Hortonworks Data Platform

Installing HDP Using Shell Scripts (gsInstaller)

(Jan 14, 2013)

docs.hortonworks.com

Hortonworks Data Platform: Installing HDP Using Shell Scripts (gsInstaller)

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. Getting Ready to Install

This section describes the information and materials you need to get ready to install the Hortonworks Data Platform (HDP) using command line installer (gsInstaller).

Important

gsInstaller is **deprecated** as of HDP 1.2.0 and will not be made available in future minor and major releases of HDP. We encourage you to consider Manual Install (RPMs) or Automated Install (Ambari).

1.1. Meet Minimum System Requirements

Ensure that you meet the following system requirements before installing HDP:

- Hardware recommendations
- Operating system
- Software requirements
- JDK requirements
- Database requirements

1.1.1. Hardware recommendations

The hardware on all Hadoop host machines is assumed to be 64-bit hardware.

Although there is no single hardware requirement for installing HDP, there are some basic guidelines. You can see sample setups here: Suggested Hardware for a Typical Hadoop Cluster.

1.1.2. Operating systems

The following operating systems are supported:

- 64-bit Red Hat Enterprise Linux (RHEL) 5 or 6
- 64-bit CentOS 5 or 6
- 64-bit SUSE Linux Enterprise Server (SLES) 11, SP1

Important

The installer pulls many packages from the base OS repos. If you do not have a complete base OS available to all your machines at the time of installation, you may run into issues. For example, if you are using RHEL 6 your hosts must be able to access the "Red Hat Enterprise Linux Server 6 Optional (RPMs)" repo. If this repo is disabled, the installation is unable to access the rubygems package, which is necessary for HMC to operate. If you encounter problems with base OS

repos being unavailable, please contact your system administrator to arrange for these additional repos to be proxied or mirrored. For more information see Deploying HDP in Production Data Centers with Firewalls.

1.1.3. Software requirements

On each of your hosts:

- yum (RHEL)
- zypper (SLES)
- rpm
- scp
- curl
- wget
- unzip
- tar
- pdsh

Note

Optionally, you can use the auxiliary script `gsPreRequisites.sh` to install `wget` and `curl` on all the nodes.

Ensure that all the ports listed here are available to the Installer.

1.1.4. Database requirements

- To use external database for Hive or Oozie metastore, ensure that a MySQL or Oracle database is deployed and available. (By default, Oozie uses Derby database for its metastore.)
  - For instructions on deploying and/or configuring MySQL database, see here [3].
  - For instructions on configuring an existing Oracle database instance, see here [4].

Note

To deploy an Oracle instance, consult your database administrator.

- Create database users for Hive Metastore and/or Oozie Metastore using one of the following options:
  - **Option I:** Allow HDP to create database users
    1. Ensure that you have root privileges to the database instance.

2. Ensure that you deploy either a MySQL or an Oracle client on the master install machine.
3. On the master install machine, edit the *`master-install-machine/`*`gsInstaller/gsInstaller.properties` file and provide values for the following properties:

Table 1.1. Properties to create database users

Property Name	Notes
dbsysuser	Database system user credentials for Oracle instance. Required if value of `dbflavor` property is set to `oracle`.
dbsyspasswd	
hive_dbuser	Database user credentials for Hive Metastore. Required if installing Hive.
hive_dbpasswd	
oozie_dbname	Database name for Oozie Metastore. Required if installing Oozie and if `oozie_use_external_db` property is set to `yes`. See Reference - Configuration Properties (oozie_use_external_db).
oozie_dbuser	Database user credentials for Oozie Metastore. Required if installing Oozie and if `oozie_use_external_db` property is set to `yes`.
oozie_dbpasswd	

4. On the master install machine, execute the `setupDatabaseUsers` auxiliary script file.

```
sh master-install-machine/gsInstaller/setupDataseUsers.sh
```

- **Option II:** Manually create database users

1. Ensure that your database administrator creates the following databases and users:
   - For Hive:
     a. hive_dbname
     b. hive_dbuser
     c. hive_dbpasswd
   - For Oozie:
     a. oozie_dbname
     b. oozie_dbuser
     c. oozie_dbpasswd
2. On the master install machine, edit the *`master-install-machine/`*`gsInstaller/gsInstaller.properties` file and provide values for the Hive and/or Oozie users (as listed in the table above).

Instructions to setup MySQL database

You can deploy a MySQL instance using one of the following options:

- **Option I:** Allow HDP to deploy MySQL instance

1. On the master install machine, edit the *master-install-machine*/gsInstaller/gsInstaller.properties file and provide values for the following properties:

   ```
   dbflavor=mysql
   ```

   ```
   dbhost=$FQDN_of_MySQL_host_machine
   ```

2. On the master install machine, deploy a MySQL client.
3. Execute the auxiliary script `startMySql.sh`:

   ```
   sh master-install-machine/gsInstaller/startMySql.sh
   ```

- **Option II:** Manually deploy MySQL

1. Connect to the host machine where you plan to deploy MySQL instance and from a terminal window, type:
   - For RHEL and CentOS:

     ```
     yum install mysql-server
     ```

   - For SLES:

     ```
     zypper install mysql
     ```

2. Start the instance.
   - For RHEL and CentOS:

     ```
     /etc/init.d/mysqld start
     ```

   - For SLES:

     ```
     /etc/init.d/mysqld start
     ```

3. Set the `root` user password and remove unnecessary information from `log` and STDOUT

   ```
   mysqladmin -u root password $password
   ```

   ```
   mysqladmin -u root 2>&1 >/dev/null
   ```

Instructions to configure Oracle database

To configure Oracle database instance, use the following instructions:

Note

The following instructions are for OJDBC driver for Oracle 11g.

- On the master install machine, edit the *master-install-machine*/gsInstaller/gsInstaller.properties file and provide values for the following properties:

  ```
  dbflavor=oracle
  ```

  ```
  dbhost=$FQDN_of_Oracle_host_machine
  ```

- Download the Oracle JDBC driver from here.
- If you are manually creating the database users, ensure that your database administrator deploys the Hive schema. The Hive schema file is located here:

```
master-install-location/gsInstaller/confSupport/sql/oracle/hive-schema-0.10.0.oracle.sql
```

1.1.5. JDK requirements

Your system must have the correct JDK installed on all the nodes of the cluster. HDP requires Oracle JDK 1.6 update 31. For more information, see Install the Java Development Kit.

1.2. Decide on Deployment Types

While it is possible to deploy all of HDP on a single host (single node deployment), this is appropriate only for initial evaluation. In general you should use at least three hosts: one master host and two slaves.

Also, see: Typical Hadoop Cluster.

Single node deployment uses separate processes for each of the Hadoop services (NameNode, DataNode, JobTracker, TaskTracker) on a single machine. However, this Hadoop cluster is not truly distributed, because no processing or data storage is performed on remote nodes.

1.3. Prepare the Environment

- Ensure you use the fully qualified domain name (FQDN) for all the host machines. If you are deploying on EC2, use the **Internal hostname**.

Note

Only alphanumeric, hyphen ("-"), and period (“.”) characters are allowed in a valid FQDN. For more details, see: Fully qualified domain name.

- All the host machines in your cluster must be configured for DNS and Reverse DNS.

Note

If you are unable to configure DNS and Reverse DNS, you must edit the hosts file on every host in your cluster to contain each of your hosts.

- Ensure that the Network Time Protocol (NTP) is enabled for your cluster.
- In environments with no access to the Internet, ensure that you make one of your master nodes as NTP server.

1.4. Collect Information

To deploy your HDP installation, you need to collect the following information:

- The fully qualified domain name (FQDN) for each host in your system, and which component(s) you wish to set up on which host. You can use `hostname -f` to check for the FQDN if you do not know it.
- The flavor of database to be used for Hive metastore or Oozie metastore, or Sqoop. (Currently, gsInstaller supports MySQL and Oracle databases).
- FQDN of your database host name.
- If you are using Oracle database, ensure that you have credentials for the database system user.

1.5. Optional - Configure the Local Repositories

If your cluster does not have access to the Internet, or you are creating a large cluster and you want to conserve bandwidth, you need to provide access to the bits using an alternative method. For more information, see Deploying HDP In Production Data Centers with Firewalls

2. Deploying Non Secure Hadoop Cluster

This section provides detailed instructions to deploy a non-secure Hadoop cluster.

2.1. Set Up the Bits

1. Download the HDP Installer from here.

Note

To access the optional Talend tool set:

```
wget http://public-repo-1.hortonworks.com/HDP-1.2.0/tools/HDP-ETL-TOS_BD-V5.1.1.tar.gz
```

2. Expand the archive on the single host machine (also referred as the `master-install-location` in this document):

```
tar zxvf HDP-gsInstaller-1.2.0.21.tar.gz
```

2.2. Define Cluster Details

1. Use `hostname -f` to identify the FQDN for all the host machines.

Note

If you are deploying on Amazon EC2, use the Internal FQDN.

2. On the master-install-location, change directory to `master-install-location/gsInstaller`.
3. Create the following flat text files:

Note

The mandatory files are required for minimal install (Apache Hadoop core components). The optional files are needed if you wish to install that component (for example, HBase, Hive, WebHCat, etc.) in your cluster.

- **Mandatory files:** gateway, namenode, snamenode, jobtracker, nodes

Note

The `nodes` file is used to define the DataNodes and TaskTrackers.

- **Optional files:** hbasemaster, hivemetastore, webhcatnode, nagiosserver, gangliaserver, oozieserver, hbasenodes, zknodes

Note

The `hbasenodes` file is used to define the RegionServers for your HBase cluster.

4. Provide FQDN of your host machines in each these text files:

   - **Option I (single node installations):** Provide the FQDN of the same host machine for all of the text files.
   - **Option II (multi node installations):**

     a. For the following files, provide FQDN of **EXACTLY one host machine**:

     `gateway`, `namenode`, `snamenode`, `jobtracker`, `hbasemaster`, `hivemetastore`, `oozieserver`, `webhcatnode`, `nagiosserver`, `gangliaserver`.

     b. For the following files, provide FQDN (separated by a new-line character) for a **MINIMUM of three host machines:**:

     `nodes`, `hbasenodes`

     c. For the `zknodes` file, provide FQDN for a **MINIMUM of one host machine.**:

     

     Note

     Multiple host machines must follow the Zookeeper ensemble rule.

2.3. Configure Properties

1. Edit the `master-install-location/gsInstaller/gsInstaller.properties` file and specify values for all of the following properties to install all HDP components:

   

   Note

   To perform minimal install (Apache Hadoop core components), specify values for Mandatory properties only (see third column in table).

Table 2.1. gsInstaller Configuration Properties

Property Name	Notes	Example	Mandatory/ Optional/Conditional
java64home	Location of JAVA_HOME for 64-bit JDK v 1.6 update 31 in your environment.		Mandatory
datanode_dir	Comma-separated list of the DataNode's data directories residing on separate disks.	/hdp/1/hadoop/ hdfs/data,/hdp/2/ hadoop/hdfs/data	Mandatory
namenode_dir	Comma-separated list of the NameNode's data directories. Provide multiple directories on separate physical file systems and on separate mount points to preserve the NameNode metadata.	/hdp/1/hadoop/ hdfs/namenode,/ hdp/2/hadoop/hdfs/ namenode	Mandatory
snamenode_dir	Comma-separated list of the NameNode's checkpointing directories residing on separate disks.	/hdp/1/hadoop/ hdfs/snamenode,/ hdp/2/hadoop/hdfs/ snamenode	Mandatory
mapred_dir	Comma-separated list of the MapReduce's data directories residing on separate disks.	/hdp/1/hadoop/ mapred,/hdp/2/ hadoop/mapred	Mandatory

Property Name	Notes	Example	Mandatory/ Optional/Conditional
log_dir	Full path to Hadoop log directory.	`/var/log/hadoop`	Mandatory
pid_dir	Full path to Hadoop PID directory.	`/var/run/hadoop`	Mandatory
sshkey	Either provide full path to the sshkey which allows you to perform passwordless SSH OR Set this field to empty when passwordless SSH is set-up		Mandatory
installpig	To install Pig, set the value to yes. (Default: yes)		Optional
installhbase	To install HBase, set the value to yes. (Default: yes)		Optional
hbase_log_dir	Location for HBase log directory.	`/var/log/hbase`	Conditional. Required if `installhbase` is set to yes.
hbase_pid_dir	Location for HBase PID directory.	`/var/run/hbase`	
zk_log_dir	Location for ZooKeeper log directory.	`/var/log/zookeeper`	Conditional. Required if `installhbase` is set to yes.
zk_pid_dir	Location for ZooKeeper PID directory.	`/var/run/zookeeper`	
zk_data_dir	Location for Zookeeper data directory.	`/hdp/1/hadoop/zookeeper`	
dbflavor	Database flavor for Hive, Oozie, and Sqoop.	Permissible values are mysql or oracle	Conditional. Required only if installing either Hive, Oozie, or Sqoop
dbhost	FQDN for database host machine		Conditional. Required only if installing either Hive, Oozie, or Sqoop
jdbc_jar	Location of the Oracle JDBC JAR file (see instructions here)		Conditional. Required only if installing either Hive, Oozie, or Sqoop and if `dbflavor` is set to `oracle`
dbsysuser	User credentials for Oracle database system users		
dbsyspasswd			
installhive	To install Hive, set the value to yes. (Default: no). You must first install and configure MySQL or Oracle for your cluster as instructed here. Also ensure that `installhcat` property is set to yes.		Optional.
hive_log_dir	Location for Hive log directory.	`/var/log/hive`	Conditional. Required only if `installhive` is set to yes.
hive_dbname	Database name for Hive metastore.		Conditional. Required only if `installhive` is set to yes.
hive_dbuser	Credentials for Hive Database user.		
hive_dbpasswd			
installhcat	To install HCatalog, set the value to yes. (Default: yes)		Optional. Ensure that `installhive` property is set to yes.
smoke_test_user	Provide the value for the smoke test user. (Default: `hdptestuser`)		Mandatory
installwebhcat	To install WebHCat, set the value to yes. (Default: yes)		Conditional. Ensure that `installpig`, `installhive`, and `installhcat` properties are set to `yes`.

Property Name	Notes	Example	Mandatory/ Optional/Conditional
webhcat_log_dir	Location for WebHCat log directory.	`/var/log/webhcat`	Required only if installing WebHCat.
webhcat_pid_dir	Location for WebHCat PID directory.	`/var/run/webhcat`	
installsqoop	To install Sqoop, set the value to yes. (Default: yes)		Optional
installoozie	To install Oozie, set the value to yes. (Default: yes)		Optional
oozie_use_external_db	Set this to yes, to use MySQL or Oracle database for Oozie metastore (default database is Derby). (Default: no)		Conditional. Required only if installing Oozie
oozie_log_dir	Location for Oozie log directory.	`/var/log/oozie`	
oozie_pid_dir	Location for Oozie PID directory.	`/var/run/oozie`	
oozie_dbname	Database name for Oozie metastore.		
oozie_dbuser	Database user credentials for Oozie metastore.		
oozie_dbpasswd			
installFlume	To download Apache Flume RPM, set the value to yes. (Default: no). To deploy Flume, use the instructions available here.		Optional
installMahout	To install Apache Mahout set the value to yes.(Default: no)		Optional
enablemon	Required if you want to install the HDP components for Monitoring (Ganglia and Nagios). (Default: yes)		Optional. You must configure and deploy HDP Monitoring components as instructed in the next section.

2.4. Optional - Configure and deploy HDP components for Monitoring

.

Note

This step is mandatory only if `enablemon` property is set to `yes`.

1. Edit the `master-install-location/gsInstaller/monInstaller.properties` file and provide values for the following properties:

Table 2.2. gsInstaller Monitoring Properties

Property Name	Notes
installnagios	Set this to yes, to install Nagios Server. (Default: yes)
installsnmp	Set this to yes, if `installnagios` property is set to yes. (Default: yes)
installganglia	Set this to yes, to install Ganglia Server. (Default: yes)
snmpcommunity	Provide the name of the SNMP community. (Default: `hadoop`).
snmpsource	Used to configure source in `snmpd.conf`. You can use either a host or network addresses in CIDR notation. (For example: `192.168.0.0/24` means all the machines from `192.168.0.0` to `192.168.0.255` are allowed to access data from snmp daemons). Ensure that both the Gateway and the Nagios host machine belong to snmpsource address range.

Property Name	Notes
nagioscontact	Provide valid email address for receiving Nagios alerts. (Default: `monitor \@monitor.com`).
gmetad_user	Provide value for the Ganglia gmetad_user.(Default: `nobody`)
gmond_user	Provide value for the Ganglia gmond_user (Default: `nobody`)
webserver_group	Provide the value for Ganglia webserver group. (Default: `apache`). If you are deploying HDP on SLES, change the default value to `www`.

2. Deploy HDP components for monitoring:

```
cd master-install-location/gsInstaller
sh monInstaller.sh
```

2.5. Set Up Your Cluster

- Option I: Allow HDP to set-up the cluster.
  1. Set the `localyumrepo` property in the `gsInstaller.properties` file to `yes`.
  2. Execute the auxiliary script file - `gsPreRequisites.sh`.

     ```
     sh master-install-location/gsInstaller/gsPreRequisites.sh
     ```
- Option II: Manually set up the cluster.
  1. Configure the local mirror repository as instructed here.
  2. For Red Hat compatible systems only:
     - Disable SELinux on all the host machines:

       ```
       sed 's/SELINUX=enforcing/SELINUX=disabled/g'/etc/selinux/config/usr/
       sbin/setenforce 0
       ```
     - Disable firewall on all the host machines:

       ```
       /etc/init.d/iptables stop
       ```

2.6. Create Users

1. Execute the auxiliary helper script `createUsers.sh`:

```
sh master-install-location/gsInstaller/createUsers.sh
```

Note

By default, the home directory for the service users created by HDP Installer will point to `/usr/lib/hadoop`. Use the auxiliary helper script (`createUsers.sh`) to change the home directory for the service users.

2.7. Start the Installation

1. As root user, execute the following command:

```
sh master-install-location/gsInstaller/gsInstaller.sh
```

2. Confirm the set-up properties. (Type **y** or **Y** and press **Enter**.)

   This step launches the HDP Installer. Depending on the number of nodes in your cluster, this step can take couple of minutes to complete the smoke tests.

Note

The value of NameNode new generation size (default size of Java new generation for NameNode (Java option `-XX:NewSize`)) should be 1/8 of maximum heap size (`-Xmx`). To change the default setting, modify the `namenode_opt_newsize` property in the `master-install-location/gsInstaller/gsCluster.properties` file. Ensure that the value of the `namenode_opt_newsize` property is 1/8 the value of maximum heap size (`-Xmx`). For more details on `gsCluster.properties` file, see Configuration Cluster Properties. Also ensure that your NameNode and secondary NameNode have identical memory settings.

2.8. Verify Installation

- Your HDP deployment is successful if the smoke tests for all the components pass successfully. To verify that your map-reduce tasks were successfully completed, browse the web interfaces for the NameNode, JobTracker, and HBase. The default locations for these interfaces are as listed below:
  - NameNode - `http://$NameNodeHost:50070/`
  - JobTracker - `http://$JobTrackerHost:50030/`
  - HBase Master Web Interface -`http://$HBaseMasterHost:60010/`
- Test access to the Ganglia server. Browse to the Ganglia server:

  ```
  http://$FQDN_for_ganglia_server/ganglia
  ```

  where `$FQDN_for_ganglia_server` is specified in the `gangliaserver` flat text file.
- Test access to the Nagios server.
  1. Browse to the Nagios server:

     ```
     http://$FQDN_for_nagios_server/nagios
     ```

     where `$FQDN_for_nagios_server` is specified in the `nagiosserver` flat text file.
  2. Login using the Nagios admin username (`nagiosadmin`) and password.
  3. Click on

     ```
     hosts
     ```

     to validate that all the hosts in the cluster are listed.
  4. Click on

     ```
     services
     ```

to validate all the Hadoop services are listed for each host.

3. Deploying Secure Hadoop Cluster

This section describes deploying Hadoop in a secure cluster environment.

3.1. Prerequisites

In addition to the prerequisites provided here, ensure that you also meet the following prerequisites for secure deployments:

- Ensure that the UNIX users (responsible for job submission) have the user ID greater than `1000`.

Note

We strongly discourage usage of Hadoop service users (hdfs, hbase, hcat, mapred) for submitting jobs. You must instead use separate UNIX users for job submissions.

- Install security policy JAR files. Download the security policy JAR files from here.

Note

These JAR files must be present under `$JAVA_HOME/jre/lib/security/s` directory

- Ensure that you replace the instances of `EXAMPLE.COM` (under the `realm` property) in `gsInstaller.propertiess` file with the actual value of the `realm` defined in your `krb5.conf` file.

3.2. Install Implications for Deploying Secure Hadoop Clusters

Security in Hadoop

With Hadoop's new security features and its integration with Kerberos, it is possible to verify that the user is who they claim to be and ensure they only have the correct access to data or resources. This allows corporations to allow finer grained access to information and reduce their operational overhead by coalescing their distinct clusters.

Secure Hadoop clusters provide solutions for the following threats:

- Prevent unauthorized access to HDFS and MapReduce communication
- Prevent unauthorized access to the jobs submitted through Oozie
- Prohibit the fraudulent servers to access your Hadoop cluster
- Prevent impersonation attacks
- Prevent access to root accounts

Deployment options for secure Hadoop cluster

Depending on your environment set-up, following are the two different options to install a secure Hadoop cluster:

- **OPTION I:** Set-up a new Kerberos Key Distribution Center

  Use the auxiliary script - `setupKerberos.sh`. This auxiliary script file is responsible for performing following tasks:

  - Sets up a new Key Distribution Center (KDC) on the host machine specified in `kdc-server` file.
  - Creates service keytabs for all processes - NameNode, JobTracker, Secondary NameNode, DataNodes, TaskTrackers, HBase Master, HBase Regionserver, and Hive Metastore
  - Places all the service keytabs (for respective hosts) under `/etc/security/keytabs` directory
  - Generates user keytabs for HDFS and `Smoke Test` users and places these keytab files to `/tmp` directory on all the nodes.

- **OPTION II:** Add existing Kerberos Key Distribution Center

  You also have the option of adding an existing Kerberos Key Distribution Center for your Hadoop cluster.

3.3. Option I - Set up New KDC

This section provides instructions to set up a new KDC and deploy a secure Hadoop cluster.

3.3.1. Set Up the Bits

1. Download the HDP Installer from here.

Note

To access the optional Talend tool set:

```
wget http://public-repo-1.hortonworks.com/HDP-1.2.0/tools/HDP-ETL-TOS_BD-V5.1.1.tar.gz
```

2. Expand the archive on the single host machine (also referred as the `master-install-location` in this document):

```
tar zxvf HDP-gsInstaller-1.2.0.21.tar.gz
```

3.3.2. Install New KDC Server

1. Create the `kdcserver` flat text file under the `master-install-location/gsInstaller` directory.
2. Populate the hostname of your master-install-machine for KDC in the `kdcserver` file.

3. Execute the auxiliary script file `setupKerberos.sh`:

```
sh master-install-location/gsInstaller/setupKerberos.sh
```

4. Provide the database master key.

   This step will push the generated keytabs to the respective host machines.

3.3.3. Define the Cluster Details

1. Use `hostname -f` to identify the FQDN for all the host machines.

Note

If you are deploying on Amazon EC2, use the Internal FQDN.

2. On the master-install-location, change directory to `master-install-location/gsInstaller`.

3. Create the following flat text files:

Note

The mandatory files are required for minimal install (Apache Hadoop core components). The optional files are needed if you wish to install that component (for example, HBase, Hive, WebHCat, etc.) in your cluster.

- **Mandatory files:** gateway, namenode, snamenode, jobtracker, nodes

Note

The `nodes` file is used to define the DataNodes and TaskTrackers.

- **Optional files:** hbasemaster, hivemetastore, webhcatnode, nagiosserver, gangliaserver, oozieserver, hbasenodes, zknodes

Note

The `hbasenodes` file is used to define the RegionServers for your HBase cluster.

4. Provide FQDN of your host machines in each these text files:

   - Option I (single node installations): Provide the FQDN of the same host machine for all of the text files.

   - Option II (multi node installations):

     a. For the following files, provide FQDN of **EXACTLY one host machine**:

        `gateway`, `namenode`, `snamenode`, `jobtracker`, `hbasemaster`, `hivemetastore`, `oozieserver`, `webhcatnode`, `nagiosserver`, `gangliaserver`.

b. For the following files, provide FQDN (separated by a new-line character) for a **MINIMUM of three host machines:**:

   `nodes, hbasenodes`

c. For the `zknodes` file, provide FQDN for a **MINIMUM of one host machine**. Ensure that multiple host machines follow the Zookeeper ensemble rule.

3.3.4. Configure Properties

1. Edit the `master-install-location/gsInstaller/gsInstaller.properties` file and specify values for all of the following properties to install all HDP components:

Note

To perform minimal install (Apache Hadoop core components), specify values for Mandatory properties only (see third column in table).

Table 3.1. gsInstaller Configuration Properties

Property Name	Notes	Example	Mandatory/ Optional/Conditional
java64home	Location of JAVA_HOME for 64-bit JDK v 1.6 update 31 in your environment.		Mandatory
security	Set the value to "yes". (Default: No)		Mandatory
datanode_dir	Comma-separated list of the DataNode's data directories residing on separate disks.	`/hdp/1/hadoop/hdfs/data,/hdp/2/hadoop/hdfs/data`	Mandatory
namenode_dir	Comma-separated list of the NameNode's data directories. Provide multiple directories on separate physical file systems and on separate mount points to preserve the NameNode metadata.	`/hdp/1/hadoop/hdfs/namenode, /hdp/2/hadoop/hdfs/namenode`	Mandatory
snamenode_dir	Comma-separated list of the NameNode's checkpointing directories residing on separate disks.	`/hdp/1/hadoop/hdfs/snamenode, /hdp/2/hadoop/hdfs/snamenode`	Mandatory
mapred_dir	Comma-separated list of the MapReduce's data directories residing on separate disks.	`/hdp/1/hadoop/mapred,/hdp/2/hadoop/mapred`	Mandatory
log_dir	Full path to Hadoop log directory.	`/var/log/hadoop`	Mandatory
pid_dir	Full path to Hadoop PID directory.	`/var/run/hadoop`	Mandatory
hdfs_user_keytab	Change this path to`/tmp/${hdfsuser}.headless.keytab`		Mandatory
keytabdir	Path to NameNode, Secondary NameNode, JobTracker, DataNode, TaskTracker, HBase Master, and RegionServer keytab file. (Default: `/etc/security/keytabs`)		Mandatory
realm	Provide the Kerberos realm. (Default: `EXAMPLE.COM`)		Mandatory
sshkey	Either provide full path to the sshkey which allows you to perform passwordless SSH OR		Mandatory

Property Name	Notes	Example	Mandatory/ Optional/Conditional
	Set this field to empty when passwordless SSH is set-up		
installpig	To install Pig, set the value to yes. (Default: yes)		Optional
installhbase	To install HBase, set the value to yes. (Default: yes)		Optional
hbase_log_dir	Location for HBase log directory.	`/var/log/hbase`	Conditional. Required if `installhbase` is set to yes.
hbase_pid_dir	Location for HBase PID directory.	`/var/run/hbase`	
zk_log_dir	Location for ZooKeeper log directory.	`/var/log/zookeeper`	Conditional. Required if `installhbase` is set to yes.
zk_pid_dir	Location for ZooKeeper PID directory.	`/var/run/zookeeper`	
zk_data_dir	Location for Zookeeper data directory.	`/hdp/1/hadoop/zookeeper`	
dbflavor	Database flavor for Hive, Oozie, and Sqoop.	Permissible values are mysql or oracle	Conditional. Required only if installing either Hive, Oozie, or Sqoop
dbhost	FQDN for database host machine		Conditional. Required only if installing either Hive, Oozie, or Sqoop
jdbc_jar	Location of the Oracle JDBC JAR file (see instructions here)		Conditional. Required only if installing either Hive, Oozie, or Sqoop and if `dbflavor` is set to `oracle`
dbsysuser	User credentials for Oracle database system users		
dbsyspasswd			
installhive	To install Hive, set the value to yes. (Default: no)		Optional. You must first install and configure MySQL for your cluster as instructed here. Also ensure that `installhcat` property is set to yes.
hive_log_dir	Location for Hive log directory.	`/var/log/hive`	Conditional Required only if `installhive` is set to yes.
hive_dbname	Database name for Hive metastore.		Conditional. Required only if `installhive` is set to yes.
hive_dbuser	Credentials for Hive Database user.		
hive_dbpasswd			
installhcat	To install HCatalog, set the value to yes. (Default: yes)		Optional. Ensure that `installhive` property is set to yes.
smoke_test_user	Provide the value for the smoke test user. (Default: `hdptestuser`)		Mandatory
smoke_test_user_keytab	Location of the keytab file for the smoke test user. (Default: `/homes/${smoke_test_user}/${smoke_test_user}.headless.keytab`)		Mandatory
installsqoop	To install Sqoop, set the value to yes. (Default: yes)		Optional

Property Name	Notes	Example	Mandatory/ Optional/Conditional
installoozie	To install Oozie, set the value to yes. (Default: yes)		Optional
oozie_use_external_db	Set this to yes, to use MySQL or Oracle database for Oozie metastore (default database is Derby). (Default: no)		Conditional. Required only if installing Oozie
oozie_log_dir	Location for Oozie log directory.	`/var/log/oozie`	
oozie_pid_dir	Location for Oozie PID directory.	`/var/run/oozie`	
oozie_dbname	Database name for Oozie metastore.		
oozie_dbuser	Database user credentials for Oozie metastore.		
oozie_dbpasswd			
installFlume	To download Apache Flume RPM, set the value to yes. (Default: no). To deploy Flume, use the instructions available here.		Optional
installMahout	To install Apache Mahout set the value to yes.(Default: no)		Optional
enablemon	Required if you want to install the HDP components for Monitoring (Ganglia and Nagios). (Default: yes)		Optional. You must configure and deploy HDP Monitoring components as instructed in next section.

2. To deploy Flume, use the instructions available here.

3.3.5. Optional - Configure and deploy HDP components for Monitoring

Note

This step is mandatory only if `enablemon` property is set to `yes`.

1. Edit the `master-install-location/gsInstaller/monInstaller.properties` file and provide values for the following properties:

Table 3.2. gsInstaller Monitoring Properties

Property Name	Notes
installnagios	Set this to yes, to install Nagios Server.(Default: yes)
installsnmp	Set this to yes, if `installnagios` property is set to yes. (Default: yes)
installganglia	Set this to yes, to install Ganglia Server. (Default: yes)
snmpcommunity	Provide the name of the SNMP community. (Default: `hadoop`).
snmpsource	Used to configure source in `snmpd.conf`.You can use either a host or network addresses in CIDR notation. (For example: `192.168.0.0/24` means all the machines from `192.168.0.0` to `192.168.0.255` are allowed to access data from snmp daemons). Ensure that both the Gateway and the Nagios host machine belong to snmpsource address range.
nagioscontact	Provide valid email address for receiving Nagios alerts. (Default: `monitor \@monitor.com`).
gmetad_user	Provide value for the Ganglia gmetad_user.(Default: `nobody`)
gmond_user	Provide value for the Ganglia gmond_user (Default: `nobody`)

Property Name	Notes
webserver_group	Provide the value for Ganglia webserver group. (Default: apache). If you are deploying HDP on SLES, change the default value to www.

2. Deploy HDP components for monitoring:

```
cd master-install-location/gsInstaller
sh monInstaller.sh
```

3.3.6. Set Up Your Cluster

- Option I: Allow HDP to set-up the cluster

  1. Set the `localyumrepo` property in the `gsInstaller.properties` file to `yes`.
  2. Execute the auxiliary script file - `gsPreRequisites.sh`.

     ```
     sh master-install-location/gsInstaller/gsPreRequisites.sh
     ```

- Option II: Manually set up the cluster.

  1. Configure the local mirror repository as instructed here.
  2. For Red Hat compatible systems only:

     - Disable SELinux on all the host machines:

       ```
       sed 's/SELINUX=enforcing/SELINUX=disabled/g'/etc/selinux/config/usr/
       sbin/setenforce 0
       ```

     - Disable firewall on all the host machines:

       ```
       /etc/init.d/iptables stop
       ```

3.3.7. Create Users

1. Execute the auxiliary helper script `createUsers.sh`:

   ```
   sh master-install-location/gsInstaller/createUsers.sh
   ```

Note

By default, the home directory for the service users created by HDP Installer will point to `/usr/lib/hadoop`. Use the auxiliary helper script (`createUsers.sh`) to change the home directory for the service users.

3.3.8. Start the Installation

1. As root user, execute the following command:

   ```
   sh master-install-location/gsInstaller/gsInstaller.sh
   ```

2. Confirm the set-up properties. (Type **y** or **Y** and press **Enter**.)

   This step launches the HDP Installer. Depending on the number of nodes in your cluster, this step can take couple of minutes to complete the smoke tests.

Note

The value of NameNode new generation size (default size of Java new generation for NameNode (Java option `-XX:NewSize`)) should be 1/8 of maximum heap size (`-Xmx`). To change the default setting, modify the `namenode_opt_newsize` property in the `master-install-location/gsInstaller/gsCluster.properties` file. Ensure that the value of the `namenode_opt_newsize` property is 1/8 the value of maximum heap size (`-Xmx`). For more details on `gsCluster.properties` file, see Configuration Cluster Properties. Also ensure that your NameNode and secondary NameNode have identical memory settings.

3.3.9. Verify Installation

- Your HDP deployment is successful if the smoke tests for all the components pass successfully. To verify that your map-reduce tasks were successfully completed, browse the web interfaces for the NameNode, JobTracker, and HBase. The default locations for these interfaces are as listed below:
  - NameNode - `http://$NameNodeHost:50070/`
  - JobTracker - `http://$JobTrackerHost:50030/`
  - HBase Master Web Interface -`http://$HBaseMasterHost:60010/`
- Test access to the Ganglia server. Browse to the Ganglia server:

  ```
  http://$FQDN_for_ganglia_server/ganglia
  ```

  where `$FQDN_for_ganglia_server` is specified in the `gangliaserver` flat text file.
- Test access to the Nagios server.
  1. Browse to the Nagios server:

     ```
     http://$FQDN_for_nagios_server/nagios
     ```

     where `$FQDN_for_nagios_server` is specified in the `nagiosserver` flat text file.
  2. Login using the Nagios admin username (`nagiosadmin`) and password.
  3. Click on

     ```
     hosts
     ```

     to validate that all the hosts in the cluster are listed.
  4. Click on

     ```
     services
     ```

     to validate all the Hadoop services are listed for each host.

3.4. Option II - Add existing KDC

Thsi section provides instructions to add existing KDC and deploy a secure Hadoop cluster.

3.4.1. Set up the bits

1. Download the HDP Installer from here.

Note

To access the optional Talend tool set:

```
wget  http://public-repo-1.hortonworks.com/HDP-1.2.0/tools/HDP-ETL-TOS_BD-V5.1.1.tar.gz
```

2. Expand the archive on the single host machine (also referred as the `master-install-location` in this document):

```
tar zxvf HDP-gsInstaller-1.2.0.21.tar.gz
```

3.4.2. Configure your existing KDC server

1. Modify the realm property for `krb5.conf` and `kdc.conf` file.

Note

Ensure that the realm value in your `krb5.conf` file matches with the default value for `realm` property in your `gsInstaller.properties` file.

Update the realm property in the `kdc.conf` file and copy this file to each node in your cluster.

2. Create the principals for all services in your Hadoop cluster.

- Execute the following command:

```
kadmin: addprinc -randkey $$principal_name/fully.qualified.domain.name@YOUR-REALM.COM
```

- The *$principal_name* must use the following mandatory naming conventions:

Important

Upper case name in the FQDN part of the principals does not work with Kerberos. (JIRA: HADOOP-7988)

Table 3.3. Secure deployment - Mandatory naming conventions for principals

Service Name	Principal Name (mandatory naming convention)
NameNode	nn and HTTP
Secondary NameNode	nn, HTTP

Service Name	Principal Name (mandatory naming convention)
JobTracker	jt
TaskTracker	tt
DataNode	dn
HBase Master	hm
HBase RegionServer	rs
Hive Metastore	hive
Oozie	oozie
Oozie	HTTP
WebHCat	HTTP

3. Set permissions for keytabs directory to 755.

Note

The location of the keytab directory is specified by the `keytabdir` property in the `gsInstaller.properties` file.

4. Create keytab files for all services and assign required permissions.

```
kadmin: xst -norandkey -k $keytab_file_name $principal_name/fully.qualified.
domain.name
```

where, the *`$keytab_file_name`* uses the following mandatory naming conventions:

Table 3.4. Secure deployment - Mandatory naming conventions for keytab files

Service Name	Keytab File Name	Principal Name	Permissions	Owner
NameNode	`nn.service.keytab`	`nn`	`700`	`hdfs:hadoop`
NameNode	`spnego.service.keytab`	`HTTP`	`700`	`hdfs:hadoop`
Secondary NameNode	`nn.service.keytab`	`nn`	`700`	`hdfs:hadoop`
Secondary NameNode	`spnego.service.keytab`	`HTTP`	`700`	`hdfs:hadoop`
JobTracker	`jt.service.keytab`	`jt`	`700`	`mapred:hadoop`
TaskTracker	`tt.service.keytab`	`tt`	`700`	`mapred:hadoop`
DataNode	`dn.service.keytab`	`dn`	`700`	`hdfs:hadoop`
HBase Master	`hm.service.keytab`	`hm`	`700`	`hbase:hadoop`
HBase RegionServer	`rs.service.keytab`	`rs`	`700`	`hbase:hadoop`
Hive Metastore	`hive.service.keytab`	`hive`	`700`	`hive:hadoop`
Oozie	`oozie.service.keytab`	`oozie`	`700`	`oozie:hadoop`
Oozie	`spnego.service.keytab`	`HTTP`	`700`	`oozie:hadoop`
WebHCat	`spnego.service.keytab`	`HTTP`	`700`	`webhcat:hadoop`

5. On each of the node in your cluster, change directory to the keytab directory (the location is specified by the `keytabdir` property in the `gsInstaller.properties` file).

6. Copy the appropriate keytab file on each of node in your cluster.

7. Follow the steps listed in Step 2-2 through Step 2-4 above for creating keytab files and principals for the following users. Use the following mandatory naming conventions:

Table 3.5. Secure deployment - Mandatory naming conventions for HDFS service and Smoke test user's keytabs and host principals

User	Keytab File Name	Principal Name	Permissions
`hdfs:hadoop`	*`$hdfsuser`*`.headless.keytab`	`hdfs`	`700`
`hdptestuser:hadoop`	*`$smoke_test_user`*`.headless.keytab`	*`$smoke_test_user`*	`700`

3.4.3. Deploy HDP

1. Use the steps 3 through 9 for OPTION I to deploy secure Hadoop cluster.

Note

In addition to configuring the properties provided in OPTION I - Step 4, you must also provide value for the following property:

Table 3.6. Secure deployment - Configuring HDFS user keytab file for secure Hadoop cluster using Option II

Property Name	Notes	Example	Mandatory/ Optional/ Conditional
hdfs_user_keytab	Location for the HDFS user keytab file.	`/tmp/`*`$hdfsuser`*`.headless.keytab`	`Mandatory`

4. Troubleshooting gsInstaller Deployments

This section describes common troubleshooting steps to remedy installer issues.

4.1. Getting the logs

Use the following instructions to view the log file for gsInstaller:

Important

If you encounter any issues with your installation, DO NOT ATTEMPT to recover the any of the files manually. Follow the steps for RPM based uninstallation and start the installation again with a clean image.

1. Find the process ID appended to the HDP Installer's log file:

```
grep -ir "gsInstaller*" /tmp
```

2. Open the file in an editor of your choice.

3. The location for all the log files for each of the components (HDFS, MapReduce, HBase, HCatalog, WebHCat, Oozie, Sqoop, Nagios, Ganglia) are specified in the `gsInstaller.properties` file.

To manually start or stop Hadoop services (NameNode, DataNodes, JobTracker, TaskTrackers, HBase Master, RegionServers, Zookeeper, and HCatalog) refer to the instructions provided here: Manage Services Manually.

4.2. Quick Checks

- If you are sharing the same host machine for NameNode, Secondary NameNode, JobTracker, HCatalog Server, and HBase Master, ensure that you setup SSH for localhost.
- Make sure the directories to which gsInstaller needs to write information are writable. The locations for all the directories are specified in the `gsInstaller.properties` file.
- Make sure all the appropriate services are running. For information on how to do this, see Manage Services Manually.
- If the first HDFS **put** command fails to replicate the block, the clocks in the nodes may not be synchronized. Make sure that Network Time Protocol (NTP) is enabled for your cluster.
- If HBase does not start, check if its slaves (as specified in the `nodes` and `hbasenodes` flat text files) are running on 64-bit JVMs. The ZooKeeper service must run on a 64-bit host machine.
- Make sure the hosts specified in the flat text files are listed as FQDN, not IP addresses.

- Make sure umask is set to 0022.
- Make sure the HCatalog host can access the MySQL server. From a shell try:

```
mysqld -h $FQDN_for_MySQL_server -u $FQDN_for_HCatalog_Server -p
```

  You will need to provide the password you set up for Hive/HCatalog during the installation process.

- Make sure MySQL is running. By default, MySQL server does not start automatically on reboot.

  To set auto-start on boot, from a shell, type:

```
chkconfig --level 35 mysql on
```

  To then start the service manually from a shell, type:

```
service mysqld start
```

4.3. Specific Issues

The following are common issues you might encounter:

4.3.1. DataNodes smoke test failures

If your DataNodes are incorrectly configured, the smoke tests fail and you get this error message in the DataNode logs::

```
DisallowedDataNodeException org.apache.hadoop.hdfs.server.protocol.
DisallowedDatanodeException
```

Solution:

1. Ensure that reverse DNS look-up is properly configured.
2. Ensure that the DataNodes are correctly specified in the `nodes` file (located here: `master-install-location/gsInstaller`)).
3. Restart the installation process.

4.3.2. Metastore startup failed for HCatalog Daemon

If the HCatalog daemon is incorrectly configured, the smoke tests fail and you get the following error message on your console:

```
Metastore startup failed, see /var/log/hcatalog/hcat.err
```

Solution: You can find the root cause of this exception using the `/var/log/hcatalog/hcat.err` file on your HCatalog node. Typically, the root cause is either `Unknown Host Exception` or `Failed Initializing Database Error`.

- **Solution - Unknown Host Exception:**

1. Open `/var/log/hcatalog/hcat.err` file to confirm the root cause:

```
 at java.lang.reflect.Method.invoke (Method.java:597)
  at org.apache.hadoop.util.Runjar.main (runjar.java:156)
  Caused by: java.net.UnknownHostException:mysql.host.com atjava.net.
InetAddress.getAllByName(INetAddress.java:1157)
```

2. Open the `master-install-location/gsInstaller/gsInstaller.properties` file in edit mode and update the value for `mysqldbhost` property.

3. Restart the installation process.

- **Solution - Failed Initializing Database Error**

1. For this error, you should find a message similar to the one shown below in your `/var/log/hcatalog/hcat.err` file:

```
11/12/29 20:52:04 ERROR DataNucleus.Plugin: Bundle “org.eclipse.jdt.core”
 required
11/12/29 20:52:04 ERROR DataNucleus.Plugin: Bundle “org.eclipse.jdt.core”
 required
11/12/29 20:52:04 ERROR DataStore.Schema: Failed initialising database
```

2. Open the `master-install-location/gsInstaller/gsInstaller.properties` file in edit mode and update the value for `mysqldbuser` and `mysqldbpswd` properties.

Important

Ensure that *`$mysqluser`* has required privileges on the MySQL database as provided under the `databasename` property in the `gsInstaller.properties` file.

3. Restart the installation process.

4.3.3. Failures caused by incorrect HCatalog configurations

If the HCatalog configuration are incorrect you get an error message similar to the following on your console:

```
unzip: cannot find zipfile directory in one of /tmp/mysqljdbc.zip or /tmp/
mysqljdbc.zip.zip, and cannot find /tmp/mysqljdbc.zip.ZIP, period on x.x.x.x
```

Solution: You can find the root cause of this exception using the `/var/log/hcatalog/hcat.err` file on your HCatalog node. Typically, the root cause is either `Unknown Host Exception` or `Failed Initializing Database Error`.

1. Copy the link of the nearest accessible mirror for MySQL JDBC Connector from here.

2. Open the `master-install-location/gsInstaller/gsLib.sh` file in edit mode and update the value for `mysqljdbcurl` property to the value obtained in Step -1. This will change the default location of the MySQL JDBC Connector in gsInstaller

3. Restart the installation process.

4.3.4. Secure deployment failures

Secure deployments typically fail because of the following errors:

- No valid credential provided
- GSS Initiate failed
- **Solution - No valid credential provided:** If your `kinit` utility is incorrectly configured, you see the following error message during secure deployment:

```
Caused by: java.io.IoException: javax.security.SaslException:GSS Initiate
failed (Caused by GSSException: Failed to find any Kerberos tgt)
```

1. Locate the correct path for the kinit utility.

```
cd $Full_Path_To_KDC_Server
which kinit
```

2. Open the `master-install-location/gsInstallergsInstaller.properties` file in edit mode.
3. Replace the existing value for `kinitpath` property with the value obtained in Step-1.
4. Restart the installation process.

- **Solution - GSS Initiate failed:**

1. If your keytab file configurations are incorrect, you should see the following error message during secure deployment:

```
Caused by: java.io.IoException: javax.security.SaslException:GSS Initiate
 failed (Caused by GSSException: Failed to find any Kerberos tgt)
```

Note

The auxiliary script file `setupKerberos.sh` pushes all the keytabs under `/tmp` directory of each host in your cluster. You must ensure that the `gsInstaller.properties` reflects the correct values for the keytab properties.

2. Open the `master-install-location/gsInstaller/gsInstaller.properties` file in edit mode and update the value for `hdfsuser.headless.keytab` and `smoke_test_user.headless.keytab` properties.
3. Restart the installation process.

More specific issues for secure deployments are discussed here:

- MapReduce smoke test failures

- Secure DataNodes errors

4.3.4.1. Secure MapReduce smoke test failures

If the user IDs for the service users (HDFS, MapReduce, HCatalog, HBase) and/or unprivileged users (users responsible for job submissions, executing Pig or Hive queries) is less than 1000, the Mapreduce tasks for smoke tests fail and you get the following error message:

```
Error initializing attempt_201112292220_0001_m_000002_0:
……
Caused by: org.apache.hadoop.util.Shell$ExitCodeException:
at org.apache.hadoop.util.Shell.runCommand(Shell.java:255)
at org.apache.hadoop.util.Shell.run(Shell.java:182)
at org.apache.hadoop.util.Shell$ShellCommandExecutor.execute(Shell.
java:375) at org.apache.hadoop.mapred.LinuxTaskController.initializ
Job(LinuxTaskController.java:185)
```

Solution:

1. Open the TaskTracker log file to analyze root cause. Find the incorrectly configured user ID by looking at the error message similar to the one shown below:

```
INFO org.apache.hadoop.mapred.TaskController: Reading task controller
 configuration /etc/hadoop/taskcontroller.cfg
INFO. org.apache.hadoop.mapred.TaskController: requested user
hdfs has id 201, which is below the minimum allowed 1000
```

2. Change the User ID for the *$user_name* obtained in Step-1 above.

```
cd master-install-location/gsInstaller
usermod -u 10000 $user_name
```

3. Restart the installation process.

4.3.4.2. Secure DataNodes errors

During secure deployment, you might get a generic error message as shown below:

```
Caused by: java.io.IoException: javax.security.SaslException:GSS Initiate
failed (Caused by GSSException: Failed to find any Kerberos tgt)
```

Solution:

1. If your SE Linux is incorrectly configured, you will see the following error message in the `jsvc.err` file:

```
17/01/2012 18:49:39 19465 jsvc.exec error: Cannot dynamically link to /usr/
hadoop-jdk1.6.0_26/jre/lib/i386/server/libjvm.so
17/01/2012 18:49:39 19465 jsvc.exec error: /usr/hadoop
jdk1.6.0_26/jre/lib/i386/server/libjvm.so: cannot restore segment prot after
 reloc: Permission denied
17/01/2012 18:49:39 19426 jsvc.exec error: Service exit with a return value
 of 1
```

2. Execute following command as root user for that DataNode:

```
ssh $DataNode
su -
sed -i 's|SELINUX=enforcing|SELINUX=disabled|g' /etc/selinux/config/usr/
sbin/setenforce 0
```

3. Restart the installation process.

4.4. Hadoop streaming jobs issue with WebHCat

Solution:

1. Check if the file `/user/webhcat/hadoop-streaming.jar` exists on HDFS:

```
su - webhcat
hadoop dfs -ls /user/webhcat/hadoop-streaming.jar
```

2. If the above command fails, copy the Hadoop streaming JAR file:

```
su - webhcat

/usr/bin/hadoop --config ${hadoopconfdir} fs -copyFromLocal  /usr/share/
hadoop/contrib/streaming/hadoop-streaming*.jar /user/webhcat/hadoop-
streaming.jar
```

5. Reference

This section provides reference information for use with the gsInstaller scripts.

5.1. Configuration Properties

The Hortonworks Data Platform (HDP) Installer properties are grouped into two categories and each category, in turn, exposes several groups.

These categories are based on the essential and optional components distributed with HDP. To learn more about these components, see: About Hortonworks Data Platform.

5.1.1. Category I - HDP Essential Components Properties

This category exposes the configuration properties for the essential components (HDFS, MapReduce, HBase, HCatalog, Pig, Hive, Zookeeper, and WebHCat) in HDP. These properties are organized into the following groups:

Generic Properties: The following table provides detailed information on the general cluster related properties:

Table 5.1. Generic Properties

Property Name	Notes	Mandatory/Optional/Conditional
deployuser	Responsible for executing the HDP Installer. This user must be created on all the nodes in your cluster.	Required only for tarball based installations.
installdir	Full path to the installation directory for HDP (Example: /hdp).	
java64home	Location of JAVA_HOME for 64-bit JDK v 1.6 update 31 in your environment.	Mandatory
package	RPM packages for Red Hat compatible based systems. (Default: rpm)	Mandatory
security	If set to yes, installs secure Hadoop cluster. (Default: no) (Allowed: yes/no)	Conditional

Property Name	Notes	Mandatory/Optional/Conditional
sshkey	Either provide full path to the sshkey which allows you to perform passwordless SSH OR Set this field to empty when passwordless SSH is set-up.	Mandatory. Required when passwordless SSH is not setup. The SSH key must be passwordless SSH key.
smoke_test_user	User responsible for executing the smoke tests. (Default:hdptestuser)	Mandatory. Ensure that this user is created on all nodes in your cluster with primary group hadoop.

Hadoop Core Properties: The following table provides information on the properties required for core Hadoop components (HDFS and MapReduce):

Table 5.2. Hadoop Core Properties

Property Name	Notes	Mandatory/Optional/Conditional
enableappend	Enable this property even if `installhbase` is set to no. (Default: true) (Allowed: true/ false)	Conditional. Required only when `installhbase` is set to yes.
enablewebhdfs	Enable this property to true only if `security` is set to yes. (Default: false) (Allowed: true/false)	Conditional. Required only when `security` is set to yes.
taskscheduler	Scheduler to be used for job scheduling. (Default: org.apache.hadoop.mapred.CapacityTask-Scheduler) (Allowed: org.apache.hadoop.mapred.JobQueueTaskScheduler/ org.apache.hadoop.mapred.CapacityTaskScheduler)	Optional
enablelzo	Enable LZO compression. (Default: no) (Allowed: yes/no)	Optional. Required for compressing MapReduce jobs.
enableshortcircuit	Enable short circuit read. (Default: true) (Allowed: true/false)	Conditional Required only when `installhbase` is set to yes.

Service User Properties: The following table lists the properties for service users:

Note

For information on other service users, see (see: Hadoop Service Accounts).

Table 5.3. Service User Properties

Service User Name	Notes	Mandatory/Optional/Conditional
smoke_test_user	User responsible for executing the smoke tests. (Default: hdptestuser:hadoop)	Mandatory.

Data and Log Directory Configurations: The following properties determine the default locations for the HDFS data directories and log directories for all the components in the HDP stack:

Note

It is strongly recommended that you assign separate disks for individual data directories.

Table 5.4. Data and Log Directory Configurations

Property Name	Notes	Mandatory/Optional/Conditional
datanode_dir	Comma-separated list of full path to the Hadoop DataNode directories. (Example: /hdp/1/hadoop/hdfs/data,/hdp/2/hadoop/hdfs/data)	Mandatory
namenode_dir	Comma-separated list of full path to the Hadoop NameNode directories. (Example: /hdp/1/hadoop/hdfs/namenode,/hdp/2/hadoop/hdfs/namenode)	
mapred_dir	Comma-separated list of full path to the Hadoop MapReduce directories. (Example: /hdp/1/hadoop/mapred,/hdp/2/hadoop/mapred)	
log_dir	Full path to Hadoop log directory. (Example: /var/log/hadoop)	
pid_dir	Full path to Hadoop PID directory. (Example: /var/run/hadoop)	
hbase_log_dir	Full path to HBase log directory. (Example:/var/log/hbase)	Conditional. Required only if `installhbase` is set to true.
hbase_pid_dir	Full path to HBase PID directory. (Example:/var/run/hbase)	
hive_log_dir	Full path to Hive log directory. (Example:/var/log/hive)	Conditional. Required only if `installhive` is set to true.
zk_log_dir	Full path to ZooKeeper log directory. (Example:/var/log/zookeeper)	Conditional. Required only if `installzookeeper` is set to true.
zk_pid_dir	Full path to ZooKeeper PID directory. (Example:/var/run/zookeeper)	
zk_data_dir	Full path to ZooKeeper data directory. (Example: /hdp/1/hadoop/zookeeper)	
webhcat_log_dir	Full path to log directory for	Conditional. Required only if `installwebhcat` is set to true.

Property Name	Notes	Mandatory/Optional/Conditional
	WebHCat. (Example: /var/log/webhcat)	
webhcat_pid_dir	Full path to PID directory for WebHCat. (Example: /var/run/webhcat)	

HDP Stack Components Properties: All the properties listed below are **Optional**.

Table 5.5. HDP Stack Components Properties

Property Name	Notes
installpig	Enter yes to install Pig and no otherwise. (Default: yes) (Allowed: yes/no)
installhbase	Enter yes to install HBase and no otherwise. (Default: yes) (Allowed: yes/no)
installhive	Enter yes to install Hive and no otherwise. Ensure that you have installed MySQL server instance. (Default: yes) (Allowed: yes/no)
installwebhcat	Enter yes to install WebHCat and no otherwise. You must also set `installhbase` to `yes`. (Default: yes) (Allowed: yes/no)
mysqldbhost	Hostname and database name for the MySQL server. Required only if `installhive` is set to `yes`.
databasename	
mysqldbuser	MySQL credentials to connect to the MySQL database specified in `databasename` property. Ensure that the this user has been granted ALL privileges on the database. Required only if `installhive` is set to `yes`.
mysqldbpasswd	
tickTime	ZooKeeper uses this time unit to regulate heartbeats and timeouts. For example, if the tickTime is set to 2000, the minimum session timeout will be two ticks. (Default: 2000 milliseconds). Required only if `installzookeeper` is set to `yes`. Note, you must also set `installhbase` to `yes`.
initLimit	Amount of time (in ticks) to allow followers to connect and sync to a leader. Increase this value only if ZooKeeper manages large amount of data in your cluster. (Default: 10). Required only if `installzookeeper` is set to `yes`. Note, you must also set `installhbase` to `yes`.
syncLimit	Amount of time (in ticks) to allow followers to sync with ZooKeeper. All those followers that fall too far behind a leade will be dropped. (Default: 5). Required only if `installzookeeper` is set to `yes`. Note, you must also set `installhbase` to `yes`.
clientPort	Default port used for listening to the client connections. (Default: 2181). Required only if `installzookeeper` is set to `yes`. Note, you must also set `installhbase` to `yes`.

Secure Hadoop Deployment Properties: All the properties listed below are **Mandatory**.

Table 5.6. Secure Hadoop Deployment Properties

Property Name	Notes
security	Required to deploy secure Hadoop cluster. (Default: no) (Allowed: yes/no)
kinitpath	Full path to kinit executable. (Default: /usr/kerberos/bin/kinit) Ensure that you provide the correct value for this property.
keytabdir	Full path to service keytab files. These files are stored for all services (NameNode, DataNode, JobTracker, TaskTracker, Hive Metastore, HBase Master, and RegionServer) (Default:/etc/security/keytabs)
realm	Ensure that you replace the default value (`EXAMPLE.COM`) with correct realm. Use the `krb5.conf` file in your Kerberos Key Distribution Center to determine correct value for this property.
hdfs_user_keytab	Full path to the keytab file for `hdfsuser` service user. (Default: /homes/hdfs/hdfs.headless.keytab

Property Name	Notes
smoke_test_user_keytab	Full path to the keytab file for `smoke_test_user`. (Default: /homes/*`$smoke_test_user`*/*`$smoke_test_user`*.headless.keytab)

5.1.2. Category II - Configuration Properties for Optional Components In HDP

This category exposes the configuration properties for the optional components (Ganglia, Nagios, Oozie, Sqoop, and Mahout) in HDP. These properties are organized into the following groups:

- Monitoring components (Ganglia and Nagios) Properties [35]
- Properties for Apache Oozie [36]
- Properties for Sqoop [36]
- Properties for Flume [36]
- Properties for Mahout [36]

Monitoring components (Ganglia and Nagios) Properties: All the properties listed below are **Mandatory** when `enablemon` property is set to `yes`.

Important

These properties can be modified from the `master-install-location/gsInstaller/monInstaller.properties` file.

Table 5.7. Monitoring components (Ganglia and Nagios) Properties

Property Name	Notes
enablemon	Must be enabled to deploy HDP Monitopring components (Nagios and Ganglia). (Default: yes Allowed: yes/no)
installnagios	If set to yes, gsInstaller deploys the Nagios server for your cluster. (Default: yes) (Allowed: yes/no)
installsnmp	Deploy the SNMP module for your cluster. (Default: yes) (Allowed: yes/no)
installganglia	If set to yes, gsInstaller deploys the Ganglia server for your cluster. (Default: yes) (Allowed: yes/no)
snmpcommunity	Name of the SNMP community. (Default: hadoop)
snmpsource	gsInstaller uses this address range to configure source in `snmpd.conf` file. You must either use a host or provide the network addresses in CIDR notation. (For example: `192.168.0.0/24` means all the machines btween `192.168.0.0` and `192.168.0.255` are allowed to access data from SNMP daemons). Ensure that the Gateway and Nagios Server belong to the `snmpsource` address range.
nagioscontact	Valid email id to receive Nagios alerts. (Default: monitor\@monitor.com)
nagios_web_login	Credentials for accessing Nagios web interface. (Default: nagiosadmin/admin)
nagios_web_password	
gmetad_user	Valid user name for Ganglia gmetad_user. (Default: nobody)
gmond_user	Valid user name for the Ganglia gmond_user. (Default: nobody)
webserver_group	Ganglia webserver group. (Default: apache)

Property Name	Notes
package	Type of installation. (Default:rpm)

Properties for Apache Oozie:

Table 5.8. Properties for Apache Oozie

Property Name	Notes	Mandatory/Optional/Conditional
installoozie	gsInstaller deploys Apache Oozie if this property is enabled. (Default: yes) (Allowed: yes/no)	Optional
oozie_use_external_db	Database used for Oozie Metastore (default database is Derby). (Default: no) (Allowed: yes/no)	
oozie_dbname	If using external database, specify Oozie database name.	Conditional. required only if oozie_use_external_db is set to no.
oozie_dbuser	If using external database, specify Oozie database user credentials	
oozie_dbpasswd		
oozie_log_dir	Full path to log directory for Oozie. (Example: /var/log/oozie)	Required only is `installoozie` is set to yes.
oozie_pid_dir	Full path to PID directory for Oozie. (Example: /var/run/oozie)	

Properties for Sqoop:

Table 5.9. Properties for Sqoop

Property Name	Notes	Mandatory/Optional/Conditional
installsqoop	gsInstaller deploys Apache Sqoop if this property is enabled. (Default: yes) (Allowed: yes/no)	Optional

Properties for Flume:

Table 5.10. Properties for Flume

Property Name	Notes	Mandatory/Optional/Conditional
installflume	gsInstaller downloads Apache Flume RPM if this property is enabled. (Default: no) (Allowed: yes/no)	Optional. For information on deploying and configuring Flume, see Deploying and Configuring Flume.

Properties for Mahout:

Table 5.11. Properties for Mahout

Property Name	Notes	Mandatory/Optional/Conditional
installmahout	gsInstaller deploys Apache Mahout if this property is enabled. (Default: no) (Allowed: yes/no)	Optional

5.2. Configuration Cluster Properties

HDP provides the option to customize different parameters to tune your Hadoop cluster.

Note

You can modify these properties from the `master-install-location/gsInstaller/gsCluster.properties` file. Any changes to this file will override all the default configurations for your Hadoop cluster. It is therefore strongly recommended to exercise caution while changing this file.

Important

The value of NameNode new generation size (default size of Java new generation for NameNode (Java option `-XX:NewSize`)) should be 1/8 of maximum heap size (`-Xmx`) above. Please check, as the default setting may not be accurate. This value is specified in the `namenode_opt_newsize` property.

Table 5.12. Hadoop-HDFS Properties

Property Name	Notes
hadoop_heap_size	JVM heap size for the balancer. (Default: 1000 MB (1000m))
namenode_javaheap	NameNode's Java heap size. (Default: 4 GB (4G))
namenode_opt_newsize	Bound for new generation size (in bytes). (Default: 640 MB (640m)).
dt_heapsize	DataNodes' heap size. (Default: 1024 MB (1024m))
jtnode_opt_newsize	Lower bound for JobTracker newgen size. (Default: 200 MB (200m))
jtnode_opt_maxnewsize	Upper bound for JobTracker newgen size. (Default: 200 MB (200m))
jt_heapsize	JobTracker heap size. (Default: 24000 MB (24000m))
fs_inmemory_size	Memory allocated for in-memory file-system. Used to merge map-outputs for the reduces. (Default: 256)
datanode_du_reserved	Reserved space in bytes per volume. Ensure that you always leave this amount of space free for non HDFS use. (Default: 1,073,741,824)
dfs_datanode_failed_volume_tolerated	The number of volumes that are allowed to fail before a DataNode stops offering service. By default, any volume failure will cause a DataNode to shutdown. (Default: 0)

Table 5.13. Hadoop-MapReduce Properties

Property Name	Notes
mapred_cluster_map_mem_mb	Size (in terms of virtual memory) for a single map slot in the Map Reduce framework. (Default: -1)
mapred_cluster_red_mem_mb	Size (in terms of virtual memory) for a single reduce slot in the Map Reduce framework. (Default: -1)
mapred_cluster_max_map_mem_mb	Maximum number of map and reduce tasks that can be executed in parallel. This parameter is dpends on the `mapred_cluster_map_mem_mb` and `mapred_cluster_map_mem_mb` properties. Ensure that the number of map tasks is always greater than the number of reduce tasks. (Default: -1)
mapred_job_map_mem_mb	Size (in terms of virtual memory) of a single map task for the job. (Default: -1)
mapred_job_red_mem_mb	Size (in terms of virtual memory) of a single reduce task for the job. (Default: -1)
mapred_child_java_opts_sz	Java opts for the map and reduce tasks. (Default: -Xmx768m)
mapred_map_tasks_max	Nmber of map tasks per TaskTracker concurrently. Ensure that the maximum slots are greateer than the number of CPU cores because map tasks consume majority of free slots. For example: If you have 6 CPU cores and 8 slots, you must set the value for this parameter to 6. (Default: 4)
mapred_red_tasks_max	Number of reduce tasks per TaskTracker concurrently. (Default: 4)
io_sort_mb	Buffer memory used while sorting files. In order to minimize seek time, each merge stream is assigned 1 MB by default. (Default: 200 MB (200m))

Property Name	Notes
io_sort_spill_percent	Soft limit for either the buffer or the record collection buffers. Once reached, a background thread starts spilling the contents to disk. Note that this does not imply any chunking of data to the spill. A value less than 0.5 is not recommended. (Default: 0.9)
mapreduce_userlog_retainhours	Maximum retention period for user-logs, post job completion. (Default: 24 hrs. (24))
maxtasks_per_job	Maximum number of tasks allowed for single job (map and reduce). (Default: -1)

Table 5.14. Hadoop-ZooKeeper Properties

Property Name	Notes
tickTime	ZooKeeper uses this time unit to regulate heartbeats and timeouts. For example, if the tickTime is set to 2000, the minimum session timeout will be two ticks. (Default: 2000 milliseconds). Required only if `installzookeeper` is set to `yes`. Note, you must also set `installhbase` to `yes`.
initLimit	Amount of time (in ticks) to allow followers to connect and sync to a leader. Increase this value only if ZooKeeper manages large amount of data in your cluster. (Default: 10). Required only if `installzookeeper` is set to `yes`. Note, you must also set `installhbase` to `yes`.
syncLimit	Amount of time (in ticks) to allow followers to sync with ZooKeeper. All those followers that fall too far behind a leade will be dropped. (Default: 5). Required only if `installzookeeper` is set to `yes`. Note, you must also set `installhbase` to `yes`.
clientPort	Default port used for listening to the client connections. (Default: 2181). Required only if `installzookeeper` is set to `yes`. Note, you must also set `installhbase` to `yes`.

5.3. Creating Kerberos Principals And Keytab Files

A Kerberos principal is a unique identity to which Kerberos can assign tickets. (See: Kerberos V5 UNIX User's Guide (MIT)) All machines hosting a Kerberos-using service need a keytab file, called /etc/krb5.keytab, to authenticate to the KDC. The keytab file is an encrypted, local, on-disk copy of the host's key. On a UNIX system, you can view the contents of a keytab file using **klist -k** command. (See: Kerberos V5 UNIX User's Guide (MIT))

Kerberos defines two different types of accounts (or Principals) as listed below:

- User Principal Name
- Service Principal Name (Example: `HTTP/$fully.qualified.domain.name@EXAMPLE.COM`)

Hortonworks Data Platform (HDP) requires each of the Hadoop service (NameNode, Secondary NameNode, JobTracker, HBase Master, HCatalog Server, DataNodes, TaskTracker, and HBase RegionServers) to have its own keytab file.

The keytab files for all the services must have the service user principal and these principals must follow the mandatory naming conventions. (For example: For MapReduce, the service user principal will be `mapred`.)

Note

The keytab files must have the following three principals for NameNode and Secondary NameNode services only:

- HDFS principal (nn)
- HTTP principal

The following sections provide more information on creating Kerberos principals and keytab files:

- Creating Kerberos Principals
- Creating Keytab files

5.3.1. Creating Kerberos Principals

Step 1: As root user, start the kadmin tool on the KDC server.

```
/usr/krb5/sbin/kadmin.local
kadmin.local:
```

Step 2: Create the principal for all services in your Hadoop cluster.

```
kadmin: addprinc -randkey $principal_name/$fully.qualified.domain.name@$YOUR-REALM.COM
```

where the *$principal_name* must use following mandatory naming conventions:

Table 5.15. Secure deployment - Mandatory naming conventions for principals

Service Name	Principal Name (mandatory naming convention)
NameNode	nn and HTTP
Secondary NameNode	nn, HTTP
JobTracker	jt
TaskTracker	tt
DataNode	dn
HBase Master	hm
HBase RegionServer	rs
Hive Metastore	hive
Oozie	oozie
Oozie	HTTP
WebHCat	HTTP

For example, to create NameNode principals, from the shell try:

```
kadmin: addprinc -randkey nn/NAMENODE@EXAMPLE.COM
kadmin: addprinc -randkey HTTP/NAMENODE@EXAMPLE.COM
```

Step 3: Follow the instructions for Step - 2 above to create keytab files according to the following mandatory naming conventions:

Table 5.16. Secure deployment - Mandatory naming conventions for principals

User Name	Principal Name (mandatory naming convention)
HDFS User	hdfs
Smoke Test User	Value specified for `smoke_test_user` property in `master-install-location/gsInstaller/gsInstaller.properties` file.

5.3.2. Creating Keytab files

Step 1: As the root user, start the kadmin tool on the KDC server:

```
/usr/krb5/sbin/kadmin.local
kadmin.local:
```

Step 2: Create the keytab files for all services in your Hadoop cluster.

1. Use `kadmin` utility to execute the following:

```
kadmin: xst -norandkey -k $keytab_file_name $principal_name/fully.qualified.
domain.name
```

The *`$keytab_file_name`* must use the following mandatory naming conventions:

Table 5.17. Secure deployment - Mandatory naming conventions for keytab files

Service Name	Keytab File Name	Principal Name	Permissions	Owner
NameNode	`nn.service.keytab`	`nn`	`700`	`hdfs:hadoop`
NameNode	`spnego.service.keytab`	`HTTP`	`700`	`hdfs:hadoop`
Secondary NameNode	`nn.service.keytab`	`nn`	`700`	`hdfs:hadoop`
Secondary NameNode	`spnego.service.keytab`	`HTTP`	`700`	`hdfs:hadoop`
JobTracker	`jt.service.keytab`	`jt`	`700`	`mapred:hadoop`
TaskTracker	`tt.service.keytab`	`tt`	`700`	`mapred:hadoop`
DataNode	`dn.service.keytab`	`dn`	`700`	`hdfs:hadoop`
HBase Master	`hm.service.keytab`	`hm`	`700`	`hbase:hadoop`
HBase RegionServer	`rs.service.keytab`	`rs`	`700`	`hbase:hadoop`
Hive Metastore	`hive.service.keytab`	`hive`	`700`	`hive:hadoop`
Oozie	`oozie.service.keytab`	`oozie`	`700`	`oozie:hadoop`
Oozie	`spnego.service.keytab`	`HTTP`	`700`	`oozie:hadoop`
WebHCat	`spnego.service.keytab`	`HTTP`	`700`	`webhcat:hadoop`

For example, to create NameNode principals' (nn, host, and HTTP) keytab files, execute the following commands:

```
kadmin: xst -k nn.service.keytab nn/NAMENODE
```

```
kadmin: xst -k spnego.service.keytab HTTP/NAMENODE
```

2. On each of the node in your cluster, change directory to the *$keytab* directory.

Note

The location is specified by the `keytabdir`property in the `gsInstaller.properties` file

3. Copy the appropriate keytab file on each of node in your cluster.

Step 3: Follow the steps listed in Step - 2 above, to create keytab files according to the following mandatory naming conventions:

Table 5.18. Secure deployment - Mandatory naming conventions for HDFS and Smoke test users' keytab files

User Name	Principal Name (mandatory naming convention)
HDFS User	Value specified for `hdfsuser.headless.keytab` property in `master-install-location/gsInstaller/gsInstaller.properties` file.
Smoke Test User	Value specified for `smoke_test_user.headless.keytab` property in `master-install-location/gsInstaller/gsInstaller.properties` file.

Step 4: Use the klist utlity on each of your service to verify that the correct keytab files and principals are associated with the correct service. For example, to verify the keytabs for the NameNode, execute the following command:

```
klist –k -t /etc/security/nn.service.keytab
```

5.4. Uninstalling gsInstaller

Step 1: Start the uninstallation.

1. Execute the following command from the master-install machine for gsInstaller.

```
sh $master-install-location/gsInstaller/gsUninstaller.sh
```

Step 2: Remove the following directories:

- Artifact directory: `/tmp/HDP-artifacts-`*`$Process_ID`*
- Log directory: `/tmp/gsinstaller-`*`$Process_ID`*
- On all the host machines in your cluster, remove the following directory: `/tmp/HDP`