

Balancer Administrator Guide

The balancer is a tool that balances disk space usage on an HDFS cluster when some datanodes become full or when new empty nodes join the cluster. The tool is deployed as an application program that can be run by the cluster administrator.

SYNOPSIS

To start:

```
bin/hadoop-daemon.sh start balancer [-threshold <threshold>]
```

Example: *bin/hadoop-daemon.sh start balancer*

start the balancer with a default threshold of 10%

```
bin/hadoop-daemon.sh start balancer -threshold 5
```

start the balancer with a threshold of 5%

To stop:

```
bin/hadoop-daemon.sh stop balancer
```

DESCRIPTION

The threshold parameter is a fraction in the range of (0, 1) with a default value of 10%. The threshold sets a target for whether the cluster is balanced. A cluster is balanced if for each datanode, the utilization of the node (ratio of used space at the node to total capacity of the node) differs from the utilization of the cluster (ratio of used space in the cluster to total capacity of the cluster) by no more than the threshold value. The smaller the threshold, the more balanced a cluster will become. It takes more time to run the balancer for small threshold values. Also for a very small threshold the cluster may not be able to reach the balanced state when applications write and delete files concurrently.

The tool moves blocks from highly utilized datanodes to poorly utilized datanodes iteratively. In each iteration a datanode moves or receives no more than the lesser of 10G bytes or the threshold fraction of its capacity.

A system property that limits the balancer's use of bandwidth is defined in the default configuration file:

```
<property>  
  <name>dfs.balance.bandwidthPerSec</name>  
  <value>1048576</value>  
  <description> Specifies the maximum bandwidth that each datanode can utilize for the  
    balancing purpose in term of the number of bytes per second. </description>  
</property>
```

This property determines the maximum speed at which a block will be moved from one datanode to another. The default value is 1MB/s. The higher the bandwidth, the faster a cluster can reach the balanced state, but with greater competition with application processes. If an administrator changes the value of this property in the configuration file, the change is observed when HDFS is next restarted.

After the balancer is started, the administrator is prompted for an output file name where the balancer progress will be recorded. The administrator can monitor the running of the balancer by reading the output file. The following is the beginning portion of a sample output:

Iteration#	Bytes Already Moved	Bytes Left To Move	Bytes Being Moved
0	0 KB	40.88 TB	2.03 TB
1	2 TB	38.29 TB	2.01 TB
2	3.98 TB	36.38 TB	1.98 TB
3	5.94 TB	34.42 TB	1.96 TB

The output shows the balancer's status iteration by iteration. In each iteration it prints the iteration number, the total number of bytes that have been moved in the previous iterations, the total number of bytes that are left to move in order for the cluster to be balanced, and the number of bytes that are being moved in this iteration. Normally "Bytes Already Moved" is increasing while "Bytes Left To Move" is decreasing.

The balancer automatically exits when any of the following three conditions is satisfied:

1. The cluster is balanced;
2. No block can be moved;
3. No block has been moved for three consecutive iterations.

Upon exit, a balancer prints one of the following messages to the output file in corresponding to the above exit reasons:

1. The cluster is balanced. Exiting...
2. No block can be moved. Exiting...
3. No block has been moved for 3 iterations. Exiting...

The administrator can interrupt the execution of the balancer at any time by running the command "*bin/hadoop-daemon.sh stop balancer*" on the machine where the balancer is running.

Running multiple instances of the balancer in an HDFS cluster is not expected although it is not prohibited by the tool.