

Instructions

Before getting started please read 'How to contribute to Hadoop Projects' in the link provided below

<http://wiki.apache.org/hadoop/HowToContribute>

How to Interpret Source Files

There are two files and a folder you have to be concerned with in the package.

1. JAR file – **hadoop-hdfs-2.3.0.jar**
2. Script – **hdfs**
3. **Hadoop HDFS Project Folder** – You can learn more about it by looking at BUILDING.txt file in the above link. It clearly describes how to compile, build and modify source code.

Steps to run the CRBalancer

1. Replace the **hadoop-hdfs-2.3.0.jar** file in following path
HADOOP_HOME_PATH/share/Hadoop/hdfs
2. Replace the script file **hdfs** in the following folder
HADOOP_HOME_PATH/bin/hdfs
3. Run the script file with the following parameters as follows

hdfs -file <full path to computation ratio file> -namenodename <Hostname of the namenode>
-port <port number to access the namenode>

How should the configuration file related to the computation ratio look like?

<hostname1> <ratio>

<hostname2> <ratio>

<hostname3> <ratio>

.

.

.

<hostnameN> <ratio>

Note: The ratios are calculated by placing entire data set on a single file and finding their least common multiple. Place the file using **hadoop fs -put HDFS_DIRECTORY_PATH/filename**

Example:

hpxeon01 0.36

jedi05 0.54

Example Command

```
hdfs crbalancer -file /user/sanket/cramp.txt -namenodename hpaxon01 -port 54310
```

How to View Source Code

1. Open Hadoop hdfs project folder.
2. Navigate to the path `hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/crbalancer`.
3. You will find three files `CRBalancer.java`, `CRBalancingPolicy.java` and `CRNamenodeConnector.java`.
4. `CRBalancer.java` is the main program which makes the decision to transfer data among nodes based on computing power.
5. `CRBalancingPolicy` calculates and stores the space occupied by each node. It is used to know the total space occupied by a node currently. It aids in making the decision.
6. `NamenodeConnector` is used for connecting to the Namenode in order to get the information about the datanodes.
7. In every Program, I have mentioned the place where code has been changed from the original balancer which balances the node based on space utilization instead of computing utilization.
8. This program is written with an assumption that there is sufficient space on faster computing nodes to move the data from slower computing nodes.