder, and then set folder permissions:

```
%HADOOP_HOME%\bin\hdfs dfs -mkdir -p /apps/hbase/data
%HADOOP_HOME%\bin\hdfs dfs -chown hadoop:hadoop /apps/hbase/data
%HADOOP_HOME%\bin\hdfs dfs -chown hadoop:hadoop /apps/hbase
%HADOOP_HOME%\bin\hdfs dfs -mkdir -p /user/hbase
%HADOOP_HOME%\bin\hdfs dfs -chown hadoop:hadoop /user/hbase
```

4.6. Configure Hive on HDFS

To use Hive, create the Hive warehouse directory, the Hive and WebHcat user directories, and the WebHCat application folder in HDFS. Then set directory permissions so all users can access them:

1. Open the command prompt with the hadoop user account:

   ```
   runas /user:hadoop cmd
   ```

2. Make a user directory for hive and the hive warehouse directory:

   ```
   %HADOOP_HOME%\bin\hdfs dfs -mkdir -p /user/hive /hive/warehouse
   ```

3. Make a user and application directory for WebHcat:

   ```
   %HADOOP_HOME%\bin\hdfs dfs -mkdir -p /user/hcat
   %HADOOP_HOME%\bin\hdfs dfs -mkdir -p /apps/webhcat
   ```

4. Change the directory owner and permissions:

   ```
   %HADOOP_HOME%\bin\hdfs dfs -chown hadoop:hadoop /user/hive
   %HADOOP_HOME%\bin\hdfs dfs -chmod -R 755 /user/hive
   %HADOOP_HOME%\bin\hdfs dfs -chown -R hadoop:users/hive/warehouse
   %HADOOP_HOME%\bin\hdfs dfs -chown -R  hadoop:hadoop /user/hcat
   %HADOOP_HOME%\bin\hdfs dfs -chmod -R 777 /hive/warehouse
   %HADOOP_HOME%\bin\hdfs dfs -chown -R hadoop:users /apps/webhcat
   %HADOOP_HOME%\bin\hdfs dfs -chmod -R 755 /apps/webhcat
   ```

4.7. Configure Tez for Hive

If your cluster properties file specifies `IS_TEZ=yes` (use Tez for Hive), perform the following steps after HDP deployment:

1. Open the command prompt with the `hadoop` account:

```
runas /user:hadoop cmd
```

2. Make a Tez application directory in HDFS:

```
%HADOOP_HOME%\bin\hdfs dfs -mkdir /apps/tez
```

3. Allow all users read and write access:

```
%HADOOP_HOME%\bin\hdfs dfs -chmod -R 755 /apps/tez
```

4. Change the owner of the file to `hadoop`:

```
%HADOOP_HOME%\bin\hdfs dfs -chown -R hadoop:users /apps/tez
```

5. Copy the Tez home directory on the local machine, into the HDFS `/apps/tez` directory:

```
%HADOOP_HOME%\bin\hdfs dfs -put %TEZ_HOME%\* /apps/tez
```

6. Remove the Tez configuration directory from the HDFS Tez application directory:

```
%HADOOP_HOME%\bin\hdfs dfs -rm -r -skipTrash /apps/tez/conf
```

7. Ensure that the following properties are set in the `%HIVE_HOME%\conf\hive-site.xml` file:

Table 4.1. Required properties

Property	Default Value	Description
hive.auto.convert.join.noconditionaltask	true	Specifies whether Hive optimizes converting common JOIN statements into MAPJOIN statements. JOIN statements are converted if this property is enabled and the sum of size for n-1 of the tables/partitions for an n-way join is smaller than the size specified with the hive.auto.convert.join.noconditionaltask.size property.
hive.auto.convert.join.noconditionaltask.size	10000000 (10 MB)	Specifies the size used to calculate whether Hive converts a JOIN statement into a MAPJOIN statement. The configuration property is ignored unless hive.auto.convert.join.noconditionaltask is enabled.
hive.optimize. reducededuplication.min.reducer	4	Specifies the minimum reducer parallelism threshold to meet before merging two MapReduce jobs. However, combining a mapreduce

Property	Default Value	Description
		job with parallelism 100 with a mapreduce job with parallelism 1 may negatively impact query performance even with the reduced number of jobs. The optimization is disabled if the number of reducers is less than the specified value.
hive.tez.container.size	-1	By default, Tez uses the java options from map tasks. Use this property to override that value. Assigned value must match value specified for mapreduce.map.child.java.opts.
hive.tez.java.opts	n/a	Set to the same value as `mapreduce.map.java.opts`.

Adjust the settings above to your environment where appropriate; `hive-default.xml.template` contains examples of the properties.

8. To verify that the installation process succeeded, run smoke tests for Tez and Hive.

4.8. Configure Node Label Support for YARN Applications

Node labels can be used to restrict YARN applications so that the applications run only on cluster nodes that have a specified node label. Node Labels are supported on Windows. To enable node label support, make the following changes. See the Linux Node Labels documentation for more information. If you do not plan to use node labels, none of these changes are needed.

1. Open the command prompt using the `hadoop` account:

```
runas /user:hadoop cmd
```

2. Create a top-level YARN application directory in HDFS:

```
%HADOOP_HOME%\bin\hdfs dfs -mkdir -p /system/yarn/node-labels
```

3. Make sure permissions are set for write access from the `hadoop` account (rwx for all the directories in the path).

4. Change the owner of the file to `hadoop`:

```
%HADOOP_HOME%\bin\hdfs dfs -chown -R hadoop:users /system/yarn
%HADOOP_HOME%\bin\hdfs dfs -chmod -R 700 /system/yarn
```

5. Add the following property values to `yarn-site.xml`:

Table 4.2. Required properties

Property	Value
yarn.node-labels.manager-class	org.apache.hadoop.yarn.server.resourcemanager.nodelabels.RMNodeLabelsManager
yarn.node-labels.fs-store.root-dir	/system/yarn/node-labels

Property	Value
yarn.node-labels.fs-store.retry-policy-spec	2000,500

6. To verify that the installation process succeeded, run smoke tests as described in Validating the Installation.

4.9. Configure Ranger Security

Apache Ranger delivers a comprehensive approach to security for a Hadoop cluster. It provides central security policy administration across the core enterprise security requirements of authorization, accounting, and data protection.

The Ranger Policy Manager and Ranger UserSync are installed in only one host (specified in the HDP Setup Ranger Host parameter); the Ranger plug-ins for corresponding components are installed wherever those components are installed.

Make sure that the MySQL database used by Ranger is set up to connect from any host in the cluster. For multiple hosts, set up the Ranger MySQL database to connect from any host in the cluster using the root password. To do this, enter:

```
grant all privileges on *.* to 'root'@'%' identified by
'RootPasswordHere' ;flush privileges;
```

Note

In HDP v2.2, MySQL is the only database supported for use with Ranger.

4.10. Configuring HDFS Compression using GZIP

You can configure HDFS comrpession using GzipCodec. Gzip is CPU intensive, but provides a high compression ratio. Gzip is recommended for long term storage.

To configure compression using Gzip, for a one time job, you can execute the following command. This does not require that you restart your cluster.

```
hadoop jar hadoop-examples-1.1.0-SNAPSHOT.jar sort
"-Dmapred.compress.map.output=true"
"-Dmapred.map.output.compression.codec=org.apache.hadoop.io.compress.
GzipCodec"
"-Dmapred.output.compress=true"
"-Dmapred.output.compression.codec=org.apache.hadoop.io.compress.GzipCodec"
-outKey org.apache.hadoop.io.Text -outValue org.apache.hadoop.io.Text input
 output
```

To configure Gzip as the default compression, edit your `core-site.xml` configuration file as follows:

```
core-site.xml
<property>
<name>io.compression.codecs</name>
<value>org.apache.hadoop.io.compress.GzipCodec,org.apache.hadoop.io.compress.
DefaultCodec,org.apache.hadoop.io.compress.BZip2Codec</value>
```

```
<description>A list of the compression codec classes that can be used
for compression/decompression.</description>
</property>
mapred-site.xml
<property>
<name>mapred.compress.map.output</name>
<value>true</value>
</property>
<property>
<name>mapred.map.output.compression.codec</name>
<value>org.apache.hadoop.io.compress.GzipCodec</value>
</property>
<property>
<name>mapred.output.compression.type</name>
<value<BLOCK</value>
</property>
<! – Enable the following two configs if you want to turn on job output
 compression. This is generally not done -->
<property>
<name>mapred.output.compress</name>
<value>true</value>
</property>
<property>
<name>mapred.output.compression.codec</name>
<value>org.apache.hadoop.io.compress.GzipCodec</value>
</property>
```

4.11. Configuring LZO Compression

LZO compression is a lossless data compression library favoring speed over compression ratio; LZO compression is recommended for temporary tables. You can enable LZO compression for HDP to optimize Hive query speed.

LZO compression is not enabled automatically. To enable it, perform the following steps on each node in your cluster:

1. Copy the `hadoop-lzo.jar` file from your installation zip package to `%HADOOP_COMMON_HOME%\share\hadoop\common`.

2. Copy `gplcompression.dll` and `lzo2.dll`from your installation zip package to the same bin folder as `hadoop.dll`.

3. Ensure that the following configuration properties are set in `core-site.xml`:

```
  <property>
    <name>io.compression.codecs</name>
<value>org.apache.hadoop.io.compress.GzipCodec,org.apache.hadoop.io.
compress.DefaultCodec,com.hadoop.compression.lzo.LzoCodec,com.hadoop.
compression.lzo.LzopCodec,org.apache.hadoop.io.compress.SnappyCodec</value>
  </property>
  <property>
    <name>io.compression.codec.lzo.class</name>
    <value>com.hadoop.compression.lzo.LzoCodec</value>
  </property>
```

4.12. Setting up the Oozie Web Console

The Oozie Web Console is not enabled automatically during installation, so you should set it up manually as part of your configuration.

1. If it is running, use the Control Panel the stop the Oozie service if it is running. It ise named Apache Hadoop oozieservice.

2. Download the extjs-2.2 zip file and copy it to `%OOZIE_HOME%\extra_libs`. For example: `D:\hdp\oozie-4.1.0.2.2.1.0-2190\extra_libs\ext-2.2.zip`.

3. Prepare Oozie using the following command:

   ```
   %OOZIE_HOME%\oozieoozie-win-distro\bin\oozie-setup.cmd prepare-war
   ```

   For example:

   ```
   D:\hdp\oozie-4.1.0.2.2.1.0-2190\oozie-win-distro\bin\oozie-setup.cmd
    prepare-war
   ```

4. Restart the Oozie service.

4.13. Using Apache Slider

Apache Slider lets you deploy distributed applications across a Hadoop cluster. On the Windows platform, Slider application packages are included in the Windows Installer MSI for HBase and Storm. (Accumulo is supported on Linux, but is not currently supported on Windows). See Running Applications on YARN Using Slider.

4.14. (Optional) Install Microsoft SQL Server JDBC Driver

If you are using MS SQL Server for Hive and Oozie metastores, you must install the MS SQL Server JDBC driver after installing Hive or Oozie.

1. Download the SQL JDBC JAR file sqljdbc_3.0.1301.101_enu.exe.

2. Run the downloaded file.

   (By default, the SQL JDBC driver file is extracted to `Downloads\Microsoft SQL Server JDBC Driver 3.0.`)

3. Copy and paste `Downloads\Microsoft SQL Server JDBC Driver 3.0\sqljdbc_3.0\enu\sqljdbc4.jar` to `HIVE_HOME/lib` (where `HIVE_HOME` can be set to `D:\hadoop\hive-0.9.0`).

4.15. Start HDP Services

Following are steps for starting local and remote services. For more information, see Managing HDP on Windows.

1. Start local services on the Master Nodes:

   ```
   %HADOOP_NODE%\start_local_hdp_services.cmd
   ```

   Wait for the Master Node services to start up before continuing.

2. At any Master Node, start all slave node services:

   ```
   %HADOOP_NODE%\start_remote_hdp_services.cmd
   ```

3. At the Knox Gateway:

   ```
   %HADOOP_NODE%\start_local_hdp_services.cmd
   ```

4. Smoke test your installation as described in Validating the Installation.

4.16. Updating Your Configuration

Once you have installed and configured HDP, you can updated the configuration at any time, by manually modifying any component configuration file. See the below table for the component configuration file default locations.

If the changes you have made impact just the master service, you only need to restart the master. If your changes are for slave services, then you need to prompagate the edited configuration file to each slave host and restart the slave and master services on all hosts.

Table 4.3. Component configuration and log file locations

Component	Configuration file location	Log file location
Hadoop (HDFS, YARN, MapReduce)	C:\hdp\hadoop-2.7.1.2.3.0.0-2543\etc\hadoop	C:\hadoop\logs\hadoop
ZooKeeper	C:\hdp\zookeeper-3.4.6.2.3.0.0-2543\conf	C:\hadoop\logs\zookeeper
Hive	C:\hdp\hive-1.2.1.2.3.0.0-2543\conf	C:\hadoop\logs\hive
HBase	C:\hdp\hbase-1.1.1.2.3.0.0-2543\conf	C:\hadoop\logs\hbase
WebHCat	C:\hdp\hive-1.2.1.2.3.0.0-2543\hcatalog\etc\webhcat	C:\hadoop\logs\webhcat
Oozie	C:\hdp\oozie-4.2.0.2.3.0.0-2543\oozie-win-distro\conf	C:\hadoop\logs\oozie
Storm	C:\hdp\storm-0.10.0.2.3.0.0-2543\conf	C:\hadoop\logs\storm
Knox	C:\hdp\knox-0.6.0.2.3.0.0-2543\conf	C:\hadoop\logs\knox
Flume	C:\hdp\flume-1.5.2.2.3.0.0-2543\conf	C:\hdp\flume-1.5.2.2.3.0.0-2543\bin\logs
Pig	C:\hdp\pig-0.15.0.2.3.0.0-2543\conf	No log files because no service is running

Component	Configuration file location	Log file location
Sqoop	C:\hdp\sqoop-1.4.6.2.3.0.0-2543\conf	No log files because no service is running
Tez	C:\hdp\tez-0.7.0.2.3.0.0-2543\conf	No log files because no service is running

5. Validating the Installation

After the HDP Cluster installation is completed, run the provided smoke tests to validate the installation. These tests validate installed functionality by executing a set of tests for each HDP component.

1. Start HDP services:

```
%HADOOP_NODE_INSTALL_ROOT%\start_remote_hdp_services.cmd
```

2. Open command prompt and execute cmd as hadoop user:

```
runas /user:hadoop cmd
```

3. Create a smoketest user directory in HDFS, if one does not already exist:

```
%HADOOP_HOME%\bin\hdfs dfs -mkdir -p /user/smoketestuser
%HADOOP_HOME%\bin\hdfs dfs -chown -R smoketestuser /user/smoketestuser
```

4. Open a command prompt and run the smoke tests as the **hadoop** user:

```
runas /user:hadoop "cmd /K
%HADOOP_NODE%\Run-SmokeTests.cmd"
```

   (You can also create a smoketest user in HDFS as described in Appendix: Adding a Smoketest User, and then run the tests as the smoketest user.)

(**Optional**) If you installed Ranger, verify that the installation was successful using any or all of the following checks:

1. Check whether the Database `RANGER_ADMIN_DB_NAME` is present in the MySQL server running on `RANGER_ADMIN_DB_HOST`
2. Check whether the Database `RANGER_AUDIT_DB_NAME` is present in the MySQL server running on `RANGER_AUDIT_DB_HOST`
3. Check whether the "ranger-admin" service is installed in `services.msc`
4. Check whether the `ranger-usersync` service is installed in `services.msc`
5. If you plan to use the Ranger Administration Console with the UserSync feature, check whether both services start.
6. Go to the Ranger Administration Console host URL and make sure you can log in using the default user credentials.

Important

If you see installation failures for any HDP component, we recommend that you reinstall HDP.

6. Managing HDP on Windows

This section describes how to manage HDP on Windows.

6.1. Starting HDP Services

The HDP Windows installer sets up Windows services for each HDP component across the nodes in a cluster. Use the following instructions to start HDP services from any host machine in your cluster.

Complete the following instructions as the administrative user:

1. Start the HDP cluster by running the following command from any host in your cluster.

Important

To Enable NameNode High Availability, do so while starting HDP services. Do not wait until all services have started.

```
%HADOOP_NODE_INSTALL_ROOT%\start_remote_hdp_services.cmd
```

2. Open the Services administration pane, `Control Panel > Administrative Tools > Services`.

   You should see a list of installed services and their status.

6.2. Enabling NameNode High Availability

If you are enabling NameNode High Availability in a multi-node cluster, you can run the following commands on the primary and standby hosts while services are starting. Log in to every host and run these commands as `administrator`.

1. On the primary host, run:

```
hdfs namenode -format -force
```

2. On each standby host, run:

```
hdfs namenode -bootstrapStandby -force hdfs zkfc -formatZK -force
```

6.3. Validating HA Configuration

1. Verify the state of each NameNode, using one the following methods:

   a. Open the web page for each NameNode in a browser, using the configured URL.

      The HA state of the NameNode should appear in the configured address label; for example, NameNode `example.com.8020` (standby).

Note

The NameNode state may be `standby` or `active`. After bootstrapping, the HA NameNode state is initially `standby`.

b. Query the state of a NameNode using `JMX(tag.HAState)`

c. Query the service state using the following command:

```
hdfs haadmin -getServiceState
```

2. Verify automatic failover.

a. Locate the Active NameNode.

Use the NameNode web UI to check the status for each NameNode host machine.

b. Cause a failure on the Active NameNode host machine.

i. Turn off automatic restart of the service.

a. In the Windows Services pane, locate the Apache Hadoop NameNode service, right-click, and choose `Properties`.

b. On the Recovery tab, select `Take No Action` for `First, Second, and Subsequent Failures`, then choose `Apply`.

ii. Simulate a JVM crash. For example, you can use the following command to simulate a JVM crash:

```
taskkill.exe /t /f /im namenode.exe
```

Alternatively, power-cycle the machine or unplug its network interface to simulate an outage. The Standby NameNode state should become `Active` within several seconds.

Note

The time required to detect a failure and trigger a failover depends on the configuration of `ha.zookeeper.session-timeout.ms` property. The default value is 5 seconds.

iii. Verify that the Standby NameNode state is `Active`.

a. If a standby NameNode does not activate, verify that the HA settings are configured correctly.

b. To diagnose issues, check log files for `zkfc` daemons and NameNode daemons.

6.4. Stopping HDP Services

The HDP on Windows installer sets up Windows services for each HDP component across the nodes in a cluster. To stop HDP services, run the following command from any host machine in your cluster, while logged on as the administrative user:

```
%HADOOP_NODE_INSTALL_ROOT%\stop_remote_hdp_services.cmd
```

7. Troubleshooting Your Deployment

Use the following information to troubleshoot issues encountered while deploying HDP on the Windows platform:

7.1. Installation Errors

This section contains fixes to some common installation errors.

7.1.1. Granting Symbolic Link Privileges

Description

You must have privileges to create sybolic links, prior to installing HDP for Windows. If you attempt to install HDP for Windows without symbolic link creation privileges, your MSI installation fails with the following error:

```
CREATE-USER: Setting password for hadoop
CREATE-USER: Granting SeCreateSymbolicLinkPrivilege
CREATE-USER: <installation_package_path> -u WIN2012\hadoop +r
 SeCreateSymbolicLinkPrivilege
CREATE-USER FAILURE: Failed to grant SeCreateSymbolicLinkPrivilege
HDP FAILURE: Failed to create Hadoop user
```

Workaround

To work around this issue, ensure that symbolic link creation privileges have been granted to the `users` group, in advance of performing your installation.

If granting symbolic link privileges to the `users` group conflicts with your company's security policies, you can create the hadoop user manually, in advance of your installation.

7.2. Cluster Information

Use the following commands to collect information about a Windows based cluster. This data helps to isolate specific deployment issues.

1. **Collect OS information**: This data helps to determine if HDP is deployed on a supported operating system (OS).

   To list the operating system, run the following command in PowerShell as an Administrator user:

   ```
   (Get-WmiObject -class Win32_OperatingSystem).Caption Microsoft Windows
   Server 2012 Standard
   ```

   To list the OS version for your host machine, enter:

   ```
   [System.Environment]::OSVersion.Version
   ```

2. **Determine installed software**This data can be used to troubleshoot performance issues or unexpected behavior for a specific node in your cluster. For example, unexpected behavior might be a situation where a MapReduce job runs for a longer duration than expected.

   To see the list of installed software on a particular host machine, go to `Control Panel -> All Control Panel Items -> Programs and Features`.

3. **Detect running processes**: This data can be used to troubleshoot either performance issues or unexpected behavior for a specific node in your cluster.

   You can either press CTRL + SHIFT + DEL on the affected host machine, or you can execute the following command on PowerShell as an Administrator user:

   ```
   tasklist
   ```

4. **Detect Java running processes**: Use this command to verify the Hadoop processes running on a specific machine.

   As HADOOP_USER, execute the following command on the affected host machine:

   ```
   su $HADOOP_USER jps
   ```

   You should see the following output:

   ```
   988 Jps
   2816 -- process information unavailable
   2648 -- process information unavailable
   1768 -- process information unavailable
   ```

   No actual name is given to any process. Ensure that you map the process IDs (pid) from the output of this command to the `.wrapper` file within the `c:\hdp \hadoop-1.1.0-SNAPSHOT\bin` directory.

   Note

   Ensure that you specify the complete path to the Java executable, if the Java bin directory's location is not set within your PATH.

5. **Detect Java heap allocation and usage**: Use the following command to list Java heap information for a specific Java process. This data can be used to verify the heap settings and thus analyze whether a specific Java process is reaching the threshold.

   Execute the following command on the affected host machine:

   ```
   jmap -heap pid_of_Hadoop_process
   ```

   You should see output similar to the following:

   ```
   C:\hdp\hadoop-1.1.0-SNAPSHOT>jmap -heap 2816
   Attaching to process ID 2816, please wait...
   Debugger attached successfully.
   Server compiler detected.
   JVM version is 20.6-b01

   using thread-local object allocation.
   ```

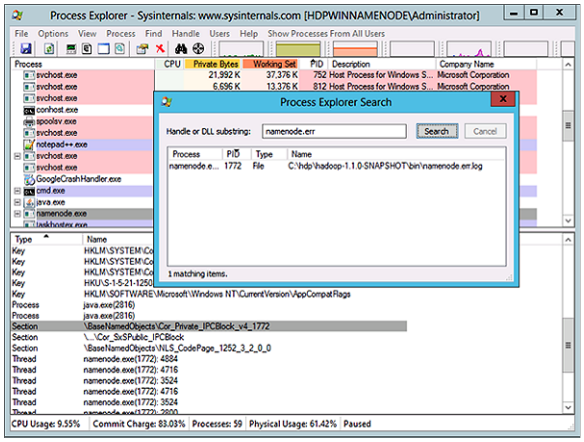
```
Mark Sweep Compact GC

Heap Configuration:
 MinHeapFreeRatio = 40
 MaxHeapFreeRatio = 70
 MaxHeapSize = 4294967296 (4096.0MB)
 NewSize = 1310720 (1.25MB)
 MaxNewSize = 17592186044415 MB
 OldSize = 5439488 (5.1875MB)
 NewRatio = 2
 SurvivorRatio = 8
 PermSize = 21757952 (20.75MB)
 MaxPermSize = 85983232 (82.0MB)

Heap Usage:
New Generation (Eden + 1 Survivor Space):
 capacity = 10158080 (9.6875MB)
 used = 4490248 (4.282234191894531MB)
 free = 5667832 (5.405265808105469MB)
 44.203707787298384% used
Eden Space:
 capacity = 9043968 (8.625MB)
 used = 4486304 (4.278472900390625MB)
 free = 4557664 (4.346527099609375MB)
 49.60548290307971% used
From Space:
 capacity = 1114112 (1.0625MB)
 used = 3944 (0.0037612915039062 5MB)
 free = 1110168 (1.0587387084960938MB)
 0.35400390625% used
To Space:
 capacity = 1114112 (1.0625MB)
 used = 0 (0.0MB)
 free = 1114112 (1.0625MB)
 0.0% used
tenured generation:
 capacity = 55971840 (53.37890625MB)
 used = 36822760 (35.116920471191406MB)
 free = 19149080 (18.261985778808594MB)
 65.7880105424442% used
Perm Generation:
 capacity = 21757952 (20.75MB)
 used = 20909696 (19.9410400390625MB)
 free = 848256 (0.8089599609375MB)
 96.10139777861446% used
```

6. **Show open files**: Use Process Explorer to determine which processes are locked on a specific file. For information on how to use Process Explorer, see Windows Sysinternals - Process Explorer.

   For example, you can use Process Explorer to troubleshoot the file lock issues that prevent a particular process from starting, as shown in the following screen shot:



7. **Verify well-formed XML**:

   Ensure that the Hadoop configuration files (for example, hdfs-site.xml, etc.) are well formed. You can either use Notepad++ or third-party tools like Oxygen, XML Spy, etc., to validate the configuration files. Here are instructions for Notepad++:

   - Open the XML file to be validated in Notepad++ and select `XML Tools -> Check XML Syntax`.
   - Resolve validation errors, if any.

8. **Detect AutoStart Programs**: This information helps to isolate errors for a specific host machine.

   For example, a potential port conflict between auto-started process and HDP processes, might prevent launch for one of the HDP components.

   Ideally, the cluster administrator must have the information on auto-start programs handy. Use the following command to launch the GUI interface on the affected host machine:

   ```
   c:\Windows\System32\msconfig.exe
   ```

   Click `Startup`. Ensure that no start-up items are enabled on the affected host machine.

9. **Create a list of all mounts on the machine**: This information determines the drives that are actually mounted or available for use on the host machine. To troubleshoot disk capacity issues, use the following PowerShell command to determine if the system is violating any storage limitations:

   ```
   Get-Volume
   ```

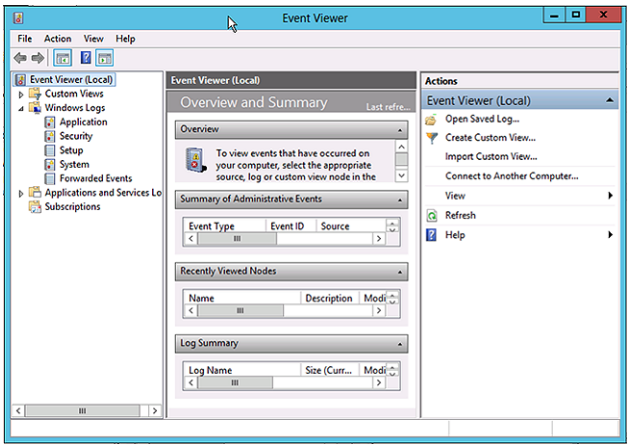
   You should see output similar to the following:

```
Drive       FileSystem FileSystem DriveType HealthStatus Size     Remaining
Letter      Label                                                  Size
----------- ---------- ---------- --------- ------------ --------
 -------------
System      Reserved   NTFS       Fixed     Healthy      108.7 MB 350 MB
C                      NTFS       Fixed     Healthy      10.74 GB 19.97 GB
D           HRM_SSS... UDF        CD-ROM    Healthy      0 B      3.44 GB
```

10. **Operating system messages**: Use `Event Viewer` to detect messages with a system or an application.

    Event Viewer can determine if a machine was rebooted or shut down at a particular time. Use the logs to isolate issues for HDP services that were non-operational for a specific time.

    Go to `Control Panel -> All Control Panel Items -> Administrative Tools` and click the `Event Viewer` icon.



11. **Hardware/system information**: Use this information to isolate hardware issues on the affected host machine.

    Go to `Control Panel -> All Control Panel Items -> Administrative Tools` and click the `System Information` icon.

12. **Network information**: Use the following commands to troubleshoot network issues.

    - **ipconfig**: This command provides the IP address, checks that the network interfaces are available, and validates whether an IP address is bound to the interfaces. To troubleshoot communication issues among host machines in your cluster, execute the following command on the affected host machine:

      ```
      ipconfig
      ```

      You should see output similar to the following:

      ```
      Windows IP Configuration

      Ethernet adapter Ethernet 2:
      ```

```
 Connection-specific DNS Suffix .  :
 Link-local IPv6 Address . . . . . : fe80::d153:501e:5df0:f0b9%14
 IPv4 Address. . . . . . . . . . . : 192.168.56.103
 Subnet Mask . . . . . . . . . . . : 255.255.255.0
 Default Gateway . . . . . . . . . : 192.168.56.100

Ethernet adapter Ethernet:

 Connection-specific DNS Suffix .  : test.tesst.com
 IPv4 Address. . . . . . . . . . . : 10.0.2.15
 Subnet Mask . . . . . . . . . . . : 255.255.255.0
 Default Gateway . . . . . . . . . : 10.0.2.2
```

- **netstat -ano**: This command generates a list of ports used within the system. To troubleshoot launch issues and resolve potential port conflicts with HDP master processes, run the following command on the host machine:

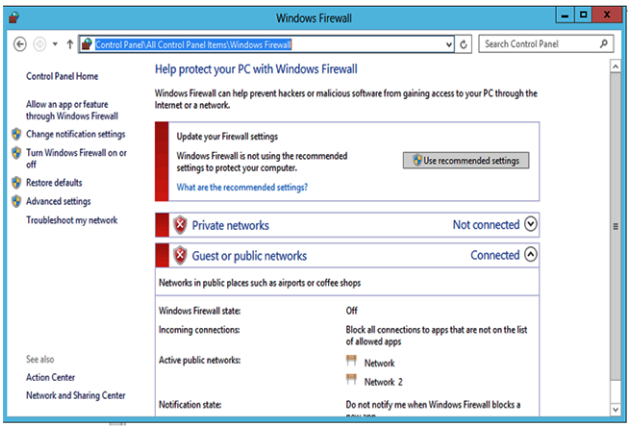
  ```
  netstat -ano
  ```

  You should see output similar to the following:

  ```
  TCP 0.0.0.0:49154 0.0.0.0:0 LISTENING 752
  TCP [::]:49154 [::]:0 LISTENING 752
  UDP 0.0.0.0:500 *:* 752
  UDP 0.0.0.0:3544 *:* 752
  UDP 0.0.0.0:4500 *:* 752
  UDP 10.0.2.15:50461 *:* 752
  UDP [::]:500 *:* 752
  UDP [::]:4500 *:* 752
  ```

- Verify if a firewall is enabled on the host machine: Go to `Control Panel -> All Control Panel Items -> Windows Firewall`.

  You should see the following GUI interface:



7.3. Component Environment Variables

It is useful to understand the environment variables that are modified during the HDP installation. The following displays the environment variables updated during installation:

Table 7.1. Component environment variables

Component	Modified environment variables
Hadoop	• HADOOP_HOME=c:\hdp • HADOOP_LOG_DIR=c:\hadoop\logs\hadoop • HADOOP_NODE=c:\hdp\ • HADOOP_NODE_INSTALL_ROOT=c:\hdp • HADOOP_OPTS= -Dfile.encoding=UTF-8 -Dhadoop.home.dir=c:\hdp\\ \hadoop-1.2.0.1.3.0.0-0380 -Dhadoop.root.logger=INFO,console,DRFA -Dhadoop.policy.file=hadoop-policy.xml -Dhadoop.log.dir=c:\hadoop\logs\hadoop -Dhadoop.log.file=hadoop-hive-HADOOP-NN.log • HADOOP_PACKAGES=c:\HadoopInstallFiles\HadoopPackages\ • HADOOP_SETUP_TOOLS=c:\HadoopInstallFiles\HadoopSetupTools\
HDFS	HDFS_DATA_DIR=c:\hdp_data\hdfs
Calcite	
DataFu	
Falcon	
Flume	FLUME_HOME=c:\hdp
HBase	• HBASE_CONF_DIR=c:\hdp\\\hbase-0.94.6.1.3.0.0-0380\conf • HBASE_HOME=c:\hdp\\\hbase-0.94.6.1.3.0.0-0380 • HBASE_LOG_DIR=c:\hadoop\logs\hbase • HCATALOG_HOME=c:\hdp
Hive	• HIVE_CONF_DIR=c:\hdp\\hive-0.11.0.1.3.0.0-0380\conf • HIVE_HOME=c:\hdp • HIVE_LIB_DIR=c:\hdp\\hive-0.11.0.1.3.0.0-0380\lib • HIVE_LOG_DIR=c:\hadoop\logs\hive • HIVE_OPTS= -hiveconf hive.querylog.location=c:\hadoop\logs\hive\history -hiveconf hive.log.dir=c:\hadoop\logs\hive
Knox	
Mahout	MAHOUT_HOME=c:\hdp
Oozie	• OOZIE_DATA=c:\hdp_data\oozie • OOZIE_HOME=c:\hdp\\\Oozie-3.3.2.1.3.0.0-0380\oozie-win-distro • OOZIE_LOG=c:\hadoop\logs\oozie • OOZIE_ROOT=c:\hdp
Phoenix	
Pig	PIG_HOME=c:\hdp
Ranger	
Slider	
Spark	
Sqoop	SQOOP_HOME=c:\hdp
Storm	
Tez	
ZooKeeper	• ZOOKEEPER_CONF_DIR=\zookeeper-3.4.5.1.3.0.0-0380\conf

Component	Modified environment variables
	• ZOOKEEPER_HOME=c:\hdp • ZOOKEEPER_LIB_DIR=\zookeeper-3.4.5.1.3.0.0-0380\lib • ZOO_LOG_DIR=c:\hadoop\logs\zookeeper
Additional environment variables	• JAVA_HOME=C:\java\jdk • TEMPLETON_HOME=c:\hdp • TEMPLETON_LOG_DIR=c:\hadoop\logs\webhcat

7.4. File Locations, Logging, and Common HDFS Commands

This section provides a list of files and their locations, instructions for enabling logging, and a list of HDFS commands that help isolate and troubleshoot issues.

7.4.1. File Locations

- **Configuration files:** These files are used to configure a hadoop cluster.
  - `core-site.xml`: All Hadoop services and clients use this file to locate the NameNode, so this file must be copied to each node that is either running a Hadoop service or is a client node. The Secondary NameNode uses this file to determine the location for storing fsimage and edits log `namefs.checkpoint.dir/name` locally, and the location of the NameNode `namefs.namedefault.name/name`.

    Use the `core-site.xml` file to isolate communication issues with the NameNode host machine.
  - `hdfs-site.xml`: HDFS services use this file, which contains a number of important properties. These include:
    - HTTP addresses for the two services
    - Replication for DataNodes `namedfs.replication/name>`
    - DataNode block storage location `namedfs.data.dir/name`
    - NameNode metadata storage `namedfs.name.dir/name`

    Use the `hdfs-site.xml` file to isolate NameNode start-up issues. Typically, NameNode start-up issues are caused when NameNode fails to load the `fsimage` and edits log to merge. Ensure that the values for the location properties in `hdfs-site.xml` are valid locations.
  - `datanode.xml`:

    DataNode services use the `datanode.xml` file to specify the maximum and minimum heap size for the DataNode service. To troubleshoot issues with DataNode: change the value for `-Xmx`, which changes the maximum heap size for the DataNode service. Restart the affected DataNode host machine.

- `namenode.xml`:

  NameNode services use the `namenode.xml` file to specify the maximum and minimum heap size for the NameNode service. To troubleshoot issues with NameNode, change the value for `-Xmx`, which changes the maximum heap size for NameNode service. Restart the affected NameNode host machine.

- `secondarynamenode.xml`:

  Secondary NameNode services use the `secondarynamenode.xml` file to specify the maximum and minimum heap size for the Secondary NameNode service. To troubleshoot issues with Secondary NameNode, change the value for -`Xmx`, which changes the maximum heap size for Secondary NameNode service. Restart the affected Secondary NameNode host machine.

- `hadoop-policy.xml`:

  Use the `hadoop-policy.xml` file to configure service-level authorization/ACLs within Hadoop. NameNode accesses this file. Use this file to troubleshoot permission related issues for NameNode.

- `log4j.properties`:

  Use the `log4j.properties` file to modify the log purging intervals of the HDFS logs. This file defines logging for all the Hadoop services. It includes, information related to appenders used for logging and layout. For more details, see the log4j documentation.

- Log Files: Following are sets of log files for each of the HDFS services. They are stored in `c:\hadoop\logs\hadoop` and `c:\hdp\hadoop-1.1.0- SNAPSHOT\bin` by default.

  - HDFS .out files: Log files with the .out extension are located in `c:\hdp \hadoop-1.1.0-SNAPSHOT\bin`. They have the following naming conventions:

    - `datanode.out.log`
    - `namenode.out.log`
    - `secondarynamenode.out.log`

    These files are created and written to when HDFS services are bootstrapped. Use these files to isolate launch issues with DataNode, NameNode, or Secondary NameNode services.

  - HDFS `.wrapper` files: The log files with the .wrapper extension are located in `c:\hdp \hadoop-1.1.0-SNAPSHOT\bin` and have the following file names:

    - `datanode.wrapper.log`
    - `namenode.wrapper.log`
    - `secondarynamenode.wrapper.log`

These files contain the start-up command string to start the service, and list process ID output on service start-up.

- HDFS `.log` and `.err` files:

  The following files are located in `c:\hdp\hadoop-1.1.0-SNAPSHOT\bin`:

  - `datanode.err.log`
  - `namenode.err.log`
  - `secondarynamenode.err.log`

  The following files are located in `c:\hadoop\logs\hadoop`:

  - `hadoop-datanode-Hostname.log`
  - `hadoop-namenode-Hostname.log`
  - `hadoop-secondarynamenode-Hostname.log`

  These files contain log messages for the running Java service. If there are any errors encountered while the service is already running, the stack trace of the error is logged in the above files.

  `Hostname` is the host where the service is running. For example, on a node where the host name is `host3`, the file would be saved as `hadoop-namenode-host3.log`.

  Note

  By default, these log files are rotated daily. Use the `c:\hdp \hadoop-1.1.0- SNAPSHOT\conf\log4j.properties` file to change log rotation frequency.

- HDFS <.date> files:

  Log files with the <.date> extension have the following format:

  - `hadoop-namenode-$Hostname.log.<date>`
  - `hadoop-datanode-$Hostname.log.<date>`
  - `hadoop-secondarynamenode-$Hostname.log.<date>`

  When a .log file is rotated, the current date is appended to the file name; for example: `hadoop-datanode- hdp121.localdomain.com.log.2013-02-08`.

  Use these files to compare the past state of your cluster with the current state, to identify potential patterns.

7.4.2. Enabling Logging

To enable logging, change the settings in the `hadoop-env.cmd` file. After modifying `hadoop-env.cmd`, recreate the NameNode service XML and then restart the NameNode.

To enable audit logging, change the `hdfs.audit.logger` value to `INFO,RFAAUDIT`. Overwrite the NameNode service XML and restart the NameNode.

1. Open the Hadoop Environment script, `%HADOOP_HOME%\etc\hadoop\hadoop-env.cmd`.
2. Prepend the following text in the HADOOP_NAMENODE_OPTS definition, for example to enable Garbage Collection logging:

   ```
   -Xloggc:%HADOOP_LOG_DIR%/gc-namenode.log -verbose:gc -XX:+PrintGCDetails -
   XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps
   ```

   For example:

   ```
   set HADOOP_NAMENODE_OPTS=-Xloggc:%HADOOP_LOG_DIR%/gc-namenode.log
   -verbose:gc
   -XX:+PrintGCDetails
   -XX:+PrintGCTimeStamps
   -XX:+PrintGCDateStamps
   -Dhadoop.security.logger=%HADOOP_SECURITY_LOGGER%
   -Dhdfs.audit.logger=%HDFS_AUDIT_LOGGER% %HADOOP_NAMENODE_OPTS%
   ```

3. Run the following command to recreate the NameNode service XML:

   ```
   %HADOOP_HOME%\bin\hdfs --service namenode > %HADOOP_HOME%\bin\namenode.xml
   ```

4. Verify that the NameNode Service XML was updated.
5. Restart the NameNode service.

7.4.3. Common HDFS Commands

This section provides common HDFS commands to troubleshoot HDP deployment on Windows platform. An exhaustive list of HDFS commands is available here.

1. **Get the Hadoop version:** Run the following command on your cluster host machine:

   ```
   hadoop version
   ```

2. **Check block information**: This command provides a directory listing and displays which node contains the block. Run this command on your HDFS cluster host machine to determine if a block is under-replicated.

   ```
   hdfs fsck / -blocks -locations -files
   ```

   You should see output similar to the following:

   ```
   FSCK started by hdfs from /10.0.3.15 for path / at Tue Feb 12 04:06:18 PST
   2013
   ```

```
 / <dir>
 /apps <dir>
 /apps/hbase <dir>
 /apps/hbase/data <dir>
 /apps/hbase/data/-ROOT- <dir>
 /apps/hbase/data/-ROOT-/.tableinfo.0000000001 727 bytes, 1 block(s):
 Under replicated blk_-3081593132029220269_1008.
 Target Replicas is 3 but found 1 replica(s). 0.
 blk_-3081593132029220269_1008
 len=727 repl=1 [10.0.3.15:50010]
 /apps/hbase/data/-ROOT-/.tmp <dir>
 /apps/hbase/data/-ROOT-/70236052 <dir>
 /apps/hbase/data/-ROOT-/70236052/.oldlogs <dir>
 /apps/hbase/data/-ROOT-/70236052/.oldlogs/hlog.1360352391409 421 bytes, 1
 block(s): Under
 replicated blk_709473237440669041_1006.
 Target Replicas is 3 but found 1
 replica(s). 0. blk_709473237440669041_1006 len=421 repl=1 [10.0.3.
15:50010] ...
```

3. **HDFS report**: Use this command to receive HDFS status. Execute the following command as the hadoop user:

```
 hdfs dfsadmin -report
```

   You should see output similar to the following:

```
-bash-4.1$ hadoop dfsadmin -report
Safe mode is ON
Configured Capacity: 11543003135 (10.75 GB)
Present Capacity: 4097507328 (3.82 GB)
DFS Remaining: 3914780672 (3.65 GB)
DFS Used: 182726656 (174.26 MB)
DFS Used%: 4.46%
Under replicated blocks: 289
Blocks with corrupt replicas: 0
Missing blocks: 0

-------------------------------------------------
Datanodes available: 1 (1 total, 0 dead)

Name: 10.0.3.15:50010
Decommission Status : Normal
Configured Capacity: 11543003135 (10.75 GB)
DFS Used: 182726656 (174.26 MB)
Non DFS Used: 7445495807 (6.93 GB)
DFS Remaining: 3914780672(3.65 GB)
DFS Used%: 1.58%
DFS Remaining%: 33.91%
Last contact: Sat Feb 09 13:34:54 PST 2013
```

4. **Safemode**: Safemode is a state where no changes can be made to the blocks. HDFS cluster is in safemode state during start up because the cluster needs to validate all the blocks and their locations. Once validated, safemode is then disabled.

   The options for safemode command are: `hdfs dfsadmin -safemode [enter | leave | get]`

To enter safemode, execute the following command on your NameNode host machine:

```
hdfs dfsadmin -safemode enter
```

8. Uninstalling HDP

Choose one of the following options to uninstall HDP.

Use the Windows GUI:

1. Open the **Programs and Features** Control Panel Pane.
2. Select the program listed: `Hortonworks Data Platform for Windows`.
3. With that program selected, click on the `Uninstall`option.

Use the Command Line Utility:

On each cluster host, execute the following command from the command shell:

```
msiexec /x MSI_PATH /lv PATH_to_Installer_Log_File DESTROY_DATA=no
```

where

- MSI_PATH is the full path to MSI.
- PATH_to_Installer_Log_File is the full path to Installer log file.

Note that this `msiexec` command retains data from your HDP installation. To delete existing HDP data, set `DESTROY_DATA=yes`.

9. Appendix: Adding a Smoketest User

Creating a smoketest user enables you to run HDP smoke tests without having to run them as the `hadoop` user.

To create a smoketest user:

1. Open a command prompt as the `hadoop` user:

```
runas /user:hadoop cmd
```

2. Change permissions on the MapReduce directory to include other users:

```
%HADOOP_HOME%\bin\hdfs fs -chmod -R 757 /mapred
```

3. Create an HDFS directory for the smoketest user:

```
%HADOOP_HOME%\bin\hdfs dfs -mkdir -p /user/smoketestuser
```

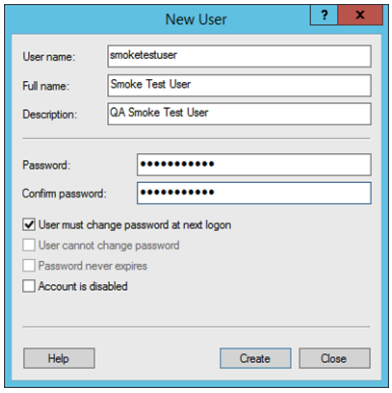
4. Change ownership to the smoketest user.

```
%HADOOP_HOME%\bin\hdfs dfs -chown -R smoketestuser /user/smoketestuser
```

5. Create a smoketest user account in Windows:

   a. Navigate to `Computer Management`.

   b. Select `Local Users and Groups > File > Action > New User` on Windows Server 2008 or `Local Users and Groups > Action > New User` on Windows Server 2012. The New User dialog displays:



   c. Create the user name and password for your `smoketest` user. Determine password requirements and select `Create`.

6. Validate the `smoketest` user by running the smoke tests as the `smoketest` user.

   a. Switch to a command prompt as the `smoketest` user. For example:

```
runas /user:smoketestuser cmd
```

   b. As the `smoketest` user, run the smoke tests:

```
%HADOOP_NODE%\Run-Smoketests.cmd
```