PseudoDistributedHadoop

Pseudo Distributed Hadoop is where Hadoop runs as set of independent JVMs, but only on a single host. It has much lower performance than a real Hadoop cluster, due to the smaller number of hard disks limiting IO bandwidth. It is, however, a good way to play with new MR algorithms on very small datasets, and to learn how to use Hadoop. Developers working in the Hadoop codebase usually test their code in this mode before deploying their build of Hadoop to a local test cluster.

If you are running in this mode (and don't have a proxy server fielding HTML requests), and have not changed the default port values, then both the NameNode and JobTracker can be reached from this page

Ports in Use

These are the standard ports; if the configuration files are changed then they will not be valid.

- NameNode: http://localhost:50070/dfshealth.jsp
- HDFS filesystem browser http://localhost:50075/browseDirectory.jsp?namenodeInfoPort=50070&dir=/
- Server logs http://localhost:50070/logs/
- JobTracker: http://localhost:50030/jobtracker.jsp
- TaskTracker: http://localhost:50060/tasktracker.jsp

Recommended Configuration Parameters for Pseudo-Distributed Hadoop

With only a single HDFS datanode, the replication factor should be set to 1 the same goes for the replication factor of submitted jars. You also need to tell the Job tracker to not try handing a failing task to another task tracker, or to blacklist a tracker that appears to fail a lot. While those options are essential in large clusters with many machines -some of which will start to fail, on a single node cluster they do more harm than good.

```
mapred.submit.replication=1
mapred.skip.attempts.to.start.skipping=1
mapred.max.tracker.failures=10000
mapred.max.tracker.blacklists=10000
mapred.map.tasks.speculative.execution=false
mapred.reduce.tasks.speculative.execution=false
tasktracker.http.threads=5
```