Apache Hadoop for System Administrators

Allen Wittenauer



Allen Wittenauer

–

San Francisco Bay Area | Computer Software

Current	The Apache Software Foundation
Previous	LinkedIn, Yahoo!, SearchMe, Inc.
Education	Southern Illinois University, Carbondale

Edit

455
connections

www.linkedin.com/in/allenwittenauer/

Contact Info

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:

```
@items=(1,2,3,4,5);
sub sqr {return $_**2);
print join(‘,’,map(sqr,@items));
1,4,9,16,25
```

  - 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

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

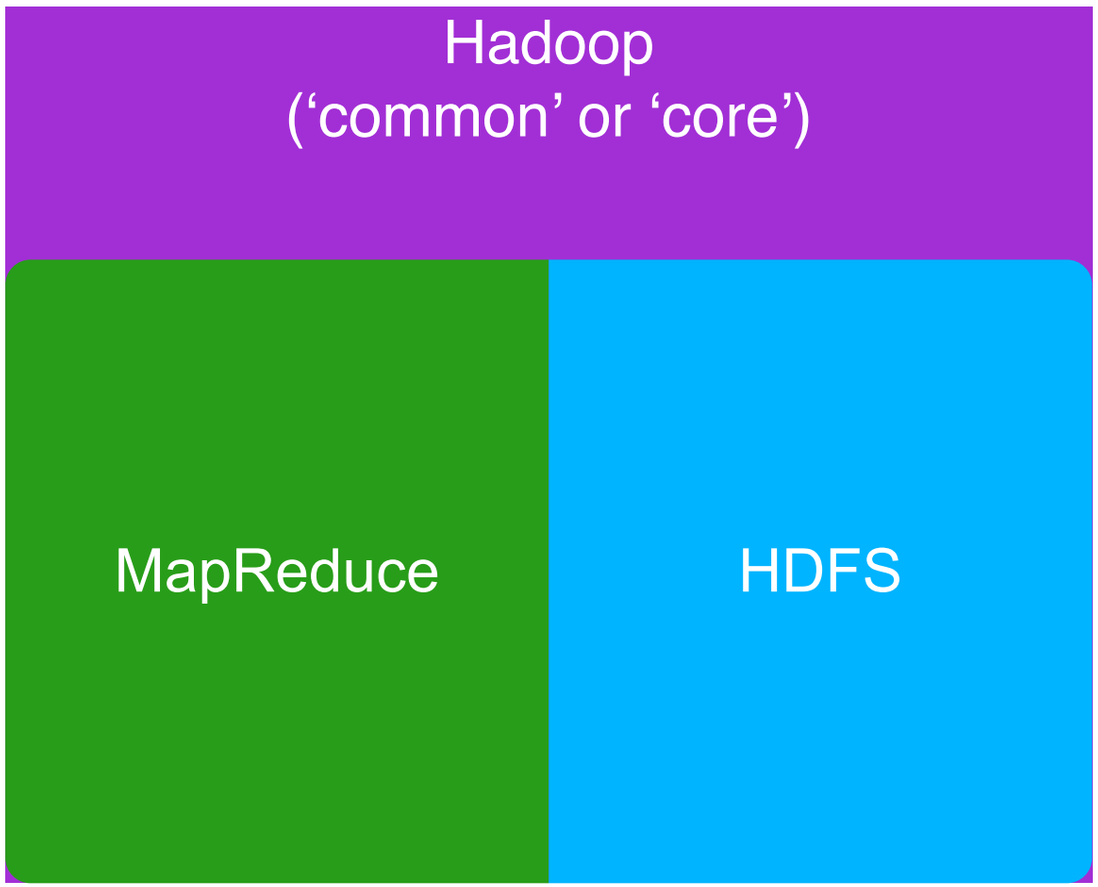
  - Python

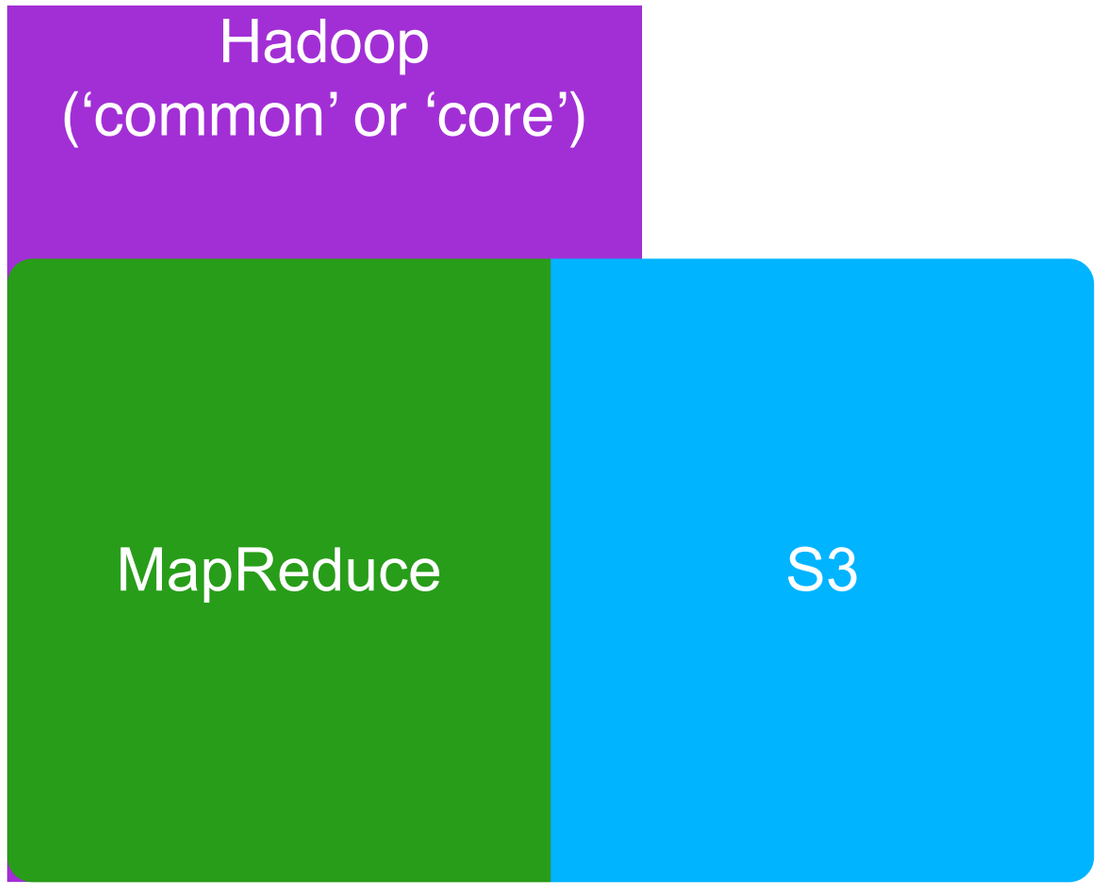
```
from functools import reduce
items = [1,4,9,16,25]
print reduce ((lambda x,y: x if
   (x>y) else y), items)
25
```

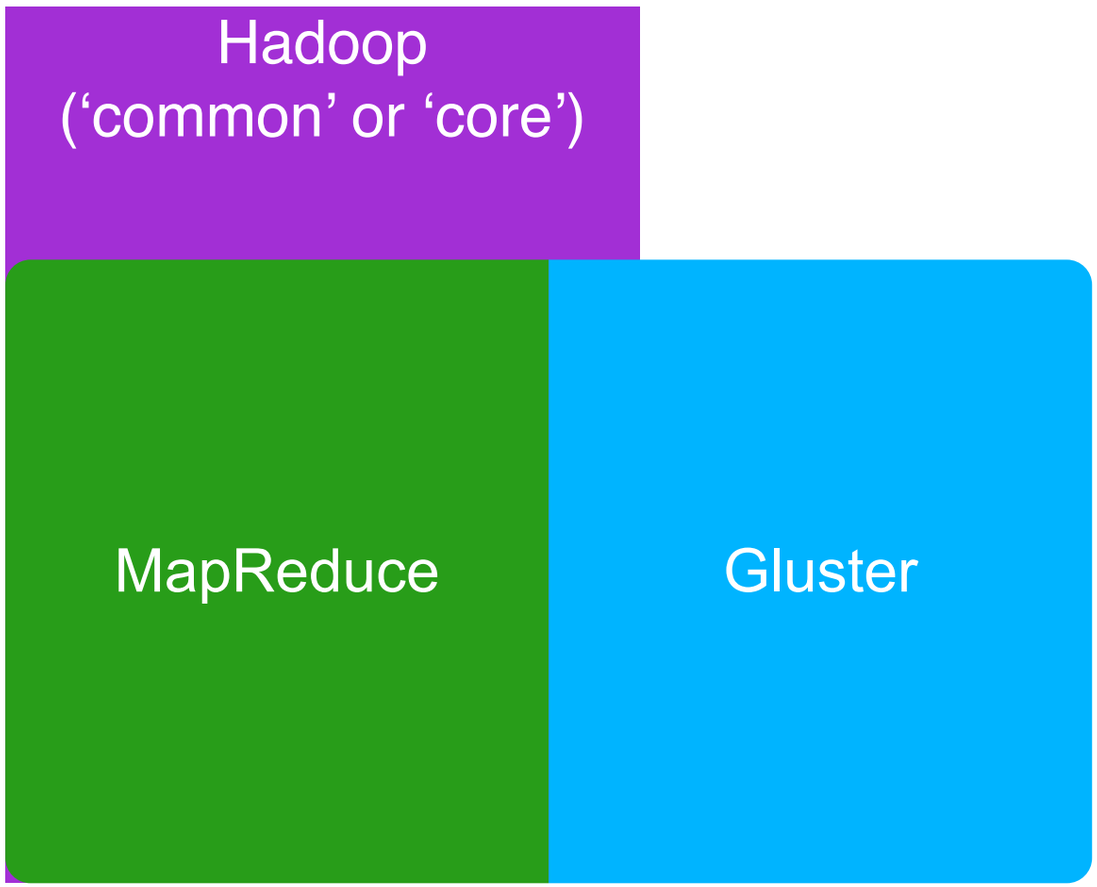
NEVER GONNA

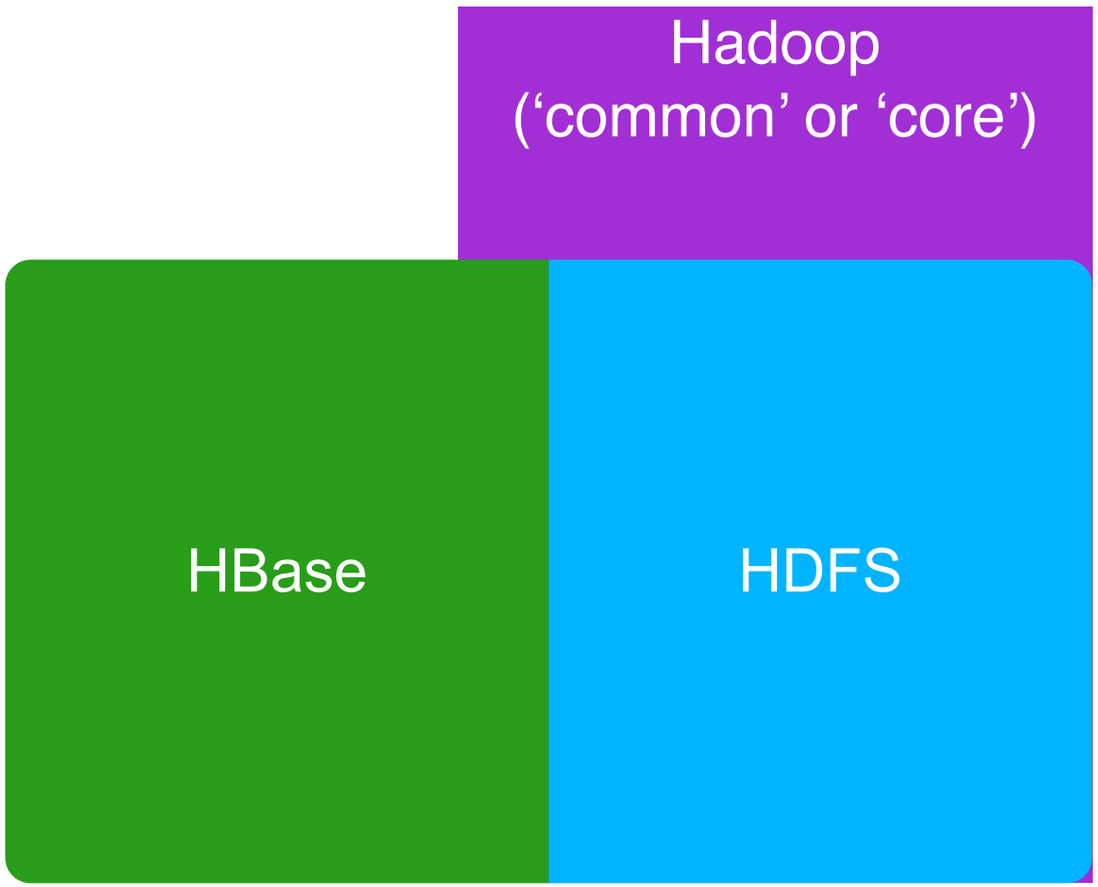


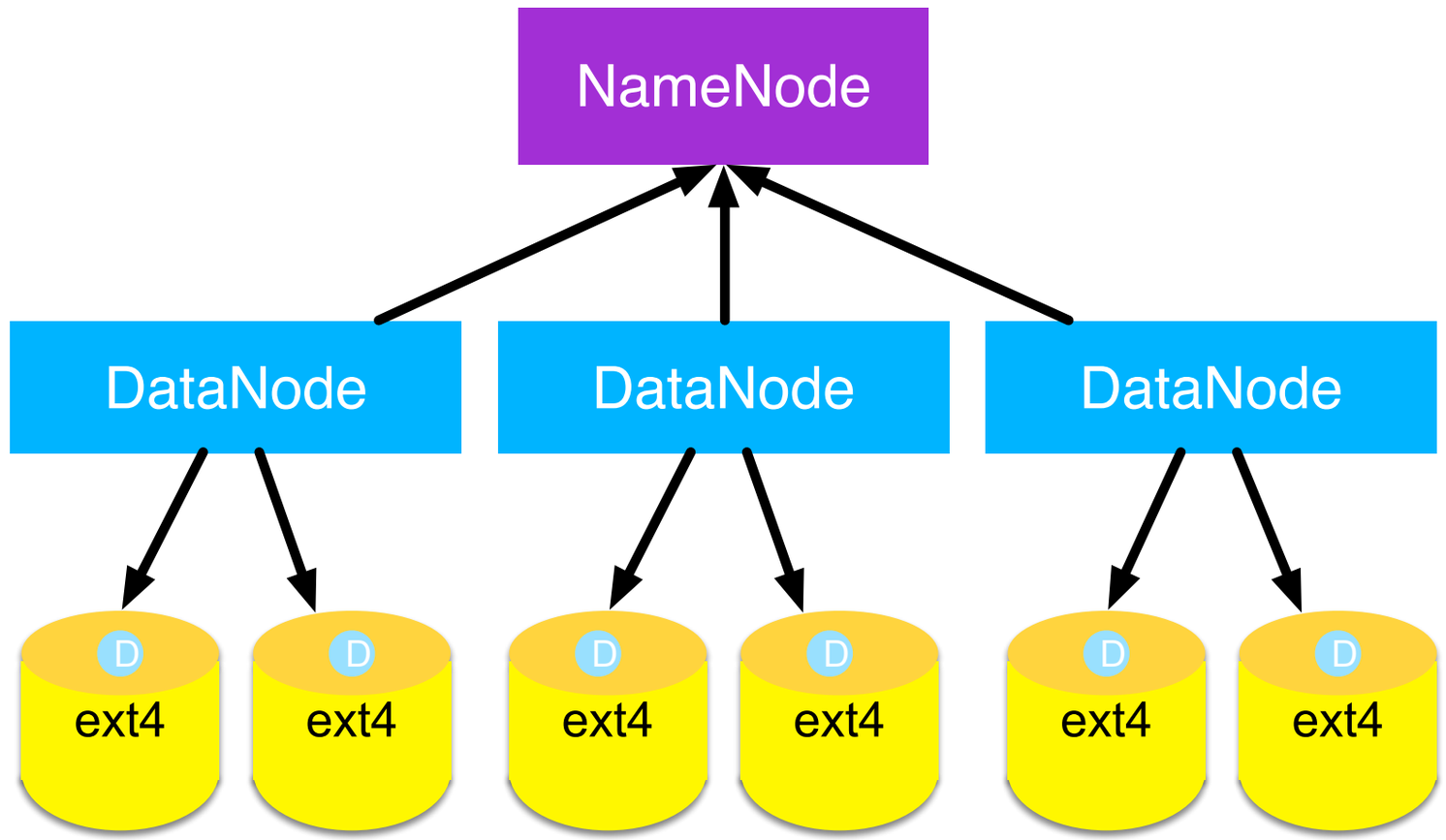
GIVE YOU UP

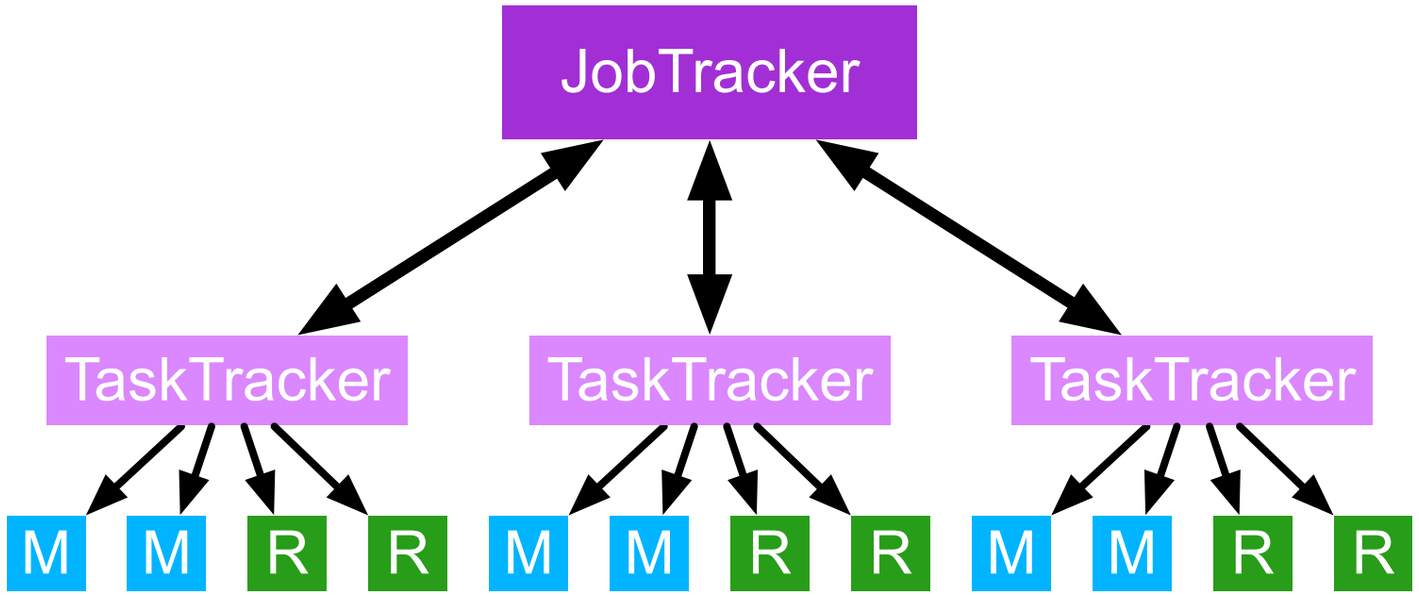

Hadoop
(‘common’ or ‘core’)
MapReduce
HDFS

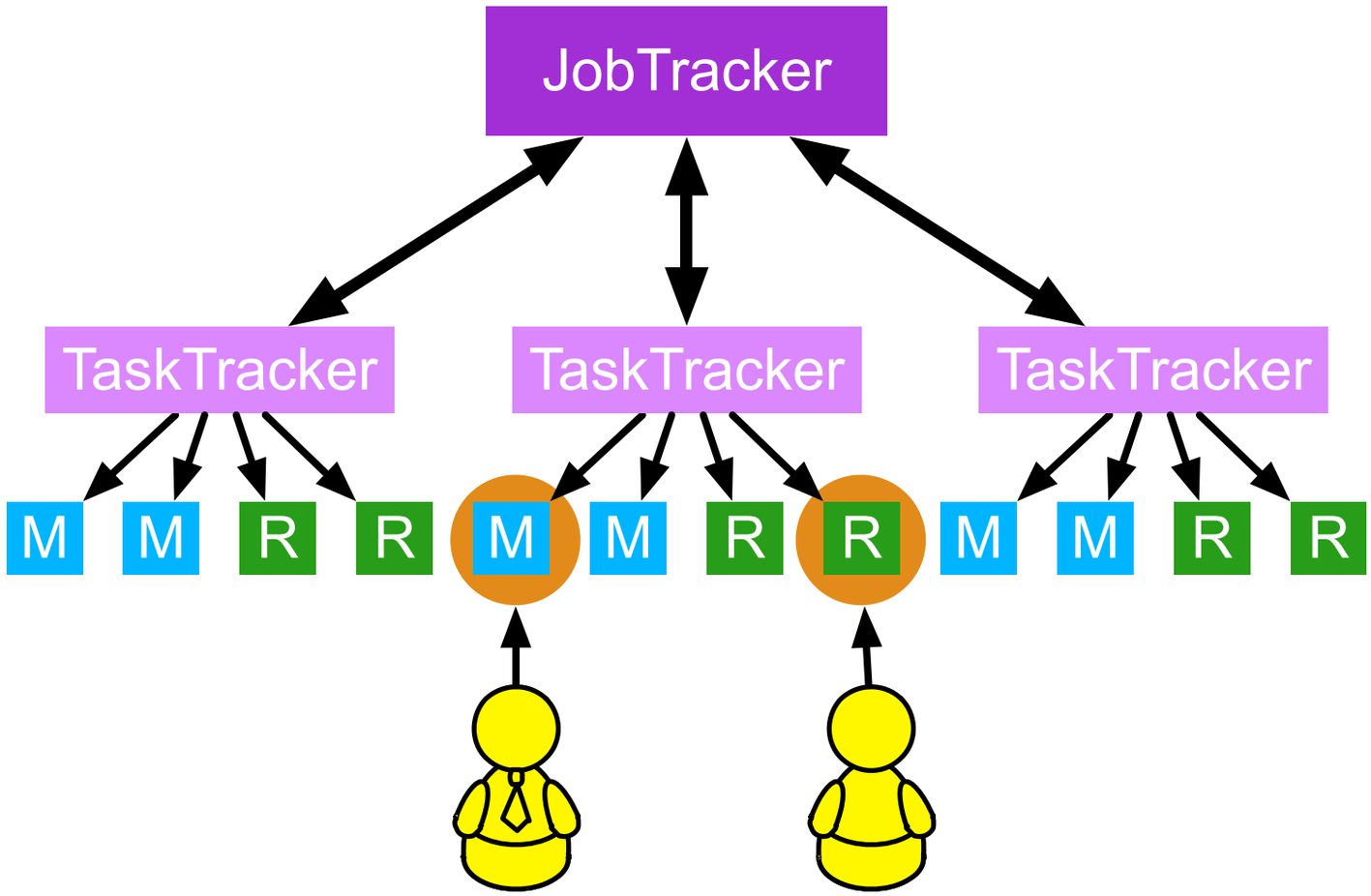

Hadoop
(‘common’ or ‘core’)
MapReduce
S3

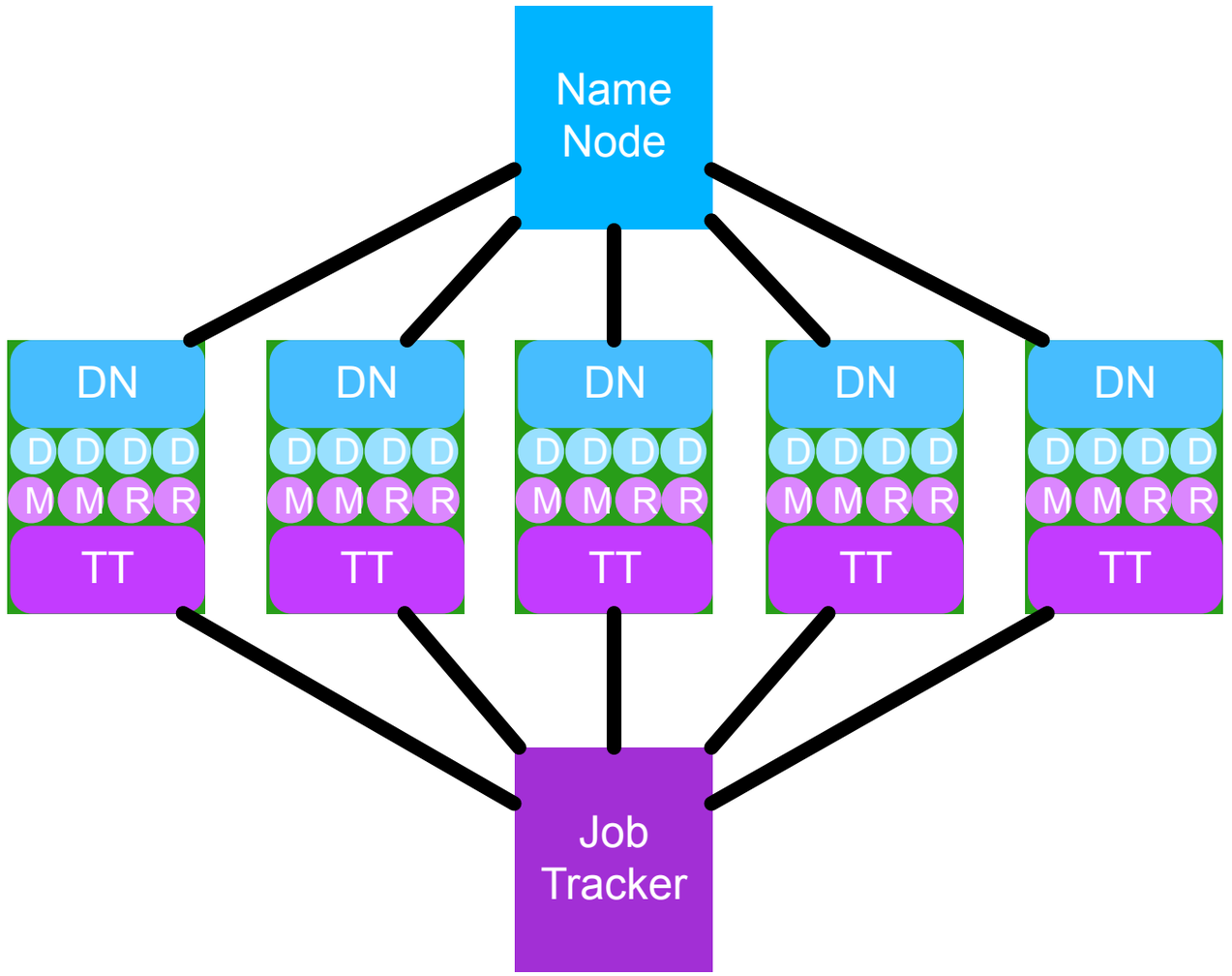

Hadoop
(‘common’ or ‘core’)
MapReduce
Gluster


Hadoop
(‘common’ or ‘core’)
HBase
HDFS


NameNode
DataNode
DataNode
DataNode
D
ext4
D
ext4
D
ext4
D
ext4
D
ext4
D
ext4


JobTracker
TaskTracker
TaskTracker
TaskTracker
M
M
R
R
M
M
R
R
M
M
R
R


JobTracker
TaskTracker
TaskTracker
TaskTracker
M
M
R
R
M
M
R
R
M
M
R
R


Name Node
DN
DN
DN
DN
DN
D D D D
M M R R
TT
TT
TT
TT
TT
Job Tracker

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



CAPTAIN OBVIOUS

Don’t Make Assumptions

tail’ing the logs won’t
tell you the whole story.

%

Monitor the masters!









Host Status Totals

Up	Down	Unreachable	Pending
54	0	0	0

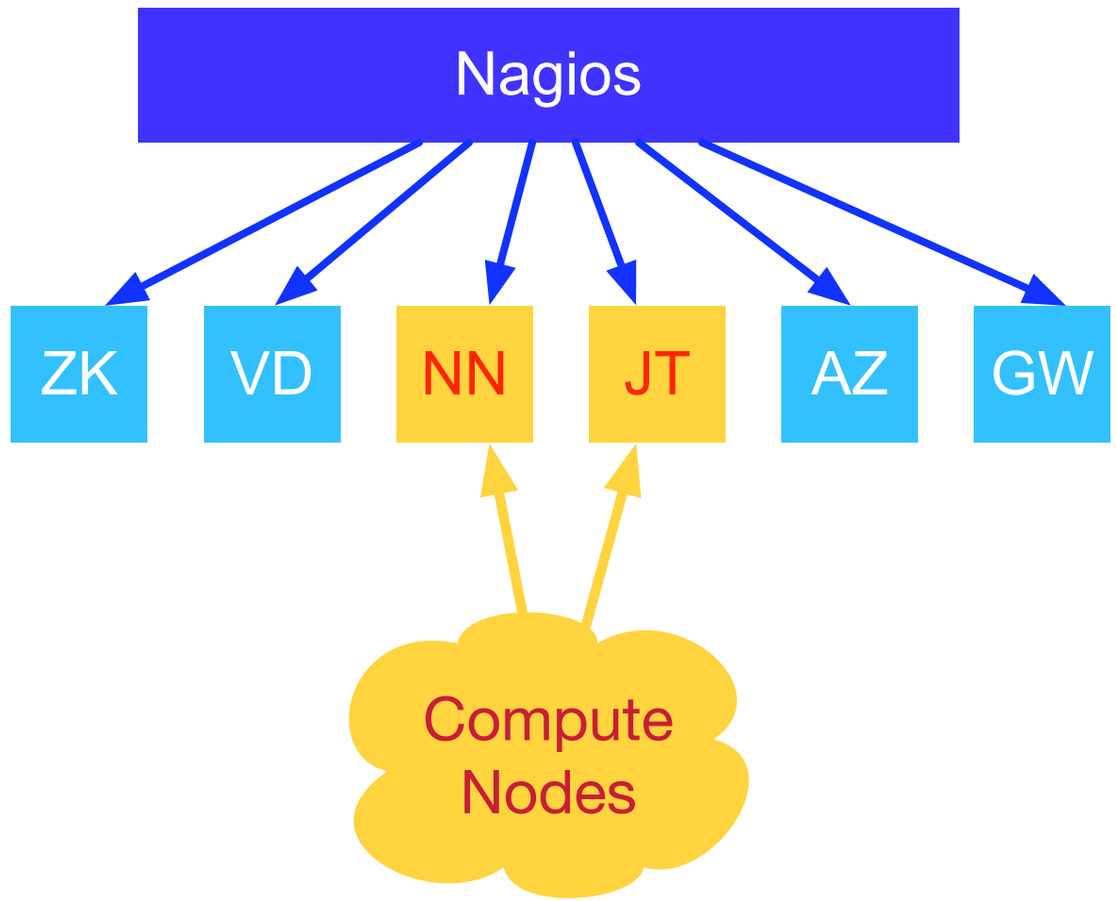
All Problems	All Types
0	54

Ok	W
220	

Service Status Details For All Hosts

Service	Status	Last Check	Duration	Attempt	Status Information
Canasta AZKABAN_v2 Port Check	OK	02-24-2013 20:20:11	11d 3h 7m 50s	1/3	OK: Port 8443
Disk Space Check - /export	OK	02-24-2013 20:20:55	11d 3h 7m 6s	1/3	DISK OK - free space: /export 7
Disk Space Check - /export/home	OK	02-24-2013 20:20:50	11d 3h 7m 11s	1/3	DISK OK - free space: /export/h
Disk Space Check - /grid/azlogs	OK	02-24-2013 20:20:16	11d 3h 7m 45s	1/3	DISK OK - free space: /grid 533
Disk Space Check - /export/home	OK	02-24-2013 20:22:51	11d 3h 6m 32s	1/3	DISK OK - free space: /export/h
Disk Space Check - /grid	OK	02-24-2013 20:21:43	11d 3h 6m 18s	1/3	DISK OK - free space: /grid 954
SSH Check	OK	02-24-2013 20:19:06	11d 3h 6m 32s	1/3	SSH OK - OpenSSH_5.3 (proto
Blacklisted Nodes	OK	02-24-2013 20:22:31	4d 15h 35m 30s	1/3	OK: Blacklisted Nodes Count: 0
CPU Load Check	OK	02-24-2013 20:22:22	11d 3h 5m 39s	1/3	OK - load average: 2.25, 2.37, 2
Disk Space Check - /export/home	OK	02-24-2013 20:22:49	11d 3h 5m 12s	1/3	DISK OK - free space: /export/h
Disk Space Check - /grid	OK	02-24-2013 20:22:47	11d 3h 5m 14s	1/3	DISK OK - free space: /grid 954
Disk Space Check - /grid/hadoop/logs	OK	02-24-2013 20:22:54	11d 3h 6m 24s	1/3	DISK OK - free space: /grid/had
JT-port-50030	OK	02-24-2013 20:19:51	11d 3h 8m 11s	1/3	OK: Port 50030
JobTracker Process	OK	02-24-2013 20:21:24	11d 3h 6m 37s	1/3	PROCS OK: 1 process with com
Lost Tracker Check	OK	02-24-2013 20:22:20	0d 21h 0m 41s	1/3	OK: No lost trackers
jobSubmitCheckDefault PASV	OK	02-24-2013 20:20:42	11d 2h 31m 31s	1/1	OK: Expected result : Sum = 18

- 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
- ...

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

```
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

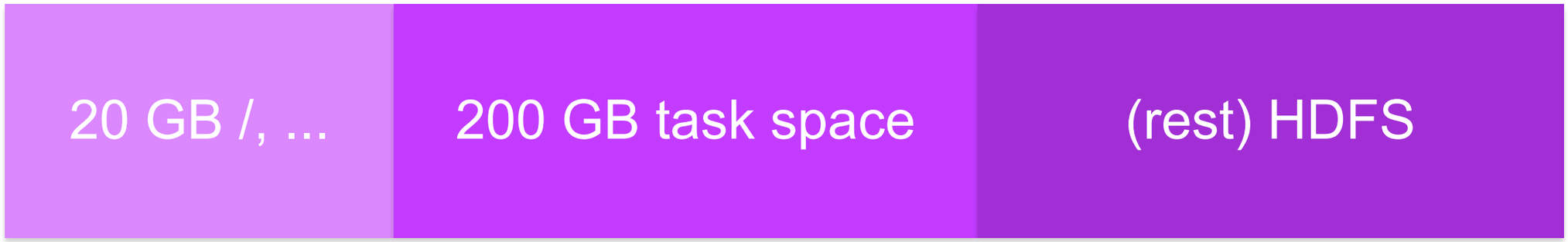
- Quota **everything**
  - Yes, including /tmp
  - No “show me all quotas” functionality

```
dfsadmin -setQuota
dfsadmin -setSpaceQuota
```

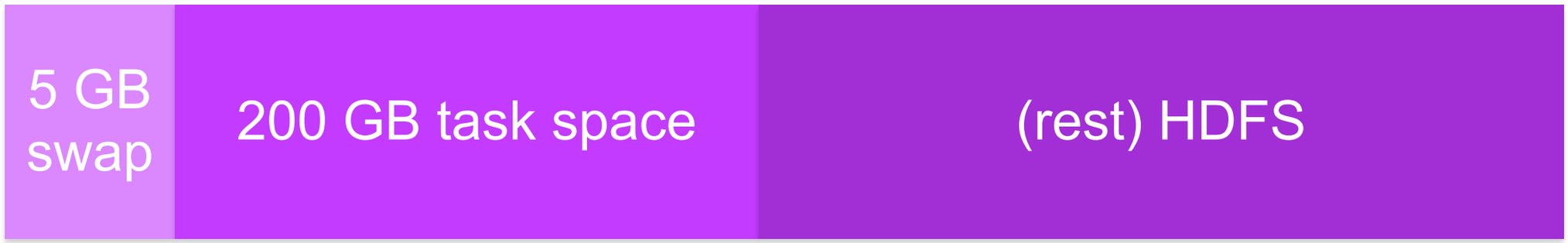
    - Be consistent:
      - `/user/*` all get same quota
    - Be flexible:
      - Make another dir for user’s to store big projects (e.g., `/project`)
- Be smart:
  - Have a policy that content in /tmp gets deleted after X days. Automate this!
  - Build reporting that shows files that are replicated less than 3 times

Compute Node Disk Partitioning as Protective Measure

- root partitioning



- non-root partinioning

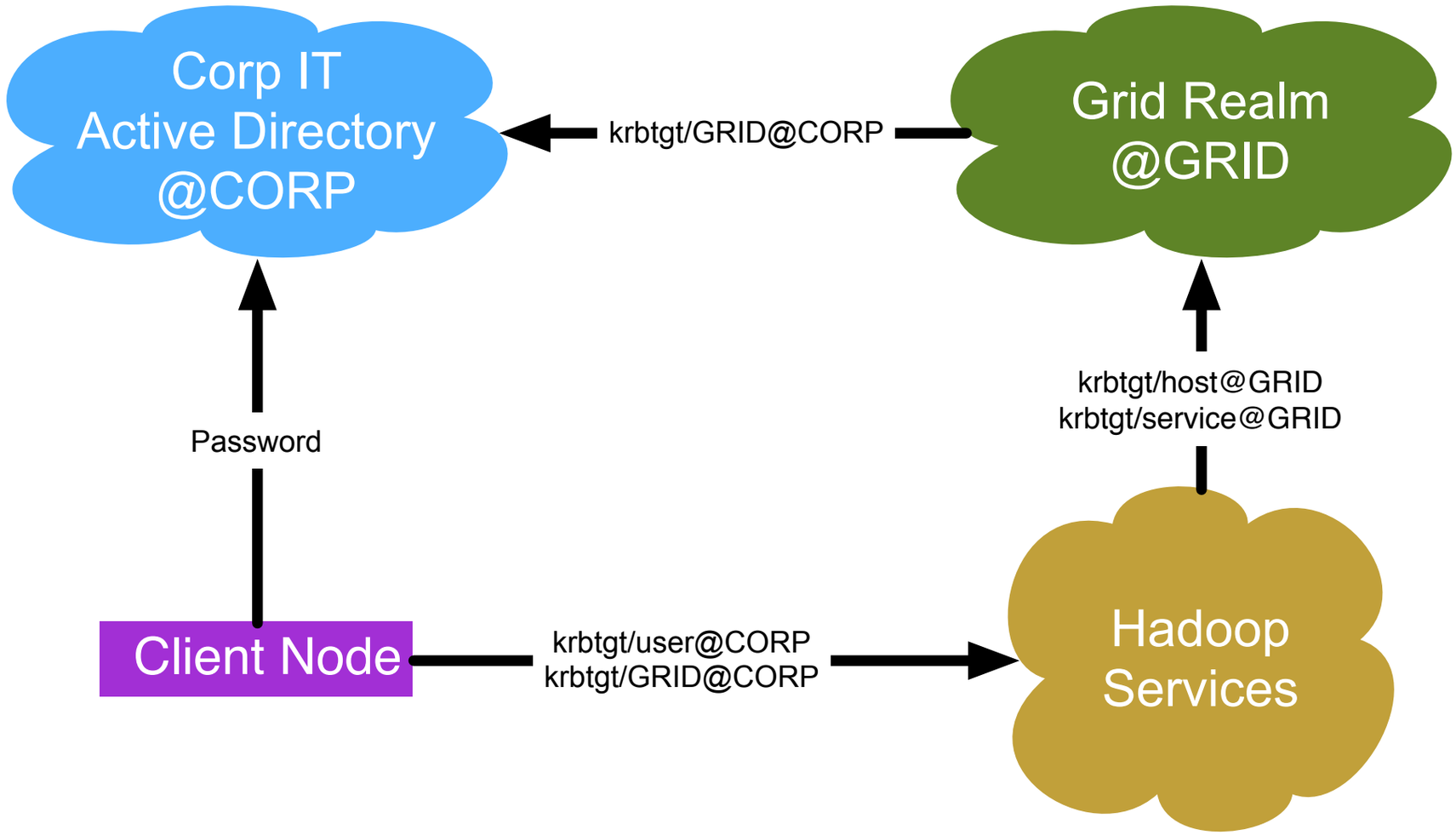


Security!

- Queue Level ACLs `mapred-queue-acls.xml`
  - users
  - groups
  - netgroups
- Service Level ACLs `hadoop-policy.xml`
  - hosts
  - users
  - groups
  - netgroups
  - Limitation: Web services are all or nothing! :(
  - Be aware: Hadoop uses ephemeral ports all over the place! :(

Kerberos!


I CAN HAZ KINIT?


Corp IT
Active Directory
@CORP
krbtgt/GRID@CORP
Grid Realm
@GRID
Password
krbtgt/host@GRID
krbtgt/service@GRID
Client Node
krbtgt/user@CORP
krbtgt/GRID@CORP
Hadoop
Services

Hadoop Self Service | Portal | **Account** | Group | Welcome Allen Wittenauer (awittena)

Account Management

Help

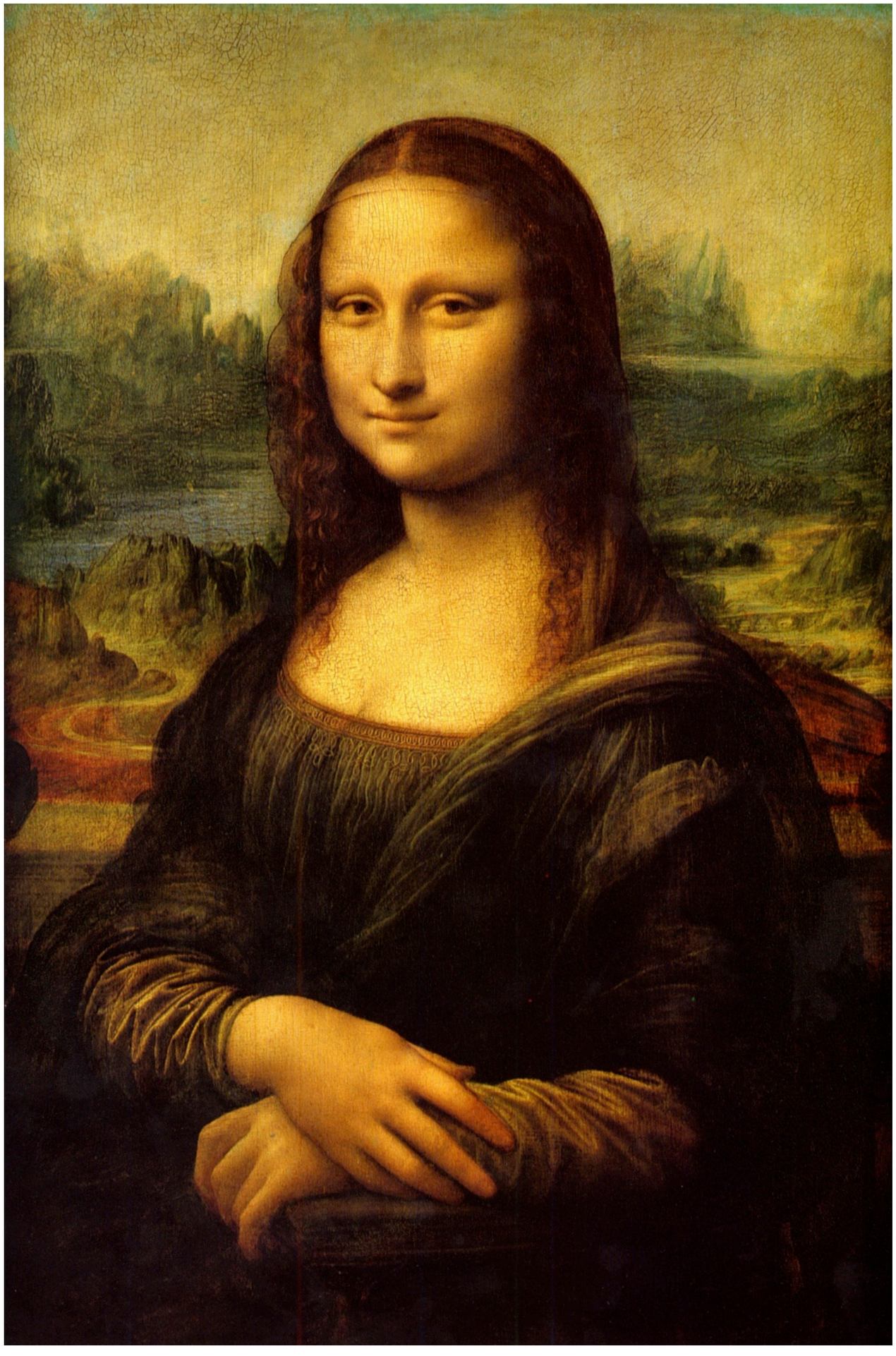
This tab is for managing your account on the development Hadoop grids at LinkedIn. If you don't already have an account, click the "Activate" button under status of eat1-magic. It can take up to 72 hours for your account request to be processed, if it takes longer, file a JIRA ticket.

My Account | Headless Accounts

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

Grid	Type	Status
eat1-magic	Development	✔ Active

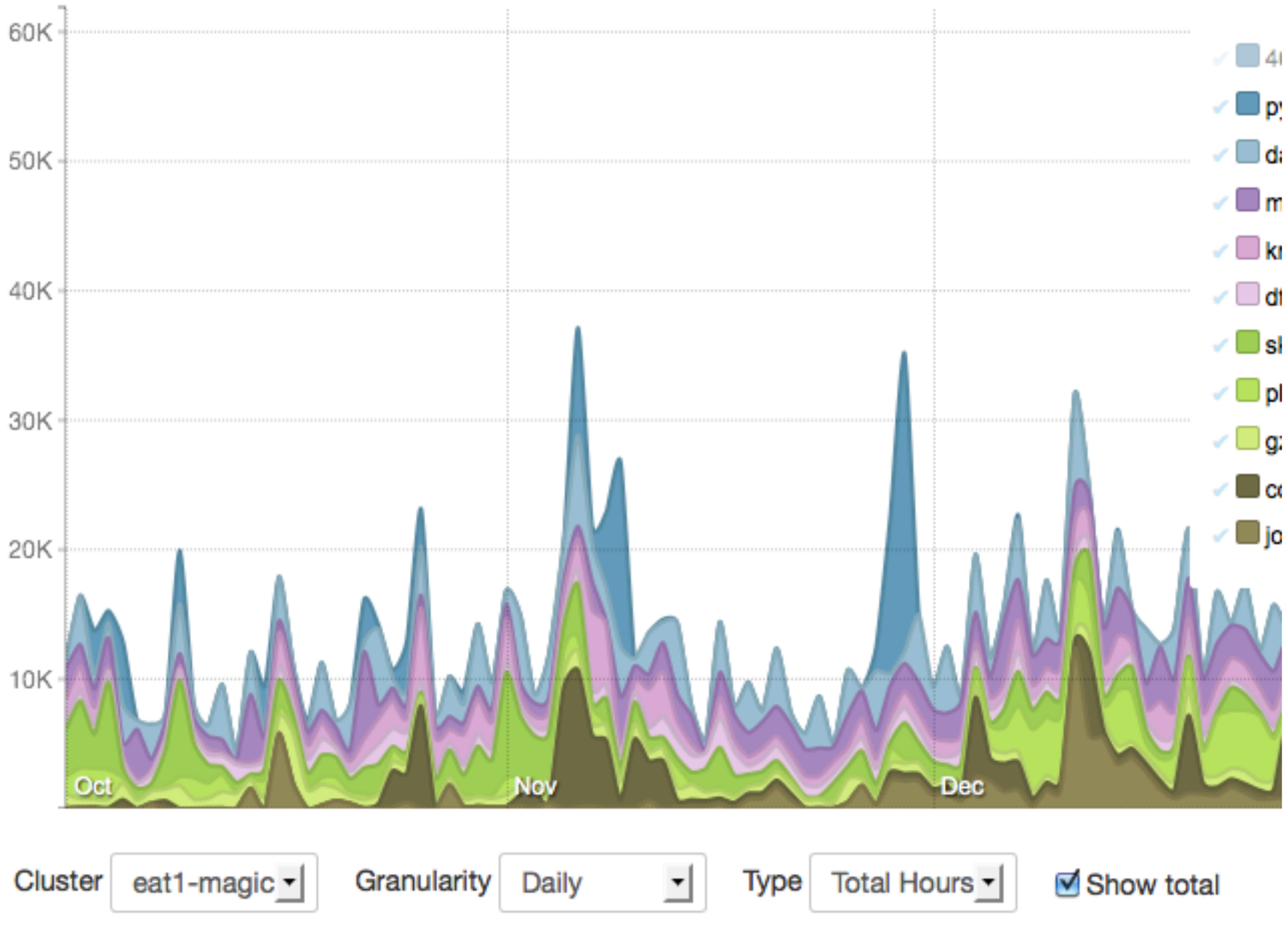
Version 2.0.1 © LinkedIn 2012





White Elephant

http://data.linkedin.com/opensource/white-elephant



- Fonzi: http://www.flickr.com/photos/elzey/7224689810
- Captain Obvious by artist Stuart McGhee. http://stuartmcghee.com/
- 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