Apache Hadoop for System Administrators
Allen Wittenauer
Twitter: @_a___w_
Email: aw@apache.org
Hadoop Deployed Now?
Planning Hadoop Deployment?
Needed some place to sit before lunch?
An Extremely Quick & Incomplete Intro to Hadoop
Map (“transform”)

- Perl:
  ```perl
  @items=(1,2,3,4,5);
  sub sqr {return $_**2);
  print join(',,',map(sqr,@items));
  1,4,9,16,25
  ```

- Python:
  ```python
  items = [1,2,3,4,5]
  def sqr(x) : return x**2

  print list(map(sqr,items))
  [1, 4, 9, 16, 25]
  ```
Reduce ("compress" or "fold")

- Perl

```perl
use List::Util qw/reduce/;
@items=(1,4,9,16,25);
print reduce { $a>$b ? $a:$b } @items;
25
```

- Python

```python
from functools import reduce
items = [1,4,9,16,25]
print reduce ((lambda x,y: x if (x>y) else y), items)
25
```
NEVER                  GONNA

GIVE                           YOU UP
Hadoop
(‘common’ or ‘core’)

MapReduce

HDFS
Hadoop
(‘common’ or ‘core’)

MapReduce

S3
Hadoop
('common' or 'core')

MapReduce

Gluster
Hadoop
('common' or 'core')

HBase

HDFS
Hadoop isn’t designed for system administrators and/or support staff.
“Hadoop is not a developer problem; it’s an operations problem.”
-- Hadoop vendor ex-employee
That's a rock!
Don’t Make Assumptions
tail’ing the logs won’t tell you the whole story.
Monitor the masters!
### Host Status Totals

<table>
<thead>
<tr>
<th></th>
<th>Up</th>
<th>Down</th>
<th>Unreachable</th>
<th>Pending</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### All Problems

- 0

### All Types

- 54

## Service Status Details For All Hosts

<table>
<thead>
<tr>
<th>Service</th>
<th>Status</th>
<th>Last Check</th>
<th>Duration</th>
<th>Attempt</th>
<th>Status Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canasta AZKABAN _v2 Port Check</td>
<td>OK</td>
<td>02-24-2013 20:20:11</td>
<td>11d 3h 7m 50s</td>
<td>1/3</td>
<td>OK: Port 8443</td>
</tr>
<tr>
<td>Disk Space Check - /export</td>
<td>OK</td>
<td>02-24-2013 20:20:55</td>
<td>11d 3h 7m 6s</td>
<td>1/3</td>
<td>DISK OK - free space: /export</td>
</tr>
<tr>
<td>Disk Space Check - /export/home</td>
<td>OK</td>
<td>02-24-2013 20:20:50</td>
<td>11d 3h 7m 11s</td>
<td>1/3</td>
<td>DISK OK - free space: /export/home</td>
</tr>
<tr>
<td>Disk Space Check - /grid/azlogs</td>
<td>OK</td>
<td>02-24-2013 20:20:16</td>
<td>11d 3h 7m 45s</td>
<td>1/3</td>
<td>DISK OK - free space: /grid 533</td>
</tr>
<tr>
<td>Disk Space Check - /export/home</td>
<td>OK</td>
<td>02-24-2013 20:22:51</td>
<td>11d 3h 6m 32s</td>
<td>1/3</td>
<td>DISK OK - free space: /export</td>
</tr>
<tr>
<td>Disk Space Check - /grid</td>
<td>OK</td>
<td>02-24-2013 20:21:43</td>
<td>11d 3h 6m 18s</td>
<td>1/3</td>
<td>DISK OK - free space: /grid 954</td>
</tr>
<tr>
<td>SSH Check</td>
<td>OK</td>
<td>02-24-2013 20:19:06</td>
<td>11d 3h 6m 32s</td>
<td>1/3</td>
<td>SSH OK - OpenSSH_5.3 (proto)</td>
</tr>
<tr>
<td>Blacklisted Nodes</td>
<td>OK</td>
<td>02-24-2013 20:22:31</td>
<td>4d 15h 35m 30s</td>
<td>1/3</td>
<td>OK: Blacklisted Nodes Count: 0</td>
</tr>
<tr>
<td>CPU Load Check</td>
<td>OK</td>
<td>02-24-2013 20:22:22</td>
<td>11d 3h 5m 39s</td>
<td>1/3</td>
<td>OK - load average: 2.25, 2.37, 2</td>
</tr>
<tr>
<td>Disk Space Check - /export/home</td>
<td>OK</td>
<td>02-24-2013 20:22:49</td>
<td>11d 3h 5m 12s</td>
<td>1/3</td>
<td>DISK OK - free space: /export</td>
</tr>
<tr>
<td>Disk Space Check - /grid</td>
<td>OK</td>
<td>02-24-2013 20:22:47</td>
<td>11d 3h 5m 14s</td>
<td>1/3</td>
<td>DISK OK - free space: /grid 954</td>
</tr>
<tr>
<td>Disk Space Check - /grid/hadoop/ogs</td>
<td>OK</td>
<td>02-24-2013 20:22:54</td>
<td>11d 3h 6m 24s</td>
<td>1/3</td>
<td>DISK OK - free space: /grid/hadoop</td>
</tr>
<tr>
<td>JT-port-50030</td>
<td>OK</td>
<td>02-24-2013 20:19:51</td>
<td>11d 3h 8m 11s</td>
<td>1/3</td>
<td>OK: Port 50030</td>
</tr>
<tr>
<td>JobTracker Process</td>
<td>OK</td>
<td>02-24-2013 20:21:24</td>
<td>11d 3h 6m 37s</td>
<td>1/3</td>
<td>PROCES OK: 1 process with cor</td>
</tr>
<tr>
<td>Lost Tracker Check</td>
<td>OK</td>
<td>02-24-2013 20:22:20</td>
<td>0d 21h 0m 41s</td>
<td>1/3</td>
<td>OK: No lost trackers</td>
</tr>
<tr>
<td>jobSubmitCheckDefault</td>
<td>OK</td>
<td>02-24-2013 20:20:42</td>
<td>11d 2h 31m 31s</td>
<td>1/1</td>
<td>OK: Expected result: Sum = 18</td>
</tr>
</tbody>
</table>
- LinkedIn’s Configuration
  - 30+ Health Checks per Grid
    - Masters, canary report, daily fsck, etc
  - 10+ Health Checks per DC
    - LDAP, Kerberos, etc ...
  - Cross-DC Nagios Server Checks

- Warn: 5% down nodes
- Panic: 30% down
- HDFS: 20% Free Space
- Gateway home dir: 10% free space
- ...

Nagios

ZK  VD  NN  JT  AZ  GW

Compute Nodes
- **Health Check Script**
  - “OK” - good status
  - “ERROR (message)” - bad status

  ```bash
  mapred.healthChecker.script.path
  ```

- **Consider checking ...**
  - critical software
  - ownership & permissions
  - network connection speed
  - drive count
  - file system space
  - RO file systems
  - IO errors
  - missing memory
- Use the tools most of your user’s code is written in!

- Pig
  - testfile:
    
    100
  
  - Code:
    
    A = load 'testfile' using PigStorage(',,')
    as (i: int);
    B = foreach C generate i;
    C = distinct B;
    dump C;

  - Output:
    
    (100)
Reactive
Proactive
Resource Controls
▪ JobTracker Memory Resource Controls
  – Limit jobs stored in JT heap:
    mapred.jobtracker.completeuserjobs.maximum
  – Limit total # of job tasks: mapred.jobtracker.maxtasks.per.job

▪ Job Memory Resource Controls
  – Scheduler-level: mapred.cluster.*.memory.mb
  – TT-level: auto-calculated based upon MR slot counts & scheduler level settings
  – MR Job-level: mapred.job.*.memory.mb
  – Linux only: /proc memory calculator and task killer
“I set the heap to 1G but my process ran out of memory?”
Treat HDFS like any other multi-tenant FS
- **Quota everything**
  - Yes, including /tmp
  - No “show me all quotas” functionality

- **Be consistent:**
  - /user/* all get same quota

- **Be flexible:**
  - Make another dir for user’s to store big projects (e.g., /project)

- **Be smart:**
  - Have a policy that content in /tmp gets deleted after X days. Automate this!
  - Build reporting that shows files that are replicated less than 3 times

dfsadmin -setQuota
dfsadmin -setSpaceQuota
Compute Node Disk Partitioning as Protective Measure
- root partitioning

- non-root partitioning
Security!
- Queue Level ACLs
  - users
  - groups
  - netgroups

- Service Level ACLs
  - hosts
  - users
  - groups
  - netgroups

- Limitation: Web services are all or nothing! :( 
- Be aware: Hadoop uses ephemeral ports all over the place! :(
Kerberos!
Corp IT
Active Directory
@CORP

Client Node

Password

krbtgt/GRID@CORP

Grid Realm
@GRID

krbtgt/host@GRID
krbtgt/service@GRID

Hadoop Services

krbtgt/GRID@CORP
krbtgt/user@CORP
Account Management

Help

This tab is for managing your account on the development Hadoop grids at LinkedIn. If you don't already have an account, click the "Activate" button under status of eat1-magic. It can take up to 72 hours for your account request to be processed, if it takes longer, file a JIRA ticket.

My Account

Login Name: awittena
Real Name: Allen Wittenauer
Unix Home Dir: /export/home/awittena
Unix Shell: /bin/bash

Headless Accounts

<table>
<thead>
<tr>
<th>Grid</th>
<th>Type</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>eat1-magic</td>
<td>Development</td>
<td>✔️ Active</td>
</tr>
</tbody>
</table>
- Fonzi: http://www.flickr.com/photos/elzey/7224689810
- Ant on flower: http://www.flickr.com/photos/bolonski/6116358907
- Ant Colony: http://www.flickr.com/photos/klearchos/2821230516
- Ant Queen: http://commons.wikimedia.org/wiki/File:Camponotus_crispulus_queen_ant.jpg
- Canary: http://www.flickr.com/photos/nathan_and_jenny/2454127424
- Mona Lisa: Leonardo Da Vinci
- White Elephant: http://data.linkedin.com/opensource/white-elephant
- Ecce Homo:
  - Elías García Martínez (original)
  - Cecilia Giménez (restored)
Thanks!

Contact:
Twitter: @a__w_
Email: aw@apache.org

More info:
Quora: www.quora.com/user/allenwittenauer
SlideShare: www.slideshare.net/allenwittenauer