cloudera®

Apache Accumulo Installation Guide

for using the Cloudera packaging of Accumulo for CDH

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: 1.7.2-cdh5.5.0

Date: Oct 20, 2016

Table of Contents

About this Guide

This guide describes how to install the Cloudera packaging of Apache Accumulo for use with CDH.

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 data storage 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

—Accumulo service name discrepancy in different versions of Cloudera Manager

In Cloudera Manager 5.5.x through 5.8.x, the Accumulo service is labeled **Accumulo 1.6**. In Cloudera Manager releases higher than 5.8.x, the Accumulo service is labeled **Accumulo**.

—Performance regression in large cells and wide tables

Batch write performance for Accumulo 1.7.2-cdh5.5.0 shows a regression of up to approximately 30 percent, depending on table shape, when compared to Accumulo 1.6.0-cdh5.1.4. The performance decrease is more severe for exceptionally large cells (100k and larger) or exceptionally wide rows (10k columns). Carefully consider the performance impact for your environment when deciding to upgrade to Accumulo 1.7.2-cdh5.5.0.

—Unsupported features

The following upstream Apache Accumulo 1.7.2 features are not supported in the Cloudera packaging of Accumulo:

- User-initiated compaction strategies.
- GroupBalancer.
- User-specified durability.
- Multi-volume deployments (including custom VolumeChoosers, including the per-table VolumeChooser).
- Data-center replication (experimental and unsupported).

For more information about problems and workarounds specific to running an Accumulo service, see the known issues document for your release of Cloudera Manager.

Prerequisites

- Accumulo depends on HDFS and ZooKeeper libraries and configuration information. For more information about configuring HDFS, see Managing HDFS. For more information about the ZooKeeper service, see Managing Zookeeper.
- Tablet Servers should be collocated with DataNodes. Optionally, you can use Accumulo with MapReduce and Sqoop 1. Cloudera recommends that MapReduce users use YARN in CDH 5.
- The current release of the Cloudera packaging of Apache Accumulo is tested for use with CDH 5.5.0 and higher. Cloudera Manager has been tested for managing this release with both parcels and package (RPM) installations with CDH 5.5.0 and higher.
- For full cluster installations, Cloudera strongly recommends following guidelines in the CDH 5 Installation Guide.

Installing Accumulo 1.7 Using Cloudera Manager

Note: Managing a cluster installed with packages

The instructions in this section use parcels to install Accumulo and require that you install CDH using parcels. You can also use Cloudera Manager to manage the Accumulo 1.7 service when installing packages. To do so, see Install Apache Accumulo from Distribution Packages.

This section describes how to install the Cloudera packaging of Accumulo by using Cloudera Manager 5.5.0 or higher.

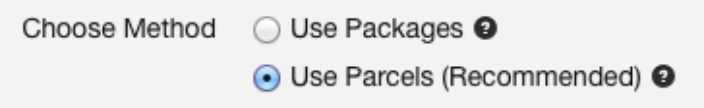
Note: Upgrading CDH and Accumulo

- You must be running CDH 5 on your cluster to run Accumulo 1.7. For information about upgrading from CDH 4 to CDH 5, see the upgrade documentation.
- If you are running Accumulo 1.4, you must first upgrade to Accumulo 1.6 before upgrading to Accumulo 1.7. For information about upgrading to Accumulo 1.6, see the *Apache Accumulo 1.6.0 Installation Guide*.

Step 1: Install and Configure Cloudera Manager and CDH

Follow the documentation to install and configure Cloudera Manager 5 with CDH 5. During the installation, select compatible CDH and Accumulo parcels while following these instructions.

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



2. Select version **CDH-5.5.0-1.cdh5.5.0.p0.8** or higher for the CDH parcel.
3. Click **Continue** and follow the rest of the installation steps as described in the documentation. Accumulo requires that you set up HDFS and Zookeeper. Other services are optional.

Step 2: Install the Accumulo Parcel

1. From the **Hosts** tab, select **Parcels.**
2. Under the parcel entry for **ACCUMULO-1.7.2-5.5.0.ACCUMULO5.5.0.p0.7** or later, click **Download.**
3. Under the cluster you want to install on (for example, Cluster 1), find the Accumulo parcel and click **Distribute.**
4. Under the cluster you want to install on (for example, Cluster 1), find the Accumulo parcel and click **Activate.**

You will be prompted to restart the cluster. Because the Accumulo parcel was not previously in use, you can safely skip this step and click **Close.**

Step 3: Configure HDFS

Cloudera strongly recommends that you establish an HDFS nameservice on the cluster that will run Accumulo. Because of the way Accumulo manages files within HDFS, establishing a nameservice greatly reduces administrative tasks in the future if a NameNode needs to be replaced or moved. To set up an HDFS nameservice, follow the instructions for enabling HDFS high availability.

To prevent data loss, you must configure HDFS to durably write data on file close. If the following configuration changes are not made, Accumulo issues warning messages until the problem is corrected.

1. Navigate to your cluster's **HDFS service** page.
2. Click the **Configuration** tab.
3. Search for `hdfs-site.xml`.
4. Search the Service-Wide / Advanced section property for "HDFS Service Advanced Configuration Snippet (Safety Valve) for hdfs-site.xml"**.**
5. Click the field and add this snippet:

```
<property>
    <name>dfs.datanode.synconclose</name>
    <value>true</value>
</property>
```

6. Search for any Gateway group properties labeled **HDFS Client Advanced Configuration Snippet (Safety Valve) for hdfs-site.xml**.
7. Click the field and add this snippet.

```
<property>
    <name>dfs.datanode.synconclose</name>
    <value>true</value>
</property>
```

8. Save your changes with a descriptive message, such as "HDFS changes for Accumulo."
9. Redeploy client configurations for the HDFS service.
10. Restart the HDFS service.

Note: If you want to run Accumulo with HDFS encryption, you must disable HDFS trash, which Accumulo uses by default. HDFS trash is not compatible with HDFS Transparent Encryption.

To disable HDFS trash in Accumulo, see HDFS storage demands due to retained HDFS trash. For information about HDFS encryption, see HDFS Transparent Encryption.

Step 4: Add the Accumulo Service

1. Navigate to the Cloudera Manager **Home** page.

2. Click the actions menu for the cluster you want to add the Accumulo service to (for example, Cluster 1), and select **Add a Service.**
3. Select the service labeled **Accumulo** or **Accumulo 1.6**, depending on your version of Cloudera Manager, and click **Continue.**
4. Select the dependent services and click **Continue.**
5. Assign the Accumulo roles to the hosts in your cluster. Assign a **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 do not already have other Accumulo roles assigned.
6. Click **Continue.**
7. Configure the **Accumulo Instance Secret**. Protect this secret because the security of the Accumulo service relies on it.
8. Configure the **Accumulo Instance Name**.
9. Configure the **Root Password** setting. Change this from the default setting because the root user is capable of bypassing Accumulo's cell level authorizations and has full administrative access to the cluster.
10. Configure the **Trace User** and **Trace Password** settings. Do **not** leave the **Trace User** set to **root**, because this is not a secure configuration. For example, use the value **trace** for the Trace User and a randomly generated password for Trace Password.
11. Click **Continue.** Cloudera Manager does the initial service setup.
    If you entered non-default values for the Trace User and Trace Password settings, your first run will **fail to start** any **Tracer** role instances. The service is otherwise operational.
12. Follow the instructions in step 5 of Test the Accumulo Shell to create the Trace User you just configured, with the appropriate permissions. You will not see the new table **trace** until you complete these instructions.
13. Click the Cloudera Manager logo in the upper left corner to exit the wizard.
14. Click the service labeled **Accumulo** or **Accumulo 1.6**, depending on your version of Cloudera Manager.
15. Click the **Tracer** roles that are in **Down** state, which takes you to an **Instances** view.
16. Select all of the **Tracer** roles that are down.
17. Click on the **Actions for Selected** dropdown.
18. Click **Start**.
19. In the dialogue box, click **Start**.
20. Click **Close.**
21. Click **Status** to return to the service overview.

Verify your installation by following the instructions in Test the Accumulo Shell.

> **Important:**
>
> By default, Cloudera provides Tablet Servers with 64 MB of memory in a memory-constrained system. This will cause Tablet Servers to crash on startup.
>
> To prevent Tablet Servers from crashing on startup, set memory as follows:
>
> - Tablet Server roles >= 512 MiB of memory
> - tserver.cache.index.size >= 256MiB
> - tserver.cache.data.size >= 128MiB

Step 5: Performance Tuning – Relaxing WAL Durability (Optional)

For increased write throughput, use the `BatchWriter` API to ingest data into Accumulo. However, using this API reduces data durability. Use this setting only in environments with reliable UPS.

To enable this setting, perform the following configuration changes:

1. Navigate to your cluster's **Accumulo** service page.
2. Click **Configuration** and then click **View and Edit**.
3. Search for "Tablet Server accumulo-site.xml".
4. Find the Tablet Server Default Group / Advanced section's property for "**Tablet Server Advanced Configuration Snippet (Safety Valve) for accumulo-site.xml**".
5. Click the field and add the following snippet.

```
<property>
    <name>table.durability</name>
    <value>flush</value>
</property>
```

6. Save your changes with a descriptive message, such as "Accumulo WAL Durability Changes".
7. Redeploy client configurations.
8. Restart the Tablet Servers in your Accumulo 1.7 Service.

Step 6: Security Configuration – Kerberos for Accumulo Clients (optional)

If you enable Kerberos for your CDH cluster, Cloudera Manager makes the configuration changes required for the Accumulo service to run on top of secured HDFS and ZooKeeper services. Additionally, you can enable Kerberos-based authentication and authorization for clients of the Accumulo cluster by making the following changes. These changes require a cluster outage period

1. Go to your cluster's **Accumulo** service page.
2. Click **Configuration** and then click **View and Edit**.
3. Search the Accumulo (Service-wide) / Main section properties for "Trace User".

4. Click the field and set the value to the principal you will use for the trace user. It can be with an instance (for example, `trace/some.host.example.com@EXAMPLE.COM`), or without one (for example, `trace@EXAMPLE.COM`).
5. Search the Accumulo (Service-wide) / Advanced section property for "Service Advanced Configuration Snippet (Safety Valve) for accumulo-site.xml".
6. Click the field and add the following snippet.

```
 <property>
   <name>instance.rpc.sasl.enabled</name>
   <value>true</value>
 </property>
 <property>
   <name>rpc.sasl.qop</name>
   <value>auth-conf</value>
   <description>
     one of “auth”, “auth-int”, or “auth-conf”. If you use delegation
     tokens, you must set this to “auth-conf” to avoid leaking tokens
     to network observers.
   </description>
 </property>
 <property>
   <name>instance.security.authenticator</name>
   <value>
     org.apache.accumulo.server.security.handler.KerberosAuthenticator
   </value>
 </property>
 <property>
   <name>instance.security.authorizor</name>
   <value>
     org.apache.accumulo.server.security.handler.KerberosAuthorizor
   </value>
 </property>
 <property>
   <name>instance.security.permissionHandler</name>
   <value>
   org.apache.accumulo.server.security.handler.KerberosPermissionHandler
   </value>
 </property>
 <property>
   <name>trace.token.type</name>
   <value>
     org.apache.accumulo.core.client.security.tokens.KerberosToken
   </value>
 </property>
 <property>
   <name>trace.token.property.keytab</name>
   <value>accumulo16.keytab</value>
```

```
  <description>
    Path to a keytab corresponding to the principal given in trace.user.
    Will be used by the Trace and Monitor roles to write and read,
    respectively, the trace table.
</property>
```

7. Save your changes with a descriptive message, such as "Accumulo Kerberos for clients Changes".
8. Stop all Tracer roles in the cluster.
9. Redeploy client configurations.
10. Restart all roles except Tracers.
11. On each Gateway machine where you will run an Accumulo client, for each user who will run an Accumulo client, create a client configuration file in `~/.accumulo/config` with the following contents:

```
# Each of these must match what's in the server configs exactly
instance.rpc.sasl.enabled=true
rpc.sasl.qop=auth-conf
# Should match the System User configuration value
kerberos.server.primary=accumulo
```

12. Follow the instructions for setting up an administrative user described in the Administrative User section of the user manual. The remaining instructions presume this user is named `accumulo_admin@EXAMPLE.COM`.
13. Set up access to the `trace` table for the trace principal. For example, on a Gateway machine with the client configurations listed in Step 11, log in to Kerberos using the administrative user from the previous step, and then run the following Accumulo shell commands:

```
$ accumulo shell
Shell - Apache Accumulo Interactive Shell
-
- version: 1.7.2-cdh5.5.0
- instance name: cloudera
- instance id: e30831be-b22f-4dfb-99e3-40640c5ae82f
-
- type 'help' for a list of available commands
-
accumulo_admin@EXAMPLE.COM@ACCUMULO16-1> createuser trace@EXAMPLE.COM
accumulo_admin@EXAMPLE.COM@ACCUMULO16-1> grant Table.READ -t trace -u
trace@EXAMPLE.COM
accumulo_admin@EXAMPLE.COM@ACCUMULO16-1> grant Table.WRITE -t trace -u
trace@EXAMPLE.COM
accumulo_admin@EXAMPLE.COM@ACCUMULO16-1> grant Table.ALTER -t trace -u
trace@EXAMPLE.COM
```

If you configure the Trace User with a `_HOST` instance component, you must repeat the previous commands with an instance for each host that will run a Tracer or Monitor role.

14. Start all Tracer roles.

Install Apache Accumulo from Distribution Packages

This section describes how to install the Cloudera packaging of Accumulo from packages (RPM or DEB) instead of using Cloudera Manager.

> **Important:**
>
> If you are installing Accumulo from packages, you need to have installed CDH from RPM or DEB and not from parcels.

Step 1: Add or Build the Accumulo Repository

- If you are installing Accumulo on a Red Hat system, you can download the Cloudera packages using yum or your web browser.
- If you are installing Accumulo on an SLES system, you can download the Cloudera packages using zypper, YaST, or your web browser.
- If you are installing Accumulo on an Ubuntu or Debian system, you can download the Cloudera packages using apt or your web browser.

On Red Hat-compatible Systems

Use one of the following methods to add or build the Accumulo repository or download the packages on Red Hat-compatible systems by using the instructions in one of the following sections:

- Adding the Accumulo Repository
- Building a yum Repository

Do this on all systems in the cluster.

Adding the Accumulo Repository

Follow the link in the table below that matches your Red Hat or CentOS system, navigate to the repo file for your system, and save it in the `/etc/yum.repos.d/` directory.

For OS Version	Follow this Link
Red Hat/CentOS/Oracle 5	CDH 5 for Red Hat/CentOS/Oracle 5
Red Hat/CentOS/Oracle 6	CDH 5 for Red Hat/CentOS/Oracle 6
Red Hat/CentOS/Oracle 7	CDH 5 for Red Hat/CentOS/Oracle 7

Continue to Step 2: Install Accumulo.

Building a yum Repository

To create your own yum repository, download the appropriate repo file, create the repository, distribute the repo file, and set up a web server, as described in Creating a Local yum Repository.

On SLES Systems

Use one of the following methods to download the Accumulo repository or packages on SLES systems:

- Adding the Accumulo Repository
- Building a SLES Repository

Do this on all systems in the cluster.

Adding the Accumulo Repository

1. Run the following command:

```
$ sudo zypper addrepo -f
http://archive.cloudera.com/accumulo-c5/sles/11/x86_64/cdh/cloudera-accumulo.repo
```

2. Update your system package index:

```
$ sudo zypper refresh
```

Building a SLES Repository

To create your own SLES repository, create a mirror of the Accumulo SLES directory by following these instructions.

Continue to Step 2: Install Accumulo.

On Ubuntu or Debian Systems

Use one of the following methods to add or build the Accumulo repository or download the packages on Ubuntu or Debian systems by using the instructions in one of the following sections:

- Adding the Accumulo Repository
- Building a Debian Repository

Do this on all the systems in the cluster.

Adding the Accumulo Repository

Follow the link in the table below that matches your Ubuntu or Debian system, navigate to the repo file for your system, and save it in the `/etc/apt/sources.list.d/` directory.

For OS Version	Follow this Link
Ubuntu Trusty	CDH 5 for Ubuntu Trusty
Debian	CDH 5 for Debian Wheezy

Continue to Step 2: Install Accumulo.

Building a Debian Repository

To create your own apt repository, create a mirror of the Accumulo Debian directory, and then create an apt repository from the mirror.

Continue to Step 2: Install Accumulo.

Step 2: Install Accumulo

Important:

Before proceeding, you must decide where to deploy the Accumulo Master, Accumulo Monitor, Accumulo Garbage Collector, and Accumulo Tracer daemons. Follow these guidelines:

- The Accumulo Master and Accumulo Monitor should run on the same master host unless the cluster is large (more than a few tens of nodes), and the master host (or hosts) should not run the Accumulo Tablet Server service.
- In a large cluster, it is important that the Accumulo Garbage Collector and Accumulo Tracer run on machines separate from the Accumulo Master.
- Each node in the cluster **except the master host(s)** should run the Accumulo Tablet Server service. In particular, these services should be run on every DataNode.

1. Install and deploy CDH 5.
2. Install and deploy ZooKeeper.
3. Install each type of daemon package on the appropriate systems, as follows:

Installation package and OS	**Install commands**
Accumulo Master host running:	
Red Hat/CentOS compatible	sudo yum clean all; sudo yum install accumulo-master
SLES	sudo zypper clean –all; sudo zypper install accumulo-master
Ubuntu or Debian	sudo apt-get update; sudo apt-get install accumulo-master
Accumulo Monitor host running:	
Red Hat/CentOS compatible	sudo yum clean all; sudo yum install accumulo-monitor
SLES	sudo zypper clean –all; sudo zypper install accumulo-monitor
Ubuntu or Debian	sudo apt-get update; sudo apt-get install accumulo-monitor
Accumulo Garbage Collector host running:	
Red Hat/CentOS compatible	sudo yum clean all; sudo yum install accumulo-gc
SLES	sudo zypper clean –all; sudo zypper install accumulo-gc
Ubuntu or Debian	sudo apt-get update; sudo apt-get install accumulo-gc
Accumulo Tracer host running:	
Red Hat/CentOS compatible	sudo yum clean all; sudo yum install accumulo-tracer
SLES	sudo zypper clean –all; sudo zypper install accumulo-tracer
Ubuntu or Debian	sudo apt-get update; sudo apt-get install accumulo-tracer

All cluster hosts except Accumulo Master, Accumulo Monitor, Accumulo Garbage Collector, and Accumulo Tracer hosts running:	
Red Hat/CentOS compatible	sudo yum clean all; sudo yum install accumulo-tserver
SLES	sudo zypper clean –all; sudo zypper install accumulo-tserver
Ubuntu or Debian	sudo apt-get update; sudo apt-get install accumulo-tserver
All client hosts running:	
Red Hat/CentOS compatible	sudo yum clean all; sudo yum install accumulo
SLES	sudo zypper clean –all; sudo zypper install accumulo
Ubuntu or Debian	sudo apt-get update; sudo apt-get install accumulo

Step 3: Configure HDFS

Establish an HDFS nameservice on the cluster that will run Accumulo. Doing so significantly reduces future administrative tasks if a NameNode needs to be replaced or moved. To set up an HDFS nameservice, follow the instructions for enabling HDFS high availability.

To guard against data loss, you must configure HDFS to durably write data on file close. If the following configuration changes are not made, Accumulo issues warning messages until the problem is corrected.

1. Edit the hdfs-site.xml used in your cluster and ensure it contains the following snippet:

```
<property>
    <name>dfs.datanode.synconclose</name>
    <value>true</value>
</property>
```

2. Synchronize the updated hdfs-site.xml file across your cluster.
3. Restart all HDFS DataNodes.

Step 4: Configure Accumulo for Your Environment

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

1. On every host, check the following properties in `/etc/accumulo/conf/accumulo-site.xml` to ensure that they are configured for your environment. Replace `secret` as the password in the `instance.secret` property to a value for your setup.

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

<property>
  <name>instance.secret</name>
  <value>secret</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>trace.password</name>
  <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`.
3. Review the service specific options, such as Java heap size, in the `/etc/default/accumulo` file:

```
ACCUMULO_TSERVER_OPTS="-Xmx1g -Xms1g -XX:NewSize=500m
-XX:MaxNewSize=500m"
ACCUMULO_MASTER_OPTS="-Xmx2g -Xms1g"
ACCUMULO_MONITOR_OPTS="-Xmx2g -Xms256m"
ACCUMULO_GC_OPTS="-Xmx256m -Xms256m"
ACCUMULO_GENERAL_OPTS="-XX:+UseConcMarkSweepGC
-XX:CMSInitiatingOccupancyFraction=75
-Djava.net.preferIPv4Stack=true"

ACCUMULO_OTHER_OPTS="-Xmx1g -Xms256m"

ACCUMULO_KILL_CMD='kill -9 %p'
```

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 `/etc/accumulo/conf`, and add the FQDN or IP address of the Tablet Servers (one per line) to the `/etc/accumulo/conf/slaves` file.
- On a multi-host cluster, the contents of the `/etc/accumulo/conf` directory must always be synchronized across all Accumulo servers within a cluster. This can be done using configuration management, version control, or a utility such as `rsync`. Servers with out-of-sync configurations are not allowed to join the cluster.

Step 5: Initialize Accumulo

To initialize Accumulo:

1. 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 751 /accumulo
$ hadoop fs -chmod 750 /user/accumulo
$ exit
```

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.

   Override the following defaults in /etc/accumulo/conf/accumulo-site.xml:

   - Set `logger.dir.walog` to a directory on a partition with sufficient space for write-ahead logs.
   - Set `tracer.user` and `tracer.password` to your required values.

```
$ sudo -i service accumulo-master 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 are warned if you did not change your instance secret in `/etc/accumulo/conf/accumulo-site.xml`.
- If the `Hadoop Filesystem is` line contains `file://` or `fs://` instead of `hdfs://`, HDFS is not properly configured.
- If the command fails with an htrace message with a link similar to the following:

  ```
  org/apache/htrace/core/Tracer$Builder
  java.lang.NoClassDefFoundError
  ```

  you need to add the htrace jar (`/lib/hadoop/client/htace-core*.jar`) to your `accumulo` classpath in `accumulo-site.xml`.

Step 6: Start Accumulo

To start Accumulo:

1. Run the following commands on the following hosts:

For the following service	Run this command
Accumulo Master	`sudo -i service accumulo-master start`
Accumulo Monitor	`sudo -i service accumulo-monitor start`
Accumulo Garbage Collector	`sudo -i service accumulo-gc start`
Accumulo Tracer	`sudo -i service accumulo-tracer start`
All cluster hosts except Accumulo Master, Accumulo Monitor, Accumulo Garbage Collector, and Accumulo Tracer hosts	`sudo -i service accumulo-tserver start`

2. Connect to Accumulo on http://localhost:50095. You can check the status of each daemon with the following command:

```
$ sudo -i service accumulo-<service> status
```

where `<service>` is one of `master`, `monitor`, `gc`, `tracer`, or `tserver`.

3. You can stop each daemon with the following command:

```
$ sudo -i service accumulo-<service> stop
```

   where `<service>` is one of `master`, `monitor`, `gc`, `tracer`, or `tserver`.

Verify your installation by following the instructions in Test the Accumulo Shell.

Step 7: Performance Tuning – Relaxing WAL Durability (Optional)

For increased write throughput, use the `BatchWriter` API to ingest data into Accumulo. However, using this API reduces data durability. Use this setting only in environments with reliable UPS.

To enable this setting, add the following text to the `accumulo-site.xml` file and distribute the change across hosts running Accumulo roles in the cluster:

```
<property>
    <name>table.durability</name>
    <value>flush</value>
</property>
```

Restart all Tablet Servers following this change.

Upgrading from Accumulo 1.6 to Accumulo 1.7 Using Parcels

If you are running CDH 5 and have Accumulo 1.6 installed, follow these instructions to upgrade to Accumulo 1.7:

1. Install the Accumulo 1.7 parcel by following the instructions in steps 2 through 5 in Install Apache Accumulo 1.7 Using Cloudera Manager.
2. In the last step of the installation wizard, select "just activate".
3. Verify that you have checked the following:
   a. Make sure that Tablet Server roles are configured with a minimum of 512 MB of memory.
   b. Adjust WAL durability for greater write throughput. See Step 7: Performance Tuning – Relaxing WAL Durability (Optional). Note that the necessary configuration property changed names compared to Accumulo 1.6.
4. Restart the Accumulo service.
5. When you have verified that Accumulo 1.7 is working correctly, uninstall the Accumulo 1.6 parcel.

Upgrading from Accumulo 1.6 to Accumulo 1.7 Using Packages

If you are using packages for your cluster, follow these instructions to update your CDH version.

1. Stop all Accumulo services, make copies of your Accumulo configurations, and remove the Accumulo packages from all nodes in your cluster, using your system's package manager.

2. Remove the Accumulo repo from your package manager on all nodes in the cluster.
3. Upgrade CDH according to the documentation for packages.
4. Follow the CDH 5 instructions in Step 1 and Step 2 of Install Apache Accumulo from Distribution Packages.
5. Follow the instructions in Step 4 of Install Apache Accumulo from Distribution Packages, copying your previous `accumulo-site.xml` configuration and updating settings.

Troubleshooting the Accumulo Installation

Under-replicated block exceptions or cluster failure occurs on small clusters

By default, Accumulo attempts to use a replication factor of 5 for the metadata table, ignoring the "table.file.replication" setting. Normally, Cloudera Manager does not set a maximum replication factor. This causes under-replication warnings until you correct the number of nodes or until you manually adjust the replication setting on that table.

If `dfs.replication.max` has been adjusted to match the number of cluster nodes, attempts by Accumulo to create new files for its internal tables will fail.

To fix the issue:

1. Edit `dfs.replication.max` setting for HDFS to be >= 5.
2. Adjust replication on the metadata and root tables to be less than or equal to the number of DataNodes.
3. Readjust `dfs.replication.max` to lower it again.

For example, to adjust the replication in the Accumulo shell:

```
$> config -t accumulo.metadata -s table.file.replication=3
$> config -t accumulo.root -s table.file.replication=3
```

HDFS storage demands due to retained HDFS trash

By default, Accumulo uses HDFS trash if it is enabled for all files it deletes, including write-ahead logs and long-term storage files that have been obviated due to compaction. By default, the retention period for the HDFS trash is 24 hours. On Accumulo installations with a heavy write workload, this can result in a large amount of data accumulating in the trash folder for the service user.

As a workaround, periodically run the `hdfs dfs -expunge` command as the Accumulo service user. The command must be run twice each time you want to purge a backlog of data; the first time creates a trash checkpoint, and the second removes that checkpoint immediately.

Alternatively, you can tune the amount of time HDFS retains trash to control how much data Accumulo saves. This change is HDFS-wide and impacts the ability to recover from accidental deletions unrelated to Accumulo.

To change the HDFS trash setting in Cloudera Manager:

1. Go to the **HDFS service** page.
2. Click the **Configuration** tab.
3. Search for “trash”.
4. Change the **Filesystem Trash Interval** to a smaller value; for example, 4 hours.
5. Save your changes with a descriptive message, such as "HDFS changes for Accumulo."
6. Redeploy client configurations for the HDFS service.
7. Restart the HDFS service.

In some deployments, you might not want to change the system-wide retention period for HDFS trash. If that is the case, you can disable Accumulo’s use of the HDFS trash entirely. If you do so, any deletions through the Accumulo APIs are unrecoverable.

To configure Accumulo to skip the HDFS trash in Cloudera Manager:

1. Go to the **Accumulo service** page.
2. Click the **Configuration** tab.
3. Search for “accumulo-site.xml”.
4. Search the Accumulo (Service-wide) / Advanced section's property for "**Service Advanced Configuration Snippet (Safety Valve) for accumulo-site.xml**".
5. Click the field and add the following snippet.

```
<property>
    <name>gc.trash.ignore</name>
    <value>true</value>
</property>
```

6. Save your changes with a descriptive message, such as "Accumulo changed to skip HDFS trash."
7. Redeploy client configurations.
8. Restart the your Accumulo Service.

Test the Accumulo Shell

You can now run the Accumulo shell on any client hosts (for Cloudera Manager installs, hosts assigned the **Gateway** role) in your cluster. By default, the user **root** is created and given the password **secret**. If you did not set a different password during install, change the root user password.

The following steps will verify that the Accumulo shell works while allowing you to change the root user password.

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

```
$ accumulo shell -u root
Enter current password for 'root'@'accumulo': ******

Shell - Apache Accumulo Interactive Shell
-
- version: 1.7.2-cdh5.5.0-SNAPSHOT
- instance name: cloudera
- instance id: e30831be-b22f-4dfb-99e3-40640c5ae82f
-
- 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
Enter current password for 'root'@'accumulo': *********

Shell - Apache Accumulo Interactive Shell
-
- version: 1.7.2-cdh5.5.0-SNAPSHOT
- instance name: cloudera
- instance id: e30831be-b22f-4dfb-99e3-40640c5ae82f
-
- type 'help' for a list of available commands
```

```
-
root@accumulo>
```

4. Verify that you can list tables.

```
root@accumulo> tables
accumulo.metadata
accumulo.replication
accumulo.root

root@accumulo>
```

5. If the trace table does not exist, make sure that you have created the trace user. Use the same password you used for the `trace.password` setting in `/etc/accumulo/conf/accumulo-site.xml` for a manually managed cluster or the **Trace Password** setting in Cloudera Manager installations.

```
root@accumulo> tables
accumulo.metadata
accumulo.replication
accumulo.root
root@cloudera> createuser trace
Enter new password for 'trace': *****
Please confirm new password for 'trace': *****
root@cloudera> grant System.CREATE_TABLE -s -u trace
root@cloudera> tables
accumulo.metadata
accumulo.replication
accumulo.root
trace

root@cloudera> revoke System.CREATE_TABLE -s -u trace
```

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

Using Sqoop 1 with Accumulo

CDH 5.5.0 and higher include Sqoop bindings for import/export of data with Accumulo. For instructions on invoking Sqoop with Accumulo as a source or sink, see the Sqoop documentation.

When running the sqoop command, you might see warning messages about failing to create `/usr/lib/accumulo/logs`. These messages are safe to ignore.

Sqoop 1 Client Under CDH 5 and Cloudera Manager

To use Sqoop integration, you must perform the following configuration changes:

1. Navigate to your cluster's Sqoop 1 Client service page.
2. Click **Configuration**.
3. Search for "sqoop-env.sh".
4. Look for the Gateway Default Group / Advanced section's property for **Sqoop 1 Client Client Advanced Configuration Snippet (Safety Valve) for sqoop-conf/sqoop-env.sh.**
5. Click the field and add the snippet appropriate for your installation, ensuring that each line ends with a bash comment hash, '#'.
    - For parcels:
      ```
      export ACCUMULO_CONF_DIR=/etc/accumulo/conf #
      export ACCUMULO_HOME=/opt/cloudera/parcels/ACCUMULO/lib/accumulo #
      export HADOOP_CLIENT_HOME=/opt/cloudera/parcels/CDH/lib/hadoop/client #
      export HADOOP_PREFIX=/opt/cloudera/parcels/CDH/lib/hadoop #
      export ZOOKEEPER_HOME=/opt/cloudera/parcels/CDH/lib/zookeeper #
      ```
    - For packages:
      ```
      export ACCUMULO_CONF_DIR=/etc/accumulo/conf #
      export HADOOP_CLIENT_HOME=/usr/lib/hadoop/client #
      export HADOOP_PREFIX=/usr/lib/hadoop #
      export ZOOKEEPER_HOME=/usr/lib/zookeeper #
      ```
6. Save your changes with a descriptive message, such as "Sqoop changes for Accumulo."
7. Redeploy client configurations for the Sqoop 1 Client service.

Sqoop 1 Without Cloudera Manager

To use Sqoop integration, you must perform the following configuration changes:

1. If you do not already have `/etc/sqoop/conf/sqoop-env.sh`, create it.
   ```
   # cp /etc/sqoop/conf/sqoop-env-template.sh \
        /etc/sqoop/conf/sqoop-env.sh
   ```
2. Add the following exports to this `sqoop-env.sh` file. Ensure that they match your actual installation locations.
    - For parcels:
      ```
      export ACCUMULO_CONF_DIR=/etc/accumulo/conf
      export ACCUMULO_HOME=/opt/cloudera/parcels/ACCUMULO/lib/accumulo
      export HADOOP_CLIENT_HOME=/opt/cloudera/parcels/CDH/lib/hadoop/client
      export HADOOP_PREFIX=/opt/cloudera/parcels/CDH/lib/hadoop
      export ZOOKEEPER_HOME=/opt/cloudera/parcels/CDH/lib/zookeeper
      ```
    - For packages:
      ```
      export ACCUMULO_CONF_DIR=/etc/accumulo/conf
      export HADOOP_CLIENT_HOME=/usr/lib/hadoop/client
      export HADOOP_PREFIX=/usr/lib/hadoop
      export ZOOKEEPER_HOME=/usr/lib/zookeeper
      ```
3. Save your changes.
4. Synchronize this file across all nodes that will run Sqoop commands.

Using Accumulo with Maven

If you want to build applications or tools with the Cloudera packaging of Accumulo, and you are using Maven or Ivy for dependency management, you can pull the Accumulo artifacts from the Cloudera Maven repository. The repository is available at https://repository.cloudera.com/artifactory/cloudera-repos/. The following is a sample snippet from a POM (pom.xml) file:

```
<repositories>
  <repository>
    <id>cloudera</id>
    <name>Cloudera Releases Repository</name>
    <url>https://repository.cloudera.com/artifactory/cloudera-repos/</url>
  </repository>
</repositories>
```

CDH 5-Compatible Releases

The following table lists the project name, groupId, artifactId, and version required to access each CDH 5-compatible artifact. Client applications should only require the accumulo-core artifact as a dependency and might need the accumulo-maven-plugin for running integration tests.

Project	groupId	artifactId	version
Accumulo	org.apache.accumulo	accumulo	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-core	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-examples-simple	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-fate	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-gc	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-master	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-maven-plugin	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-minicluster	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-monitor	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-proxy	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-server-base	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-start	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-test	1.7.2-cdh5.5.0

	org.apache.accumulo	accumulo-trace	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-tracer	1.7.2-cdh5.5.0
	org.apache.accumulo	accumulo-tserver	1.7.2-cdh5.5.0

Default Ports

If your cluster is running firewall software, you might need to allow communication between hosts on specific ports. The following table lists the default port for each server process and the configuration property used to change that value.

Accumulo Process	Port	Property
Garbage Collector	50091	gc.port.client
Master	10010	master.port.client
Monitor (Log Forwarding)	4560	monitor.port.log4j
Monitor (Client Port)	50091	monitor.port.client
Tablet Server	10011	tserver.port.client
Tracer	12234	trace.port.client
Master Replication Service	10001	master.replication.coordinator.port
TabletServer Replication Service	10002	replication.receipt.service.port

Creating a Local yum Repository

You can set up a local yum repository to install Accumulo on the machines in your cluster. You might want to do this in the following scenarios:

- The computers in your cluster do not have Internet access. You can use yum to do an installation on those machines by creating a local yum repository.
- You want to keep a stable local repository to ensure that any new installations (or reinstallations on existing cluster members) use exactly the same bits.
- Using a local repository is the most efficient way to distribute the software to cluster members.

To set up your own internal mirror, do the following.

Note: Before You Start

These instructions assume you already have the appropriate Cloudera repo file on the system on which you are going to download the local repository. If this is not the case, follow the instructions in Adding the Accumulo Repository.

1. On a computer that has Internet access, install the `yum-utils` and `createrepo` packages if they are not already installed (`yum-utils` includes the `reposync` command):

```
$ sudo yum install yum-utils createrepo
```

2. On the same computer used in Step 1, download the yum repository to a temporary location. On Red Hat/CentOS 6, you can use a command such as:

```
$ reposync -r cloudera-accumulo
```

> **Note:**
> `cloudera-accumulo` is the name of the repository on your system. The name is in square brackets and usually is on the first line of the repo file, which in this example is `/etc/yum.repos.d/cloudera-accumulo.repo`.

3. Copy all of the RPMs to the machine that will serve the local repository and place them in a directory served by your web server. For this example, it is called `/var/www/html/accumulo/1.7.0/RPMS/x86_64` (or `i386` for 32-bit systems). Make sure you can remotely access the files in the directory you just created. The URL should look like `http://<yourwebserver>/accumulo/1.7.0/RPMS/`.
4. On the server in step 3, go to `/var/www/html/accumulo/1.7.0/` and type the following command:

```
$ createrepo .
```

   This will create or update the necessary metadata so yum can understand this new repository. You will see a new directory named `repodata`.

> **Important:**
>
> Check the permissions of the subdirectories under `/var/www/html/accumulo/1.7.0/`. Make sure they are all readable by your web server user.

5. Edit the repo file you downloaded previously and replace the line starting with `baseurl=` or `mirrorlist=` with `baseurl=http://<yourwebserver>/accumulo/1.7.0/`.
6. Save this modified repo file in `/etc/yum.repos.d/`, and check that you can install Accumulo through yum.

Example:

```
$ yum update && yum install accumulo
```

After you have confirmed that your internal mirror works, you can distribute this modified repo file to all your machines, and they should all be able to install Accumulo without needing access to the Internet. Follow the instructions in Step 2: Install Accumulo.