NameNodeFailover

Outdated

The information on this wiki page is outdated and will be deleted soon. NameNode High-Availability is present in 2.x.

Introduction

As of 0.20, Hadoop does not support automatic recovery in the case of a NameNode failure. This is a well known and recognized single point of failure in Hadoop.

Experience at Yahoo! shows that NameNodes are more likely to fail due to misconfiguration, network issues, and bad behavior amongst clients than actual hardware problems. Out of fifteen grids over three year period, only three NameNode failures were related to hardware problems.

Configuring Hadoop for Failover

There are some preliminary steps that must be in place prior to performing a NameNode recovery. The most important is the dfs.name.dir property. This setting configures the NameNode such that it can write to more than one directory. A typcal configuration might look something like this:

```
<property>
<name>dfs.name.dir</name>
<value>/export/hadoop/namedir,/remote/export/hadoop/namedir</value>
</property>
```

The first directory is a local directory and the second directory is a NFS mounted directory. The NameNode will write to both locations, keeping the HDFS metadata in sync. This allows for storage of the metadata off-machine so that one will have something to recover. During startup, the NameNode will pick the most recent version of these two directories to use and then sync both of them to use the same data.

After we have configured the NameNode to write to two or more directories, we now have a working backup of the metadata. Using this data, in the more common failure scenarios, we can use this data to bring the dead NameNode from the grave.

When a Failure Occurs

Now the recovery steps:

1. Just to be safe, make a copy of the data on the remote NFS mount for safe keeping.
2. Pick a target machine on the same network.
3. Change the IP address of that machine to match the NameNode's IP address. Using an interface alias to provide this address movement works as well. If this is not an option, be prepared to restart the entire grid to avoid hitting https://issues.apache.org/jira/browse/HADOOP-3988 .
4. Install Hadoop similarly to how you did the NameNode
5. Do **not** format this node!
6. Mount the remote NFS directory in the same location.
7. Startup the NameNode.
8. The NameNode should start replaying the edits file, updating the image, block reports should come in, etc.

At this point, your NameNode should be up.

Other Ideas

There are some other ideas to help with NameNode recovery:

1. Keep in mind that the SecondaryNameNode and/or the CheckpointNode also has an older copy of the NameNode metadata. If you haven't done the preliminary work above, you might still be able to recover using the data on those systems. Just note that it will only be as fresh as the last run and you will likely experience some data loss.
2. Instead of using NFS on Linux, it may be worth while looking into DRBD. A few sites are using this with great success.