cloudera
Ask Bigger Questions

Apache Accumulo Installation Guide

Important Notice

Cloudera, Inc.
1001 Page Mill Road
Palo Alto, CA 94304-1008
info@cloudera.com
US: 1-888-789-1488
Intl: 1-650-362-0488
www.cloudera.com

Release Information

Version: 1.4.3-cdh4.3.0

Date: January 20, 2014

Table of Contents

Introducing Apache Accumulo

Apache Accumulo™ is an ideal solution for government agencies looking for a secure, distributed NoSQL data store to serve their most performance-intensive Big Data applications. Accumulo is an open source project integrated with Hadoop and provides the ability to store data in massive tables (billions of rows / millions of columns) for fast, random access. Accumulo was created and contributed to the Apache Software Foundation by the National Security Agency (NSA). It has quickly gained adoption as a Hadoop-based key/value store for applications that have unique and stringent information security requirements.

Known Issues

——Feedback loop on trace table

Under certain circumstances, a tablet server with a small in-memory map can get into a feedback loop where flushes of a tablet for the trace table cause additional tracer entries about the flush.

Bug: ACCUMULO-2014

Severity: Low

Workaround: Increase the amount of memory available for the in-memory map.

——Spurious warnings when using Cloudera Manager

When executing certain utilities, including the Accumulo shell, against a Cloudera Manager installed Accumulo cluster, you will see warnings about missing configuration files.

Severity: Low

Workaround: Ignore warnings.

Resolution: To be fixed in an upcoming release.

——Cloudera Manager does not set the correct permissions for the /accumulo directory

After installing Accumulo using Cloudera Manager, clients will see Permission denied errors on the `/accumulo` directory.

Severity: Low

Workaround: Set the appropriate permissions on the directory as described in the following "Installing Apache Accumulo by using Cloudera Manager" section.

Resolution: To be fixed in an upcoming release.

Prerequisites

Accumulo depends on Hadoop (HDFS & MapReduce) and ZooKeeper libraries and configuration information. TabletServers and Loggers should be collocated with DataNodes. These DataNodes do not require TaskTrackers.

The Cloudera release of Apache Accumulo is tested for use with CDH 4.3.0 or later, either in parcel (Cloudera Manager) or package (RPM/DEB) form. Examples given are for a package-based install on the CDH 4.3.0 QuickStart VM (available for VMware, VirtualBox and KVM) but multi-node examples have been provided where needed.

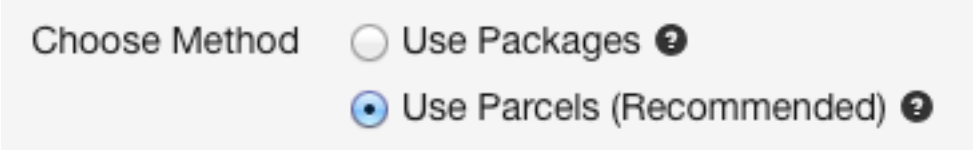
Installing Apache Accumulo by using Cloudera Manager

This section describes how to install Accumulo by using Cloudera Manager. If you prefer to install Accumulo from a tarball, skip this section and see the "Installing Apache Accumulo from the Distribution Tarball" section on page 8.

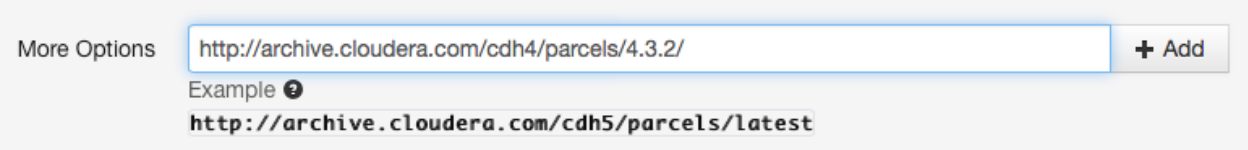
Installing and Configuring Cloudera Manager 5 and CDH4.3

Follow the documentation for installing and configuring Cloudera Manager 5 with CDH4.3. During the installation of CDH, you must choose the CDH4.3 parcel when going through the wizard.

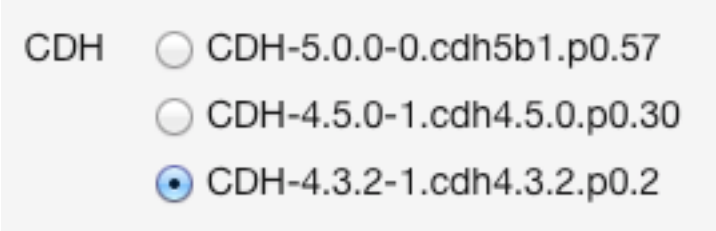
1. Be sure the **Use Parcels** option is checked.



2. Under **More Options**, add the CDH4.3 parcel repository URL (http://archive.cloudera.com/cdh4/parcels/4.3.2/) and click Add.



3. Select version **CDH-4.3.2-1.cdh4.3.2.p0.2** for the **CDH** parcel.

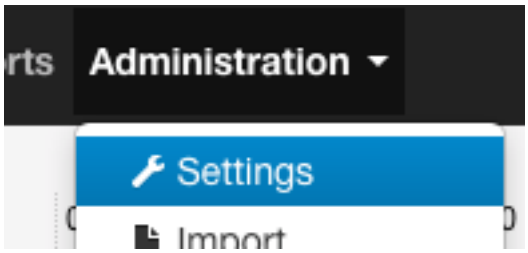


4. Click **Continue** and follow the rest of the installation steps as described in the documentation. Accumulo requires that you setup the HDFS, Zookeeper, and MapReduce services (only MRv1 is currently supported). Other services are optional.

Adding the Accumulo 1.4.3 Parcel Repository to Cloudera Manager

In order to install Accumulo, you must configure Cloudera Manager with the location of the Accumulo Parcel Repository.

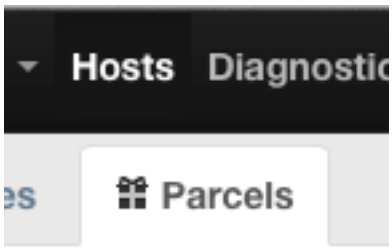
1. From the **Administration** tab, select **Settings.**



2. Select **Parcels** on the left.
3. Under **Remote Parcel Repository URLs**, click the **+** and add the Accumulo parcel URL: http://archive.cloudera.com/accumulo/parcels/1.4.3/

Installing the Accumulo Parcel

1. From the **Hosts** tab, select **Parcels.**



2. Under the **ACCUMULO 1.4.3-cdh4.3.0** parcel, click **Download.**

3. Under the **ACCUMULO 1.4.3-cdh4.3.0** parcel, click **Distribute.**
4. Under the **ACCUMULO 1.4.3-cdh4.3.0** parcel, click **Activate.**
5. You will be prompted to restart the cluster. Since the Accumulo parcel was not previously in use, you can safely skip this step and click **Close.**

Downloading and Installing the Accumulo Cloudera Service Descriptor (CSD) Plugin

Cloudera Manager 5 added a Cloudera Service Descriptor (CSD) API to extend Cloudera Manager to manage third-party services. Cloudera Manager's Accumulo support is implemented through this API so you need to install the Accumulo CSD plugin before you can add the service.

1. Create a directory named `/opt/cloudera/csd` on the host running the Cloudera Manager Server.

```
$ sudo mkdir /opt/cloudera/csd
```

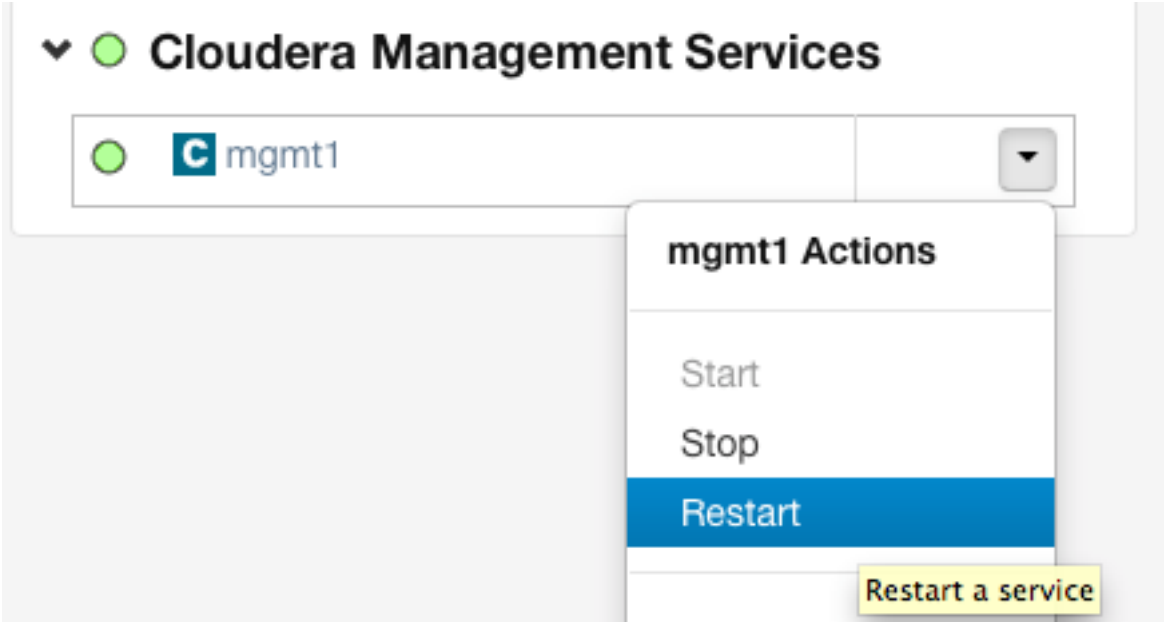
2. Download the CSD jar file, copy it into the `/opt/cloudera/csd` directory, and change ownership of the directory and its contents.

```
$ sudo cp accumulo-1.4.3.jar /opt/cloudera/csd
$ sudo chown –R cloudera-scm:cloudera-scm /opt/cloudera/csd
```

3. Restart the Cloudera Manager Server.

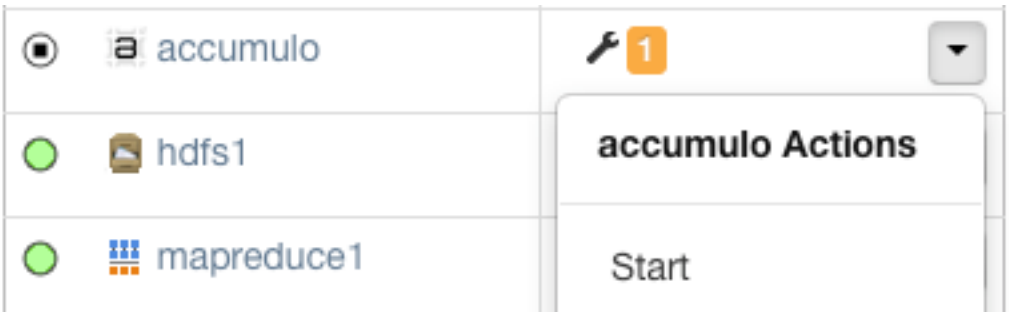
```
$ sudo service cloudera-scm-server restart
```

4. Restart the Cloudera Management Services.



Adding the Accumulo Service

1. Click the **Clusters** tab, and then choose **All Services.**
2. From the **Actions** menu, select **Add a Service.**
3. Select **Accumulo** and click **Continue.**
4. Select the dependent services and click **Continue.**
5. Assign the Accumulo roles to the hosts in your cluster. Cloudera recommends assigning a **Logger** and **Tablet Server** role on each host that is assigned the **DataNode** role. The **Monitor**, **Garbage Collector**, **Tracer**, and **Master** roles should all be assigned to **non-DataNodes**. The **Gateway** role should be assigned to any hosts where you want to use Accumulo that don't already have any other Accumulo roles assigned.
6. Click **Continue.**
7. Click **Accept.**
8. Click **Continue.**
9. Click **Continue.**
10. From the **accumulo Actions** menu, select **Start.**



11. Click on the **accumulo** service.
12. From the **Actions** menu, select **Deploy Client Configuration.**

Note: If you are following the wizard install, steps 10-12 above will be performed by Cloudera Manager.

There is a known issue with the current release of the Accumulo service descriptor where the permissions for the `/accumulo` directory are not set correctly. To fix this, execute the following from a cluster node:

```
$ sudo su - hdfs
$ hadoop fs -chmod 751 /accumulo
$ hadoop fs -chmod 750 /accumulo/*
$ hadoop fs -chmod 755 /accumulo/instance_id
```

Testing the Accumulo Shell

You should now be able to run the Accumulo shell on any of the hosts that are assigned the **Gateway** role within your cluster. By default, the user **root** is created and given the password **secret**. It is strongly recommended that you change this password.

1. Launch the Accumulo shell for the default root user.

```
$ accumulo shell -u root
grep: /etc/accumulo/conf/masters: No such file or directory
Enter current password for 'root'@'accumulo': ******

Shell - Apache Accumulo Interactive Shell
-
- version: 1.4.3-cdh4.3.0
- instance name: accumulo
- instance id: 9863d1f1-c323-4671-9e1f-69857f0d635f
-
- type 'help' for a list of available commands
-
root@accumulo>
```

2. Use the passwd command to set a new password for the root user.

```
root@accumulo> passwd
Enter current password for 'root': ******
Enter new password for 'root': *********
Please confirm new password for 'root': *********
root@accumulo>
```

3. Relaunch the shell with this new password.

```
root@accumulo> exit
$ accumulo shell -u root
grep: /etc/accumulo/conf/masters: No such file or directory
Enter current password for 'root'@'accumulo': *********

Shell - Apache Accumulo Interactive Shell
-
- version: 1.4.3-cdh4.3.0
- instance name: accumulo
- instance id: 9863d1f1-c323-4671-9e1f-69857f0d635f
```

```
-
- type 'help' for a list of available commands
-
root@accumulo>
```

4. Verify that you can list tables.

```
root@accumulo> tables
!METADATA
trace
root@accumulo>
```

For more information on using the Accumulo shell, see the Accumulo user manual.

Changing the Trace User Settings

If you followed the directions in the previous section, then you must also update the credentials used by the Tracer.

1. Click the **Accumulo** service.
2. Select the **Configuration** tab and click **Service-Wide** on the left.
3. Scroll down to the **Trace User** and **Trace Password** settings.
4. If you want to keep using the **root** user for the Tracer, then set the **Trace Password** to the same value you set for the root user's password. Alternatively, you could add a new user in the Accumulo shell to use for the Tracer and configure the **Trace User** and **Trace Password** settings here. See the Accumulo user manual for details on creating users and setting permissions with the Accumulo shell.
5. Restart the Tracer role after saving the new configuration.

Installing Apache Accumulo from the Distribution Tarball

This section describes how to install Accumulo from a tarball instead of using Cloudera Manager.

Setup and Configuration for Accumulo

Creating a Accumulo User Account

Before installing Accumulo, follow the instructions in this section to create a user `accumulo` and assign ownership of the new `/var/lib/accumulo` directory to it.

To create a Accumulo user account:

1. Run the following commands on every host in your Accumulo cluster.

```
$ sudo mkdir /var/lib/accumulo
$ sudo useradd -r -s /bin/bash -d /var/lib/accumulo accumulo
$ sudo cp /etc/skel/.bash* ~accumulo/
$ sudo chown –R accumulo:accumulo /var/lib/accumulo
```

2. Verify that the `accumulo` user on the Accumulo Master is configured for passwordless SSH to the `accumulo` account on itself as well as all other servers running Accumulo processes:

```
$ sudo su - accumulo
$ ssh-keygen –t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/var/lib/accumulo/.ssh/id_rsa):
Created directory '/var/lib/accumulo/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /var/lib/accumulo/.ssh/id_rsa.
Your public key has been saved in /var/lib/accumulo/.ssh/id_rsa.pub.

$ cat ~/.ssh/id_rsa.pub > ~/.ssh/authorized_keys
$ chmod 600 ~/.ssh/authorized_keys
```

3. On every host, verify that the `/var/lib/accumulo/.ssh/authorized_keys` file contains the key from `/var/lib/accumulo/.ssh/id_rsa.pub` as well. You should then be able to SSH from the Master to other members of the Accumulo cluster without being prompted for a password.

Installing Accumulo

To install Accumulo:

1. On every host, extract the Accumulo tarball to `/usr/lib/accumulo`:

```
$ sudo mkdir /usr/lib/accumulo
$ sudo tar –xf accumulo-1.4.3-cdh4.3.0-dist.tar.gz –C
```

```
/usr/lib/accumulo --strip=1
$ sudo chown –R accumulo:accumulo /usr/lib/accumulo/
```

2. Move the `/usr/lib/accumulo/conf` directory to `/etc/accumulo/cdh4_default` and rename the sample environment configuration script based on whether you're running a Cloudera Manager-installed cluster (`accumulo-env-cdh4-parcels.sh`) or a package-installed (RPM/DEB) cluster (`accumulo-env-cdh4-packages.sh`):

```
For either CDH installation type:
$ sudo mkdir /etc/accumulo
$ sudo mv /usr/lib/accumulo/conf /etc/accumulo/cdh4_default
$ sudo su - accumulo
$ cp /etc/accumulo/cdh4_default/examples/CDH4_parcels/* \
    /etc/accumulo/cdh4_default/
$ exit
$ sudo chown -R accumulo:accumulo /etc/accumulo/
$ sudo chmod -R og+rX /etc/accumulo/cdh4_default/examples
$ sudo alternatives --install /etc/accumulo/conf accumulo-conf \
    /etc/accumulo/cdh4_default 20

Additional Step For Package (RPM/DEB) based CDH installations:
$ sudo sed -i "s/\/opt\/cloudera\/parcels\/CDH/\/usr/g"
/etc/accumulo/conf/accumulo-env.sh
```

3. Set your `ACCUMULO_CONF_DIR`, `ACCUMULO_HOME` and `JAVA_HOME` variables, as well as define the values in `/var/lib/accumulo/.bashrc` to make them persistent:

```
$ sudo su - accumulo
$ export ACCUMULO_CONF_DIR=/etc/accumulo/conf
$ export ACCUMULO_HOME=/usr/lib/accumulo
$ export JAVA_HOME=/usr/java/jdk1.6.0_31
$ echo "export ACCUMULO_CONF_DIR=/etc/accumulo/conf" >> ~/.bashrc
$ echo "export ACCUMULO_HOME=/usr/lib/accumulo" >> ~/.bashrc
$ echo "export JAVA_HOME=/usr/java/jdk1.6.0_31" >> ~/.bashrc
$ exit
```

4. Repeat step 3 for any other users who will need to use the Accumulo shell.

Configuring Accumulo for your Environment

After installation, follow the steps in this section to configure Accumulo for your environment.

1. On every host, configure the following properties in `/etc/accumulo/conf/accumulo-site.xml` with the proper values for your environment:

```
<property>
```

```
  <name>instance.zookeeper.host</name>
  <value>localhost:2181</value>
  <description>comma separated list of zookeeper
servers</description>
</property>

<property>
  <name>logger.dir.walog</name>
  <value>/dfs/walogs</value>
  <description>The directory used to store write-ahead logs on the
local filesystem. It is possible to specify a comma-separated list of
directories.</description>
</property>

<property>
  <name>instance.secret</name>
  <value>DEFAULT</value>
  <description>A secret unique to a given instance that all servers
must know in order to communicate with one another.
               Change it before initialization. To change it later
use ./bin/accumulo org.apache.accumulo.server.util.ChangeSecret
[oldpasswd] [newpasswd],
               and then update this file.
  </description>
</property>

<property>
  <name>tserver.memory.maps.max</name>
  <value>256M</value>
</property>

<property>
  <name>tserver.cache.data.size</name>
  <value>15M</value>
</property>

<property>
  <name>tserver.cache.index.size</name>
  <value>40M</value>
</property>

<property>
  <name>trace.password</name>
  <!--
    change this to the root user's password, and/or change the user
below
    -->
  <value>secret</value>
</property>
```

```
<property>
  <name>trace.user</name>
  <value>root</value>
</property>
```

2. Review the configured values. For example, verify that you changed the value for `instance.secret` and make sure that the `logger.dir.walog` value points to where you will want your Accumulo write-ahead logs saved on your TabletServers.

Important:

If running a **multi-host cluster managed by Cloudera Manager,** you will need to add the following lines to `accumulo-site.xml`:

```
<property>
  <name>tserver.port.client</name>
  <value>10011</value>
</property>
```

This is to avoid a port conflict with the Cloudera Manager Service Monitor which also uses port 9997. In addition, be sure to open up TCP port 10011 in the firewall on all hosts running Table Servers.

Similarly, if you are running **Hue** (which is included but not started by default on Cloudera's QuickStart VM) on the same server or virtual machine as Accumulo, you will need to add the following property to `accumulo-site.xml`:

```
<property>
  <name>master.port.client</name>
  <value>10010</value>
</property>
```

You will also need to open up TCP port 10010 in your firewall if you are running Hue and the Accumulo Master on the same node in a multi-node setup (this, however, is not recommended).

3. Review any settings you may want to change in `accumulo-env.sh`, such as the location of the logs. For example, to make sure Accumulo logs to `/var/log/accumulo`, use the following commands:

```
$ sudo mkdir /var/log/accumulo
$ sudo chown accumulo:accumulo /var/log/accumulo
$ sudo ln –s /var/log/accumulo /usr/lib/accumulo/logs
```

4. Open `accumulo-env.sh` and edit the value of the `ACCUMULO_LOG_DIR` variable to be **`/var/log/accumulo.`** Make sure that this folder is created on every host of this cluster.

5. On each TabletServer, create the write-ahead log directory. For our example this would be `/dfs/walogs`.

```
$ sudo mkdir –p /dfs/walogs
$ sudo chown accumulo:accumulo /dfs/walogs
$ sudo ln –s /dfs/walogs /usr/lib/accumulo/walogs
```

6. Build the native libraries. In Accumulo 1.4.3, both 32-bit and 64-bit libraries are built which is why both the i686 and x86_64 versions of the development libraries are required.

```
$ sudo yum install gcc-c++ glibc-devel.i686 libstdc++-devel.i686 \
glibc-devel.x86_64 libstdc++-devel.x86_64
$ sudo su – accumulo
$ cd /usr/lib/accumulo
$ make -C ./src/server/src/main/c++
```

Apache Accumulo is now installed and configured. Continue to the next section to initialize Accumulo and start the Accumulo services.

Important:

On a **multi-host cluster** replace `localhost` with the fully qualified domain name (FQDN) or IP address of the Accumulo Master in the `masters`, `monitor`, gc and `tracers` files in `$ACCUMULO_CONF_DIR`, and add the FQDN or IP address of the TabletServers (one per line) to the `/usr/lib/accumulo/conf/slaves` file.

Important:

On a **multi-host cluster**, the contents of the `$ACCUMULO_CONF_DIR` directory must always be synchronized across all Accumulo servers within a cluster. This may be done using configuration management, version control or via a utility such as `rsync`. Servers with out-of-sync configurations will not be allowed to join the cluster.

Initializing Accumulo

To initialize Accumulo:

1. Since the `hdfs` user is not running Accumulo, create the `/accumulo` and `/user/accumulo` directories in `hdfs` and change their ownership to the `accumulo` user.

```
$ sudo su - hdfs
$ hadoop fs -mkdir /accumulo /user/accumulo
$ hadoop fs -chown accumulo:supergroup /accumulo /user/accumulo
$ hadoop fs –chmod 750 /accumulo /user/accumulo
```

2. On the Accumulo Master, enter the following commands to initialize Accumulo and follow the prompts to name your instance (for this example, `cloudera`) and set a root password:

```
$ sudo su - accumulo
$ cd /usr/lib/accumulo/
$ bin/accumulo init
[util.Initialize] INFO : Hadoop Filesystem is
hdfs://localhost.localdomain:8020
[util.Initialize] INFO : Accumulo data dir is /accumulo
[util.Initialize] INFO : Zookeeper server is localhost:2181
[util.Initialize] INFO : Checking if Zookeeper is available. If this
hangs, then you need to make sure zookeeper is running

Instance name : cloudera
Enter initial password for root: ****
Confirm initial password for root: ****

[conf.Configuration] WARN : dfs.replication.min is deprecated. Instead,
use dfs.namenode.replication.min
[conf.Configuration] WARN : dfs.block.size is deprecated. Instead, use
dfs.blocksize
[security.ZKAuthenticator] INFO : Initialized root user with username:
root at the request of user !SYSTEM
```

Warnings:

You will be warned here if you didn't change your instance secret in `conf/accumulo-site.xml`.

If the "`Hadoop Filesystem is`" line contains "`fs://`" instead of "`hdfs://`", your HDFS is not properly configured.

3. Make sure that users in the `accumulo` group can access Accumulo utilities, such as the shell, while leaving access to Accumulo internals restricted by altering the permissions on the Accumulo section of HDFS:

```
$ sudo su - accumulo
$ hadoop fs -chmod 751 /accumulo
```

```
$ hadoop fs -chmod 750 /accumulo/*
$ hadoop fs -chmod 755 /accumulo/instance_id
```

Starting Accumulo

To start Accumulo:

1. Run the following commands:

```
$ sudo su - accumulo
$ cd /usr/lib/accumulo/
$ bin/start-all.sh

Starting tablet servers and loggers .... done
Starting tablet server on localhost
Starting logger on localhost
[server.Accumulo] INFO : Attempting to talk to zookeeper
[server.Accumulo] INFO : Zookeeper connected and initialized,
attemping to talk to HDFS
[server.Accumulo] INFO : Connected to HDFS
Starting master on localhost
Starting garbage collector on localhost
Starting monitor on localhost
Starting tracer on localhost
```

2. You should now be able to connect to Accumulo on http://localhost:50095/. Check the number of processes via the following command:

```
$ sudo -u accumulo jps -ml | grep org.apache.accumulo.start.Main \
    | wc -l
```

Note:

If you are attempting to connect to the Accumulo Master from another host, make sure your firewall allows inbound access on TCP port 50095. Also, in a **multi-node cluster**, make sure every Accumulo node can access the following TCP ports on every other Accumulo node: 4560, 11224, 12234, 50091, and 50095.

Stopping Accumulo

If you need to stop your Accumulo cluster, run the following commands on the Accumulo Master:

```
$ sudo su - accumulo
$ cd /usr/lib/accumulo/
$ bin/stop-all.sh

Accumulo shut down cleanly
Utilities and unresponsive servers will be shut down in 5 seconds
stopping gc on localhost
stopping monitor on localhost
stopping tracer on localhost
stopping gc on localhost
stopping monitor on localhost
stopping tracer on localhost
stopping unresponsive tablet servers (if any) ...
stopping logger on localhost
stopping unresponsive tablet servers hard (if any) ...
Cleaning tablet server and logger entries from zookeeper
Cleaning all server entries in zookeeper
```

Note:

If the `bin/stop-all.sh` command is unresponsive, you may need to send it an interrupt via CTRL-C one or more times. This will cause it to forcibly stop unresponsive services.

Troubleshooting

Reviewing Accumulo Logs

When troubleshooting Accumulo, first review the logs located at `/var/log/accumulo`. There are separate logs for each server process within Accumulo. For example, on our single node configuration there are four log file extensions/types: ".log", "out" (stdout), "err" (stderr) and ".debug" for each of the Accumulo roles.

Service	**Base Log Name**
Garbage Collector	`gc.FQDN.*`
Logger	`logger.FQDN.*`
Master	`master.FQDN.*`

Service	Base Log Name
Monitor	`monitor.FQDN.*`
Tracer	`tracer.FQDN.*`
TabletServer	`tserver.FQDN.*`

Review the logs on the Master as well as the TabletServers if you encounter issues.

Example

If the native libraries Accumulo relies on aren't in `/usr/lib/accumulo/lib/native`, the TabletServers will throw an error (yet continue to run). You can see this by running the following command after starting a TabletServer where `/usr/lib/accumulo/lib/native` doesn't exist:

```
$ grep ERROR /var/log/accumulo/tserver_localhost.localdomain.log

2013-08-28 10:28:15,813 [tabletserver.NativeMap] ERROR: Failed to
load native map library
/usr/lib/accumulo/lib/native/map/libNativeMap-Linux-amd64-64.so
```

To fix this error, run the following commands to rebuild the native libraries Accumulo needs, and then restart your TabletServer:

```
$ sudo su – accumulo
$ cd /usr/lib/accumulo
$ make -C ./src/server/src/main/c++
$ bin/tdown.sh && bin/tup.sh
```

If you look at the `/var/log/accumulo/tserver_localhost.localdomain.log` file, you will see there are no additional entries related to the missing library.

Note:

If you haven't previously compiled native libraries, you may need to install additional packages via the `yum` command in order to build them:

```
sudo yum install gcc-c++ glibc-devel.i686 libstdc++-devel.i686
```