Hadoop 0.14 Upgrade

Upgrade Guide for Hadoop-0.14

This page describes upgrade information that is specific to Hadoop-0.14. The normal upgrade described in Hadoop Upgrade page still applies for Hadoop-0.14.

Upgrade Path

We have tested upgrading from 0.12 to 0.14 and 0.13.1 to 0.14. However we recommend upgrading to 0.13.1 before this upgrade. Hadoop-0.13 adds an important upgrade related feature that is very useful for upgrades like the one described here. While upgrading from 0.13, if something badly goes wrong, you can always *rollback* to pre-upgrade state by installing 0.13 again. See Hadoop Upgrade page for more information. Upgrading from 0.11 or earlier versions is very similar to upgrading from 0.12 and is expected to work, though not extensively tested.

Brief Upgrade Procedure

In most cases, upgrade to Hadoop-0.14 completes without any problems. In these cases, administrators are not required to be very familiar with rest of the sections in this document. The simple upgrade steps are same as listed in Hadoop Upgrade:

1. If you are running Hadoop-0.13.x, make sure the cluster is finalized.
2. Stop map-reduce cluster(s) and all client applications running on the DFS cluster.
3. Stop DFS using the shutdown command.
4. Install new version of Hadoop software.
5. Start DFS cluster with –upgrade option.
6. Wait for for cluster upgrade to complete.
7. Start map-reduce cluster.
8. Verify the components run properly and finalize the upgrade when convinced.

The rest of the document describes what happens once the cluster is started with `-upgrade` option.

Block CRC Upgrade

Hadoop-0.14 maintains checksums for HDFS data differently than earlier versions. Before Hadoop-0.14, checksum for a file `f.txt` is stored in another HDFS file `.f.txt.crc`. In Hadoop-0.14, there are no such *shadow* checksum files. In stead, checksum is stored with each *block* of data at datanodes. HADOOP-1134 describes this feature in great detail. In order to migrate to the new structure, each datanode reads checksum data from `.crc` files in HDFS for each of its blocks and stores the the checksum next to the block in local filesystem.

Depending on number of blocks and number of files in HDFS, upgrade can take anywhere from a few minutes to few hours.

There are three stages in Block CRC upgrade :

1. **Safe Mode** : Similar to normal restart of the cluster, namenode waits for datanodes in the cluster to report their blocks. The cluster may wait in the state for a long time if some of the datanodes do not report their blocks.
2. **Datanode Upgrade** : Once the most of the blocks are reported, namenode asks the registered datanodes to start their local upgrade. Namenode waits for for *all* the datanodes to complete their upgrade.
3. **Deleting `.crc` files** : Namenode deletes `.crc` files that were previously used for storing checksum.

Monitoring the Upgrade

The cluster stays in *safeMode* until the upgrade is complete. HDFS webui is a good place to check if safeMode is on or off. As always, log files from *namenode* and *datanode* are useful when nothing else helps.

Once the cluster is started with `-upgrade` option, the simplest way to monitor the upgrade is with '`dfsadmin -upgradeProgress status`' command.

First Stage : Safe Mode

The actual Block CRC upgrade starts after all or most of the datanodes have reported their blocks.

```
$ bin/hadoop dfsadmin -upgradeProgress status
Distributed upgrade for version -6 is in progress. Status = 0%

        Upgrade has not been started yet.
        Last Block Level Stats updated at : Thu Jan 01 00:00:00 UTC 1970
        ....
```

The message `Upgrade has not been started yet` indicates that namenode is in the first stage. When *status* is at 0%, usually it is in this stage. If some datanodes don't start, check HDFS webui to find which datanodes are listed under *Dead Nodes* table.

Second Stage : Datanode Upgrade

During this stage a typical output from `upgradeProgress` command looks like this:

```
$ bin/hadoop dfsadmin -upgradeProgress status
Distributed upgrade for version -6 is in progress. Status = 78%

        Last Block Level Stats updated at : Mon Aug 13 22:23:30 UTC 2007
        Last Block Level Stats : Total Blocks : 1054713
                                 Fully Upgragraded : 40.94%
                                 Minimally Upgraded : 52.13%
                                 Under Upgraded : 6.93% (includes Un-upgraded blocks)
                                 Un-upgraded : 6.93%
                                 Errors : 0
        Brief Datanode Status  : Avg completion of all Datanodes: 91.59% with 0 errors.
                                 274 out of 893 nodes are not done.
```

- `Status = 78%` : This is a rough approximation of how much of upgrade is completed.
- `Block Level Stats` : Once the upgrade starts, Namenode iterates through all the block to check how many of the blocks are upgraded. This information is useful on large clusters where some datanodes may never complete upgrade of their blocks (discussed in later sections).
    - `Fully Upgraded` : Percentage of blocks, where the expected number of replicas are upgraded. E.g. if a block has replication of 3, it is considered *fully upgraded* if at least three datanodes that contain this blocks have finished upgrade of their blocks.
    - `Minimally Upgraded` : Similar to above, number of upgraded replicas is at least `dfs.min.replication` (default 1) and is less than expected number of replicas.
    - `Under Upgraded` : number of upgraded replicas is less than `dfs.min.replication`.
    - `Un-upgraded` : blocks with zero upgraded replicas.
- `Brief Datanode Status` : Each datanode reports its progress to the namenode during the upgrade. This shows average of percent completion on all the datanodes. This also shows how many datanodes have completed their upgrade. For the upgrade to proceed to next stage, all the datanodes should report completion of their local upgrade.

Note that in some cases, a few blocks might be *over-replicated*. In such a case upgrade might proceed to next stage even if some of the datanodes do not complete their upgrade. If `Fully Upgraded` is calculated to be 100%, namenode will proceed to next stage even if not all the datanodes have completed their upgrade.

Potential Problems during Second Stage

- *The upgrade might seem to be stuck* : Each datanode reports its progress once every minute. If the percent completion does not change even afeter a few minutes, some datanodes might have some unexpected problems. Use `details` option with `-upgradeProgress` command to check which datanodes seem stagnant.

    ```
    $ bin/hadoop dfsadmin -upgradeProgress details
    Distributed upgrade for version -6 is in progress. Status = 72%

            Last Block Level Stats updated at : Thu Jan 01 00:00:00 UTC 1970
            Last Block Level Stats : Total Blocks : 0
                                     Fully Upgragraded : 0.00%
                                     Minimally Upgraded : 0.00%
                                     Under Upgraded : 0.00% (includes Un-upgraded blocks)
                                     Un-upgraded : 0.00%
                                     Errors : 0
            Brief Datanode Status  : Avg completion of all Datanodes: 81.90% with 0 errors.
                                     352 out of 893 nodes are not done.

            Datanode Stats (total: 893): pct Completion(%) blocks upgraded (u) blocks remaining (r) errors
    (e)

                    192.168.0.31:50010        : 54 %         2136 u  1804 r  0 e
                    192.168.0.136:50010       : 73 %         3074 u  1085 r  0 e
                    192.168.0.24:50010        : 50 %         2044 u  1999 r  0 e
                    192.168.0.214:50010       : 100 %        4678 u  0 r     0 e
                    ...
    ```

    You can run this command through '`grep -v "100 %"`' to find the nodes that have not completed their upgrade. If the problem nodes can not be corrected, as a last resort you can check *Block Level Stats* to see if the upgrade can be *forced* to next stage. E.g. if 98% are fully-upgraded and 2% are minimally-upgraded, then you can reasonably be sure that at least one copy of a block is upgraded. You can force next stage with `force` option :

```
$ bin/hadoop dfsadmin -upgradeProgress force
Distributed upgrade for version -6 is in progress. Status = 90%

        Force Proceed is ON
        Last Block Level Stats updated at : Mon Aug 13 22:43:31 UTC 2007
        Last Block Level Stats : Total Blocks : 1054713
                                 Fully Upgragraded : 99.40%
                                 Minimally Upgraded : 0.60%
                                 Under Upgraded : 0.00% (includes Un-upgraded blocks)
                                 Un-upgraded : 0.00%
                                 Errors : 0
        Brief Datanode Status  : Avg completion of all Datanodes: 99.89% with 0 errors.
                                 1 out of 893 nodes are not done.
        NOTE: Upgrade at the Datanodes has finished. Deleteing ".crc" files
        can take longer than status implies.
```

Note `Force Proceed is ON` in the status message.

Third Stage : Deleting .crc files

Once the second stage is complete, Namenode reports 90% completion. It does not have a very good way of estimating time required for deleting the files. The *status* reports 90% completion all through this stage. Later tests with larger number of files indicates that it takes one hour to delete 2 million files on a rack server. The upgrade status report looks like the following.

```
$ bin/hadoop dfsadmin -upgradeProgress status
Distributed upgrade for version -6 is in progress. Status = 90%

        Last Block Level Stats updated at : Mon Aug 20 20:24:56 UTC 2007
        Last Block Level Stats : Total Blocks : 11604180
                                 Fully Upgragraded : 100.00%
                                 Minimally Upgraded : 0.00%
                                 Under Upgraded : 0.00% (includes Un-upgraded blocks)
                                 Un-upgraded : 0.00%
                                 Errors : 0
        Brief Datanode Status  : Avg completion of all Datanodes: 100.00% with 0 errors.
        NOTE: Upgrade at the Datanodes has finished. Deleteing ".crc" files
        can take longer than status implies.
```

Note the last two lines that inform that Namenode is currently deleting `.crc` files.

Upgrade is Finally Complete

Once the upgrade is complete, *safeMode* will be turned off and HDFS runs normally. There is no need to restart the cluster. Now enjoy the new and shiny Hadoop with leaner Namenode.

```
$ bin/hadoop dfsadmin -upgradeProgress status
There are no distributed upgrades in progress.
```

Memory requirements

HDFS nodes do not require more memory during the upgrade than for normal operation before the upgrade. We observed that Namenode might use 5-10% more memory (or more GC in JVM) during the upgrade. If the namenode was operating at the edge of its memory limits before the upgrade, it could potentially have some problems. At any time, cluster can be restarted and the HDFS resumes the upgrade.

Restarting a cluster

The cluster can be restarted during any stage of the upgrade and it will resume the upgrade.

Analyzing Log Files

As a last resort while diagnosing problems, administrator could look at logs at Namenode and Datanode. It might be information overload to list all the relevant log messages here. Of course, developers most appreciate if the relevant logs are attached while reporting problems with the upgrade, along with output from `-upgradeProgress` command.

Some of the warnings on log files are expected during the upgrade. For e.g. during the upgrade, datanodes fetch checksum data located on their peers. These data transfers utilize the new protocols in Hadoop-0.14 that require checksum data to be present along with block data. Since the checksum data is not yet located next to the block you will see the following warning int the datanode logs :

```
2007-08-18 07:17:38,698 WARN org.apache.hadoop.dfs.DataNode: Could not find metadata file for
blk_2214836660875523305
```