HDFS Administration

… for those ‘whack-a-mole’ days …

(a very informal look at administering the HDFS filesystem)

Garhan Attebury
attebury@cse.unl.edu

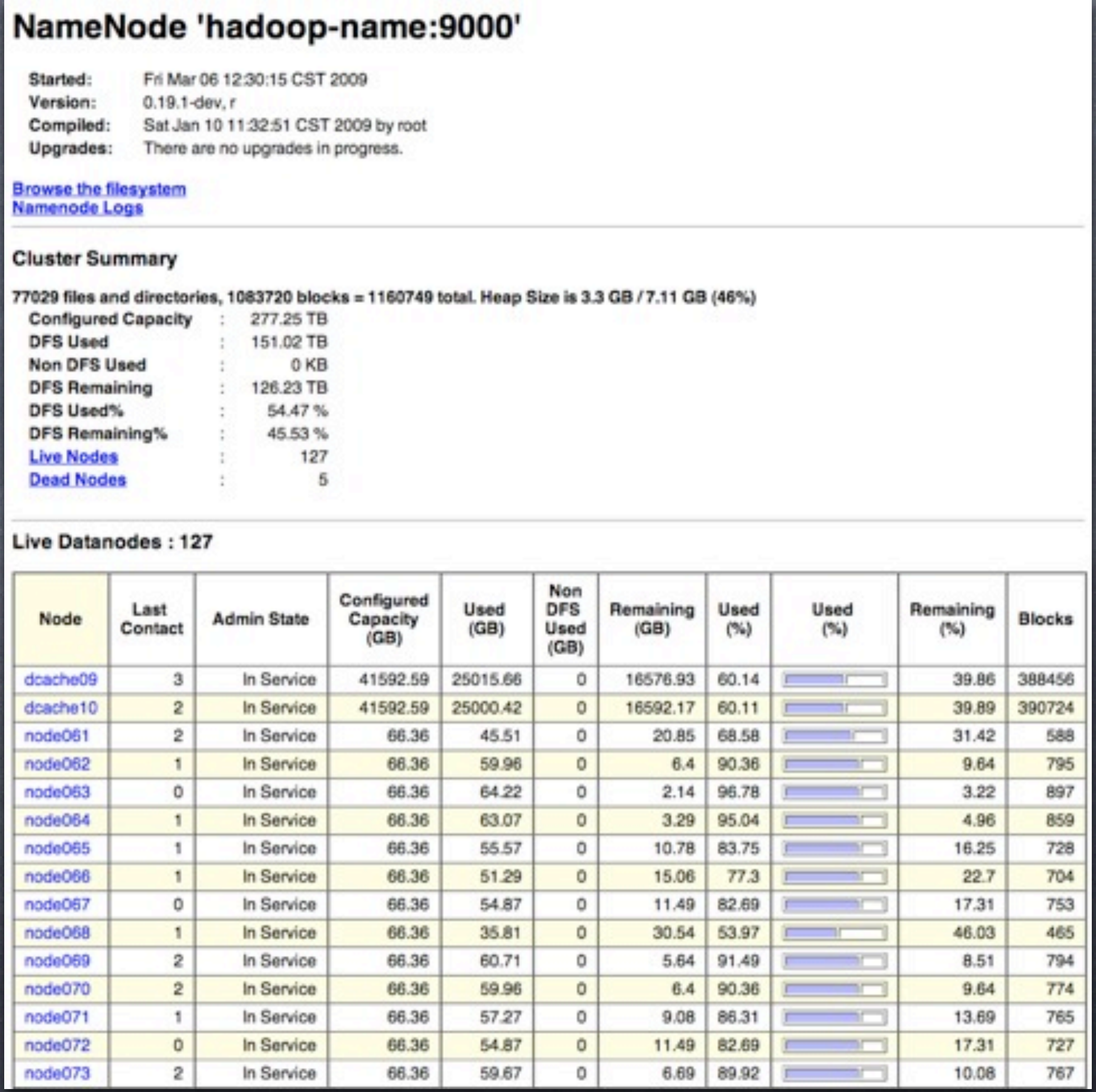
Hadoop Workshop @ UCSD
March 11-13, 2009

HDFS Administration

- Web Interfaces
- Command Line Tools
- Ganglia Integration
- Nagios Probes
- JMX Monitoring

Web Interfaces

- namenode:50070
- datanode:50075
- jobtracker:50030
- tasktracker:50060
- … might want to firewall those …



NameNode 'hadoop-name:9000'

Started:	Fri Mar 06 12:30:15 CST 2009
Version:	0.19.1-dev, r
Compiled:	Sat Jan 10 11:32:51 CST 2009 by root
Upgrades:	There are no upgrades in progress.

Browse the filesystem
Namenode Logs

Cluster Summary

77029 files and directories, 1083720 blocks = 1160749 total. Heap Size is 3.3 GB / 7.11 GB (46%)

Configured Capacity	:	277.25 TB
DFS Used	:	151.02 TB
Non DFS Used	:	0 KB
DFS Remaining	:	126.23 TB
DFS Used%	:	54.47 %
DFS Remaining%	:	45.53 %
Live Nodes	:	127
Dead Nodes	:	5

Live Datanodes : 127

Node	Last Contact	Admin State	Configured Capacity (GB)	Used (GB)	Non DFS Used (GB)	Remaining (GB)	Used (%)	Used (%)	Remaining (%)	Blocks
dcache09	3	In Service	41592.59	25015.66	0	16576.93	60.14	▬	39.86	388456
dcache10	2	In Service	41592.59	25000.42	0	16592.17	60.11	▬	39.89	390724
node061	2	In Service	66.36	45.51	0	20.85	68.58	▬	31.42	588
node062	1	In Service	66.36	59.96	0	6.4	90.36	▬	9.64	795
node063	0	In Service	66.36	64.22	0	2.14	96.78	▬	3.22	897
node064	1	In Service	66.36	63.07	0	3.29	95.04	▬	4.96	859
node065	1	In Service	66.36	55.57	0	10.78	83.75	▬	16.25	728
node066	1	In Service	66.36	51.29	0	15.06	77.3	▬	22.7	704
node067	0	In Service	66.36	54.87	0	11.49	82.69	▬	17.31	753
node068	1	In Service	66.36	35.81	0	30.54	53.97	▬	46.03	465
node069	2	In Service	66.36	60.71	0	5.64	91.49	▬	8.51	794
node070	2	In Service	66.36	59.96	0	6.4	90.36	▬	9.64	774
node071	1	In Service	66.36	57.27	0	9.08	86.31	▬	13.69	765
node072	0	In Service	66.36	54.87	0	11.49	82.69	▬	17.31	727
node073	2	In Service	66.36	59.67	0	6.69	89.92	▬	10.08	767

Command Line Tools

- start-all.sh, -dfs.sh, -mapred.sh, -balancer.sh
  (I never use these)
- hadoop-daemons.sh -> hadoop-daemon.sh

  (I use the later constantly)

  ```
  hadoop-daemon.sh --config $HADOOP_HOME/conf/ start datanode
  ```
- hadoop (who would have guessed?)
  - dfsadmin, fsck, -fs

- hadoop fsck

```
[root@hadoop-name ~]# hadoop fsck
Usage: DFSck <path> [-move | -delete | -openforwrite | -deep ] [-
files [-blocks [-locations | -racks]]]
 <path>   start checking from this path
 -move move corrupted files to /lost+found
 -delete  delete corrupted files
 -files   print out files being checked
 -openforwrite   print out files opened for write
 -blocks  print out block report
 -locations   print out locations for every block
 -racks   print out network topology for data-node locations
 -deep deeply check blocks on datanodes
```

```
[root@hadoop-name ~]# hadoop fsck /user/uscms01/pnfs/unl.edu/data4/cms/store/CSA07/2007/11/21/CSA07-CSA07Muon-Tier0-A1-Chowder/0026/92611921-
F49A-DC11-9D74-000423D6B328.root -files -locations -blocks
/user/uscms01/pnfs/unl.edu/data4/cms/store/CSA07/2007/11/21/CSA07-CSA07Muon-Tier0-A1-Chowder/0026/92611921-F49A-DC11-9D74-000423D6B328.root
1394786863 bytes, 21 block(s):  Under replicated blk_-5059841161638826489_60320. Target Replicas is 3 but found 2 replica(s).
0. blk_2870865034428690739_60318 len=67108864 repl=3 [172.16.1.140:50010, 172.16.1.138:50010, 172.16.1.123:50010]
1. blk_-7699277013711289197_60319 len=67108864 repl=3 [172.16.1.121:50010, 172.16.1.133:50010, 172.16.1.182:50010]
2. blk_-5059841161638826489_60320 len=67108864 repl=2 [172.16.1.68:50010, 172.16.1.171:50010]
3. blk_3466707575679750081_60320 len=67108864 repl=3 [172.16.1.186:50010, 172.16.1.152:50010, 172.16.1.119:50010]
4. blk_6087943856764339232_60320 len=67108864 repl=3 [172.16.1.125:50010, 172.16.1.126:50010, 172.16.1.182:50010]
5. blk_1181597584774856895_60320 len=67108864 repl=3 [172.16.1.120:50010, 172.16.1.152:50010, 172.16.1.184:50010]
6. blk_5595521708362182008_60320 len=67108864 repl=3 [172.16.1.161:50010, 172.16.1.123:50010, 172.16.1.171:50010]
7. blk_-3979303076544055_60320 len=67108864 repl=3 [172.16.1.68:50010, 172.16.1.138:50010, 172.16.1.165:50010]
8. blk_6084894110716314730_60321 len=67108864 repl=3 [172.16.1.149:50010, 172.16.1.116:50010, 172.16.1.187:50010]
9. blk_-3363090076878604174_60321 len=67108864 repl=3 [172.16.1.144:50010, 172.16.1.127:50010, 172.16.1.129:50010]
10. blk_-2961655684834144459_60321 len=67108864 repl=3 [172.16.1.188:50010, 172.16.1.138:50010, 172.16.1.172:50010]
11. blk_3724579871753852749_60321 len=67108864 repl=3 [172.16.1.136:50010, 172.16.1.115:50010, 172.16.1.191:50010]
12. blk_8839901480367319666_60321 len=67108864 repl=3 [172.16.1.160:50010, 172.16.1.186:50010, 172.16.1.181:50010]
13. blk_3915691073443560566_60322 len=67108864 repl=3 [172.16.1.190:50010, 172.16.1.125:50010, 172.16.1.115:50010]
14. blk_2110166770791579689_60322 len=67108864 repl=3 [172.16.1.123:50010, 172.16.1.116:50010, 172.16.1.182:50010]
15. blk_5130378967930757320_60322 len=67108864 repl=3 [172.16.1.132:50010, 172.16.1.172:50010, 172.16.1.191:50010]
16. blk_-4823106933471814329_60322 len=67108864 repl=3 [172.16.1.142:50010, 172.16.1.122:50010, 172.16.1.115:50010]
17. blk_1041032237295633398_60322 len=67108864 repl=3 [172.16.1.148:50010, 172.16.1.124:50010, 172.16.1.152:50010]
18. blk_5745271099540660127_60322 len=67108864 repl=3 [172.16.1.170:50010, 172.16.1.169:50010, 172.16.1.115:50010]
19. blk_3200399532031889146_60322 len=67108864 repl=3 [172.16.1.160:50010, 172.16.1.131:50010, 172.16.1.178:50010]
20. blk_-943023194480050943_60322 len=52609583 repl=3 [172.16.1.160:50010, 172.16.1.177:50010, 172.16.1.183:50010]

Status: HEALTHY
 Total size: 1394786863 B
 Total dirs: 0
 Total files:      1
 Total blocks (validated):      21 (avg. block size 66418422 B)
 Minimally replicated blocks:   21 (100.0 %)
 Over-replicated blocks:  0 (0.0 %)
 Under-replicated blocks: 1 (4.7619047 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    3
 Average block replication:     2.952381
 Corrupt blocks:         0
 Missing replicas:       1 (1.6129032 %)
 Number of data-nodes:          130
 Number of racks:        1


The filesystem under path '/user/uscms01/pnfs/unl.edu/data4/cms/store/CSA07/2007/11/21/CSA07-CSA07Muon-Tier0-A1-Chowder/0026/92611921-F49A-
DC11-9D74-000423D6B328.root' is HEALTHY
```

```
[root@hadoop-name ~]# hadoop fsck /user/uscms01/pnfs/unl.edu/data4/cms/store/CSA07/2007/11/21/CSA07-CSA07Muon-Tier0-A1-Chowder/0026/92611921-F49A-DC11-9D74-000423D6B328.root -files -locations -blocks
/user/uscms01/pnfs/unl.edu/data4/cms/store/CSA07/2007/11/21/CSA07-CSA07Muon-Tier0-A1-Chowder/0026/92611921-F49A-DC11-9D74-000423D6B328.root
1394786863 bytes, 21 block(s):  Under replicated blk_-5059841161638826489_60320. Target Replicas is 3 but found 2 replica(s).
0. blk_2870865034428690739_60318 len=67108864 repl=3 [172.16.1.140:50010, 172.16.1.138:50010, 172.16.1.123:50010]
1. blk_-7699277013711289197_60319 len=67108864 repl=3 [172.16.1.121:50010, 172.16.1.133:50010, 172.16.1.182:50010]
2. blk_-5059841161638826489_60320 len=67108864 repl=2 [172.16.1.68:50010, 172.16.1.171:50010]
3. blk_3466707575679750081_60320 len=67108864 repl=3 [172.16.1.186:50010, 172.16.1.152:50010, 172.16.1.119:50010]
4. blk_6087943856764339232_60320 len=67108864 repl=3 [172.16.1.125:50010, 172.16.1.126:50010, 172.16.1.182:50010]
5. blk_1181597584774856895_60320 len=67108864 repl=3 [172.16.1.120:50010, 172.16.1.152:50010, 172.16.1.184:50010]
6. blk_5595521708362182008_60320 len=67108864 repl=3 [172.16.1.161:50010, 172.16.1.123:50010, 172.16.1.171:50010]
7. blk_-3979303076544055_60320 len=67108864 repl=3 [172.16.1.68:50010, 172.16.1.138:50010, 172.16.1.165:50010]
8. blk_6084894110716314730_60321 len=67108864 repl=3 [172.16.1.149:50010, 172.16.1.116:50010, 172.16.1.187:50010]
9. blk_-3363090076878604174_60321 len=67108864 repl=3 [172.16.1.144:50010, 172.16.1.127:50010, 172.16.1.129:50010]
10. blk_-2961655684834144459_60321 len=67108864 repl=3 [172.16.1.188:50010, 172.16.1.138:50010, 172.16.1.172:50010]
11. blk_3724579871753852749_60321 len=67108864 repl=3 [172.16.1.136:50010, 172.16.1.115:50010, 172.16.1.191:50010]
12. blk_8839901480367319666_60321 len=67108864 repl=3 [172.16.1.160:50010, 172.16.1.186:50010, 172.16.1.181:50010]
13. blk_3915691073443560566_60322 len=67108864 repl=3 [172.16.1.190:50010, 172.16.1.125:50010, 172.16.1.115:50010]
14. blk_2110166770791579689_60322 len=67108864 repl=3 [172.16.1.123:50010, 172.16.1.116:50010, 172.16.1.182:50010]
15. blk_5130378967930757320_60322 len=67108864 repl=3 [172.16.1.132:50010, 172.16.1.172:50010, 172.16.1.191:50010]
16. blk_-4823106933471814329_60322 len=67108864 repl=3 [172.16.1.142:50010, 172.16.1.122:50010, 172.16.1.115:50010]
17. blk_1041032237295633398_60322 len=67108864 repl=3 [172.16.1.148:50010, 172.16.1.124:50010, 172.16.1.152:50010]
18. blk_5745271099540660127_60322 len=67108864 repl=3 [172.16.1.170:50010, 172.16.1.169:50010, 172.16.1.115:50010]
19. blk_3200399532031889146_60322 len=67108864 repl=3 [172.16.1.160:50010, 172.16.1.131:50010, 172.16.1.178:50010]
20. blk_-943023194480050943_60322 len=52609583 repl=3 [172.16.1.160:50010, 172.16.1.177:50010, 172.16.1.183:50010]

Status: HEA
 Total size
 Total dirs
 Total file
 Total bloc
 Minimally
 Over-repli
 Under-replicated blocks: 1 (4.7619047 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    3
 Average block replication:     2.952381
 Corrupt blocks:         0
 Missing replicas:       1 (1.6129032 %)
 Number of data-nodes:          130
 Number of racks:        1


The filesystem under path '/user/uscms01/pnfs/unl.edu/data4/cms/store/CSA07/2007/11/21/CSA07-CSA07Muon-Tier0-A1-Chowder/0026/92611921-F49A-DC11-9D74-000423D6B328.root' is HEALTHY
```

```
[root@node123 current]# find . -name *blk_2110166770791579689*
./subdir7/subdir38/subdir7/blk_2110166770791579689_60322.meta
./subdir7/subdir38/subdir7/blk_2110166770791579689

[root@node123 current]# ls -l ./subdir7/subdir38/subdir7/blk_2110166770791579689*
-rw-r--r--  1 root root 67108864 Dec 15 09:33 ./subdir7/subdir38/subdir7/blk_2110166770791579689
-rw-r--r--  1 root root   524295 Dec 15 09:33 ./subdir7/subdir38/subdir7/blk_2110166770791579689_60322.meta
```

- hfscker: http://cse.unl.edu/~attebury/hfscker
    - really basic ‘hadoop fsck’ parser
    - gives quick summary and provides -u (under replicated), -c (corrupt), and -m (missing) ouput
    - hfscker -cm | sort | uniq > retransfer.out

-
  ```
  [root@hadoop-name ~]# hfscker
  Total Problems: 0
  Problem Files: 0
     Corrupt: 0
     Missing: 0
     Under Replicated: 0
  ```

- = go back to sleep

- ok, not -that- simple, historical view from other monitoring / namenode page also

- Decommissioning a node

```
echo “node055” >> /scratch/hadoop/hosts_exclude
hadoop dfsadmin -refreshNodes
```

  - Web interface will show the node as “decommissioning” while replicating
  - Shows “decommissioned” very briefly, then appears in the ‘dead node’ list
  - Must remove from ‘hosts_exclude’ before starting datanode again

- Nagios Probes (at Nebraska)

```
command[check_fuse_dfs_count]=/usr/lib/nagios/plugins/check_procs -w
1:1 -c 0:2 -C fuse_dfs

command[check_hadoop_namenode]=/usr/lib/nagios/plugins/check_procs -w
1:1 -c 0:2 -a "org.apache.hadoop.hdfs.server.namenode.NameNode"

command[check_hadoop_jobtracker]=/usr/lib/nagios/plugins/check_procs
-w 1:1 -c 0:2 -a "org.apache.hadoop.mapred.JobTracker"

command[check_hadoop_tasktracker]=/usr/lib/nagios/plugins/check_procs
-w 1:1 -c 0:2 -a "org.apache.hadoop.mapred.TaskTracker"

#command[check_hadoop_datanode]=/usr/lib/nagios/plugins/check_procs -
w 1:1 -c 0:2 -a "org.apache.hadoop.hdfs.server.datanode.DataNode"

command[check_hadoop_datanode]=/usr/bin/sudo /usr/lib/nagios/plugins/
check_dual_datanode

command[check_red_mounts]=/usr/bin/sudo /usr/lib/nagios/plugins/
check_red_mounts
```

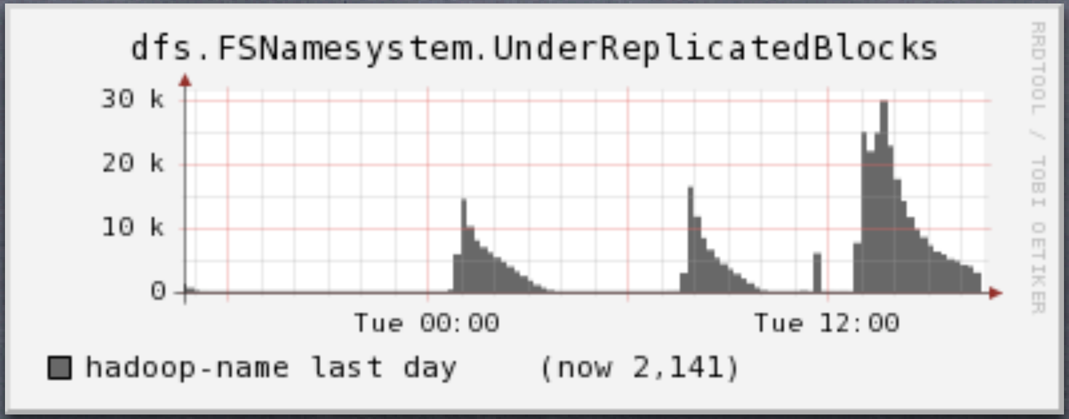
Graphing options

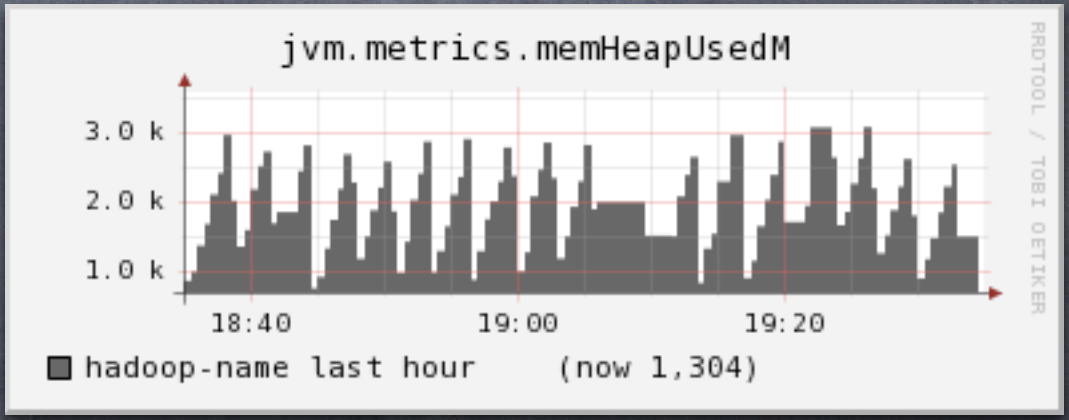
- Namenode / Jobtracker / Datanode all have Ganglia support built in

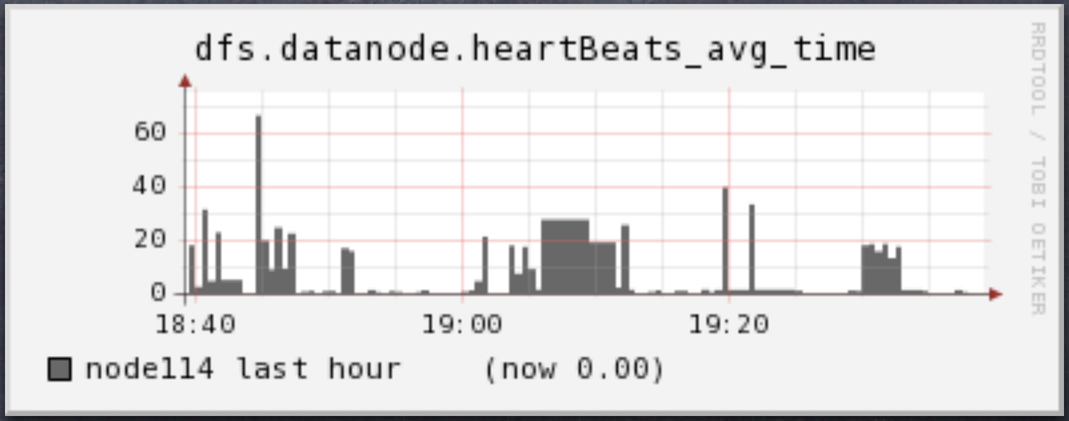
- hadoop-metrics.properties

```
# Configuration of the "dfs"
dfs.class=org.apache.hadoop.metrics.ga
nglia.GangliaContext31
dfs.period=10
dfs.servers=239.2.11.152:8649
```

- Really just JVM monitoring (JMX->Ganglia, JMX->SNMP, JMX->Whatever-you-want)







JMX Monitoring

- jconsole (GC fun)
- JVM -> Ganglia metrics is included
- JVM -> SNMP metrics (ask UCSD :)

- Daily hadoop tasks
    - hfscker (hadoop fsck)
    - glance at namenode webpage
    - pay attention to nagios (under replication or corrupt blocks a good metric)

- tweaks
  - /etc/security/limits.conf (open file limit)
  - ...

- Not perfect - annoying/strange things
  - balancer sometimes gets ‘stuck’
  - datanodes sometime get ‘stuck’
    - 2x DataNode processes, killing the bad one fixes it - no loss/corruption yet
  - node162, you -ARE- the weakest link!

- Random things
- http://hadoop.apache.org/core/mailing_lists.html "hadoop-core-user"
- Watch your log sizes -- we've got everything on, but they get BIG fast
- Disk layouts -- whatever you like really