Robinhood v3 on Sherlock

Site update

seamless Lustre/cloud storage integration

Grafana/graphite monitoring

Stéphane Thiell - Stanford Research Computing

Robinhood User Group 2016
Paris, France – Sep 19, 2016

Contents

Sherlock overview

Robinhood on Sherlock

Large filesystem challenges: query time

Robinhood v3 for Lustre/Cloud seamless integration

Robinhood Graphite/Grafana integration on Sherlock

Sherlock shared computing cluster

Sherlock

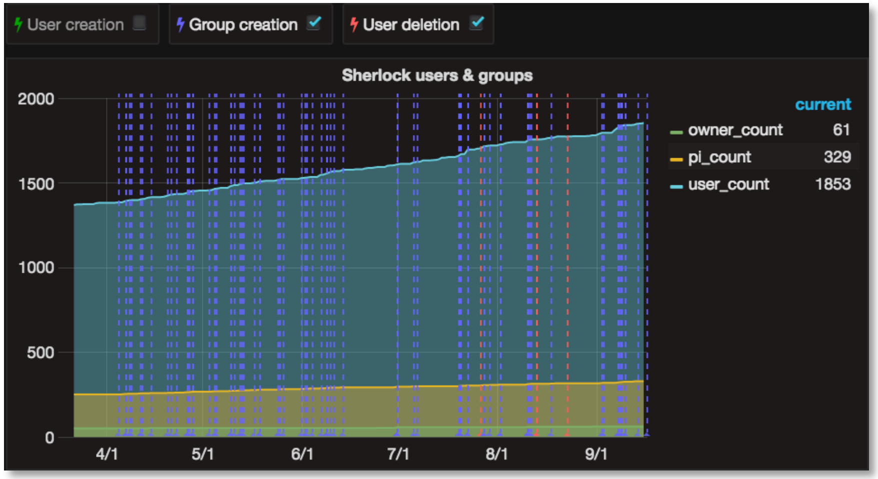
- **Condo** cluster (850+ nodes, CPU and GPU)
- Open to the Stanford community as a resource to support **sponsored research**

Sherlock's storage spaces

- Isilon (NFS) for home directories
- Lustre "scratch" behind lnet routers
  - Dell servers, MD3x60 disk arrays
  - Lustre 2.7 (IEEL 3.0)
  - 3.1 PB total, ~100 GB/s
  - **2.0PB** used, **527M inodes** used

Sherlock community

- 1853 users
- 329 sponsored faculty groups
- 61 owners



Robinhood server on Sherlock (1/2)

- We started with Robinhood v2.5 on a Sherlock's service node
  - Dell R720xd with 24 x 10K SATA drives in RAID-10 and 256GB RAM
  - CentOS 6.7
- In March 2016, /scratch had 370M inodes and we did setup a few **Robinhood Grafana dashboards** (with metrics from rbh-report and robinhood's log file)
- Also in March 2016, we started to play with HSM-to-the-cloud
  - tried an early version of Robinhood v3 (and never went back to 2.5 ☺)
  - but archiving were often faster than Robinhood archive policy run…
- Things moved forward in April 2016 as we needed to move out files from a few OSTs due to **storage hardware issues**, but we found out that rbh-find was **not usable** at that time…
- While waiting for some hardware upgrade, we:
  - upgraded MySQL 5.1 to the Community Edition v5.7, but it didn't work
  - finally installed **MySQL Community Edition v5.6** (and it worked)

Robinhood server on Sherlock (2/2)

- Finally, in late April 2016,
  - we bumped system memory from 256GB to **384GB**
  - added **2 SSDs** end of April 2016 (733GB usable)

1 x Dell R720xd 384GB 2 x Intel E5-2650 v2 8C 2.6GHz

- 2 x Intel 400GB 2.5" SATA 6Gbps MLC (RAID-0)
- 22 x 10K HDD SATA 6Gbps 1.2 TB (RAID-10)

- After that, Robinhood was finally usable for Sherlock's /scratch!
- On August 11, we recompiled v3.0 rc1 against Lustre 2.7 (IEEL 3.0.0)
  - robinhood `--alter-db` took less than 7 hours for 467M inodes
- Today we have:
  - **527M inodes** on /scratch
  - **566GB** of disk space used for MySQL
- Still, some queries take hours (like --class-info)

rbh-report command time (527,378,436 entries)

rbh-report –top-users	immediate
rbh-report –top-size	4 minutes
rbh-report –oldest-files	4 minutes
rbh-report –top-dirs	62 minutes
rbh-report --oldest-empty-dirs	86 minutes
rbh-report --class-info	90 minutes

Robinhood v3 for Lustre/Cloud seamless integration

- See tomorrow's presentation about Google Drive copytool

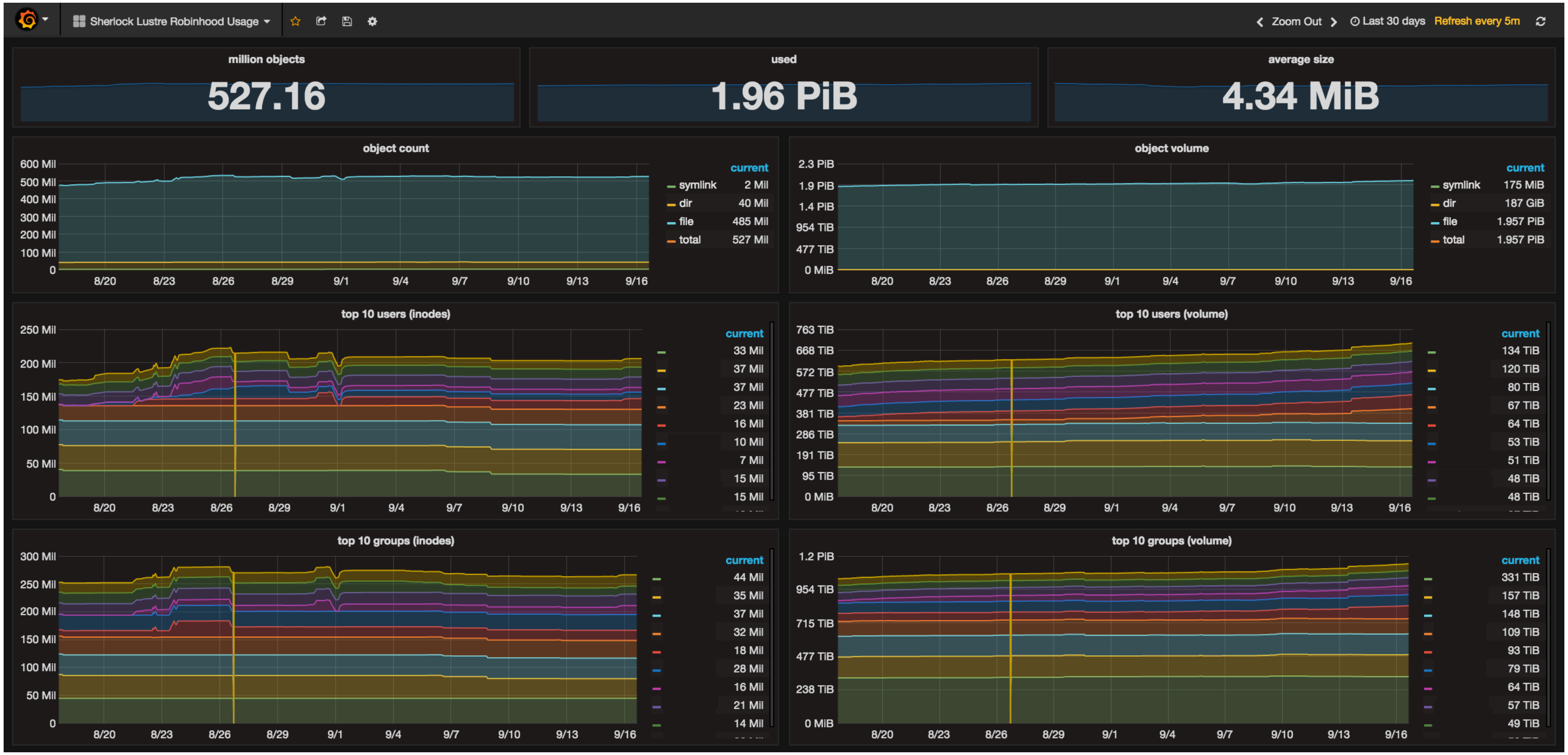
- Defined 6 different Robinhood fileclasses based on file size to help initial archiving process
  - required to find the best archive performance vs. file size
  - had to rescan a few times to make adjustment

- max_action_count has been very useful to avoid too many Lustre/HSM actions
  - reading hsm/actions takes way too much time that it is not possible to monitor it anymore

- Would love an "interleaved archiving mode" to mix smallfiles and bigfiles
  - ideally by percent of each (eg. 10% smallfiles, 90% bigfiles)
  - to push smallfiles while bigfiles are transferring, thus maximizing both transfer bandwidth and max QPS the cloud provider allows

Grafana/Graphite: Robinhood usage

- Tracking global and top 10 users/groups usage



Sherlock Lustre Robinhood Usage

Zoom Out | Last 30 days | Refresh every 5m

million objects	used	average size
527.16	1.96 PiB	4.34 MiB

object count

	current
symlink	2 Mil
dir	40 Mil
file	485 Mil
total	527 Mil

object volume

	current
symlink	175 MiB
dir	187 GiB
file	1.957 PiB
total	1.957 PiB

top 10 users (inodes)

current
33 Mil
37 Mil
37 Mil
23 Mil
16 Mil
10 Mil
7 Mil
15 Mil
15 Mil

top 10 users (volume)

current
134 TiB
120 TiB
80 TiB
67 TiB
64 TiB
53 TiB
51 TiB
48 TiB
48 TiB

top 10 groups (inodes)

current
44 Mil
35 Mil
37 Mil
32 Mil
18 Mil
28 Mil
16 Mil
21 Mil
14 Mil

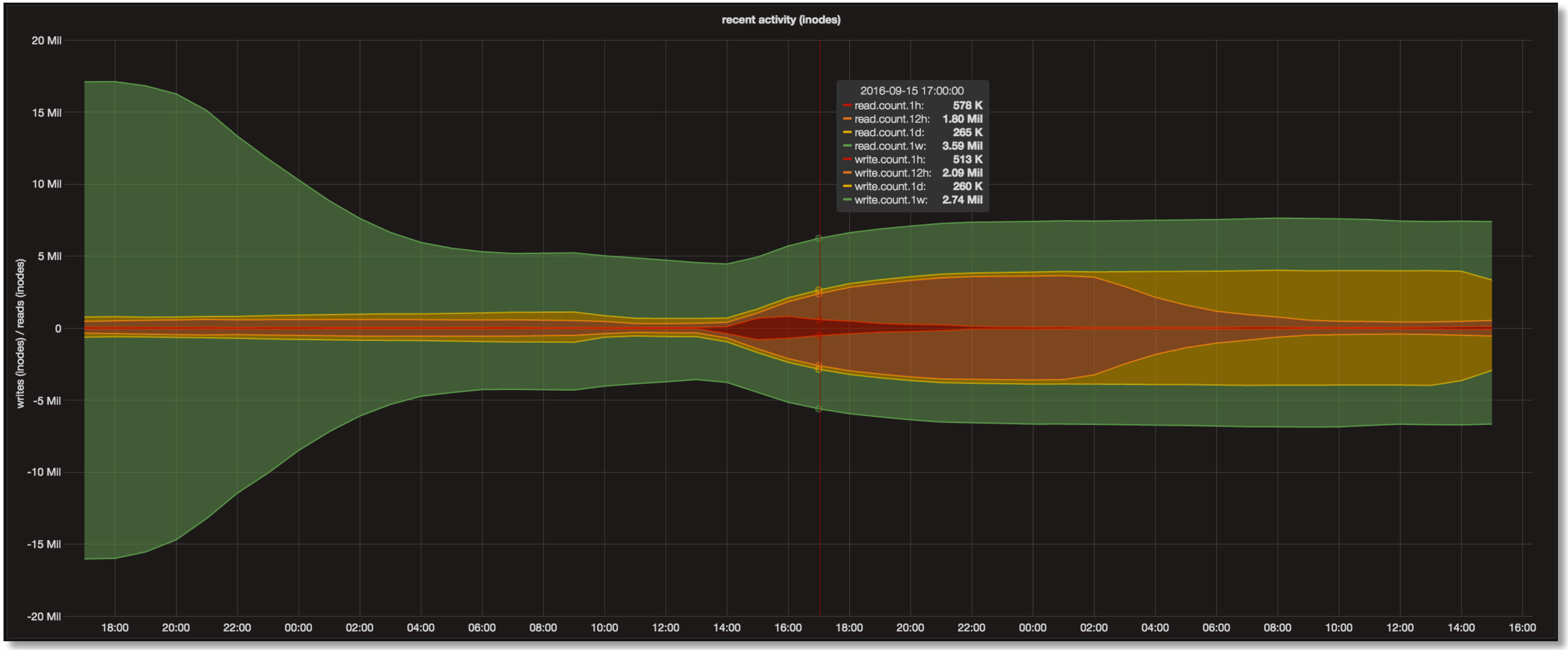
top 10 groups (volume)

current
331 TiB
157 TiB
148 TiB
109 TiB
93 TiB
79 TiB
64 TiB
57 TiB
49 TiB

Stanford University

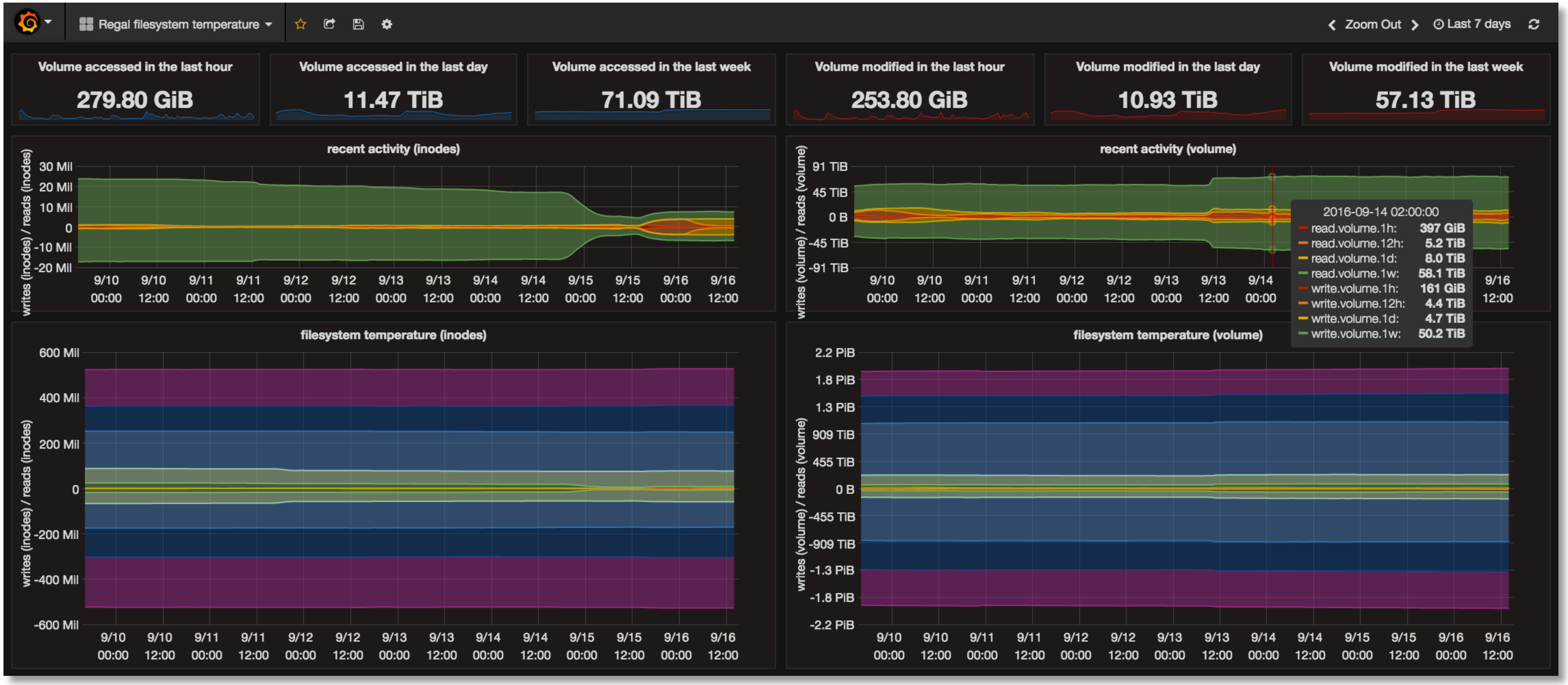
Grafana/Graphite: filesystem “temperature”

- Custom SQL queries against Robinhood DB
- Updated every hour
- Created by Kilian



Grafana/Graphite: filesystem “temperature” (cont’d)

- View of the filesystem recent activity in a single Grafana dashboard



Regal filesystem temperature

Zoom Out · Last 7 days

Volume accessed in the last hour 279.80 GiB

Volume accessed in the last day 11.47 TiB

Volume accessed in the last week 71.09 TiB

Volume modified in the last hour 253.80 GiB

Volume modified in the last day 10.93 TiB

Volume modified in the last week 57.13 TiB

recent activity (inodes)

recent activity (volume)

2016-09-14 02:00:00

read.volume.1h:	397 GiB
read.volume.12h:	5.2 TiB
read.volume.1d:	8.0 TiB
read.volume.1w:	58.1 TiB
write.volume.1h:	161 GiB
write.volume.12h:	4.4 TiB
write.volume.1d:	4.7 TiB
write.volume.1w:	50.2 TiB

filesystem temperature (inodes)

filesystem temperature (volume)

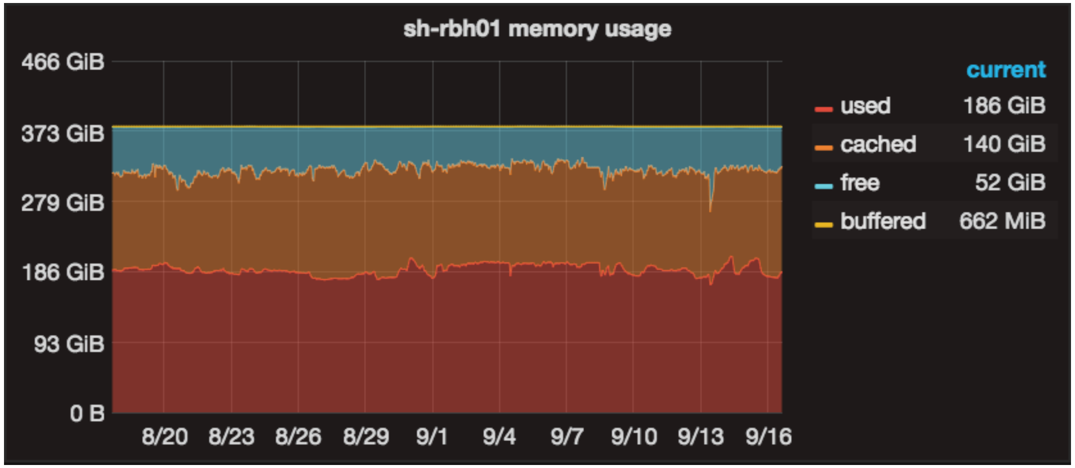
Grafana/Graphite: filesystem “temperature” (cont’d)

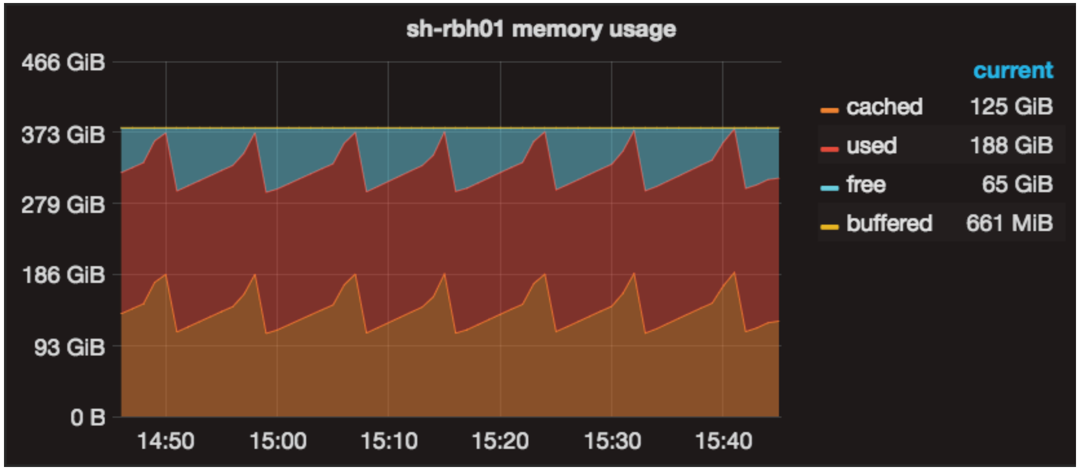
- 2 big SQL queries
- 20 minutes per query (~530M inodes)
- 1.5M read/sec seen using innotop during query run
- Example of query for modified files:

```
SELECT age, SUM(c) AS cnt, SUM(v) AS vol FROM (
    SELECT c, v, CASE
        WHEN log_age < ROUND(LOG(10,3600),5) THEN '1h'
        WHEN log_age < ROUND(LOG(10,43200),5) THEN '12h'
        WHEN log_age < ROUND(LOG(10,86400),5) THEN '1d'
        WHEN log_age < ROUND(LOG(10,604800),5) THEN '1w'
        WHEN log_age < ROUND(LOG(10,2592000),5) THEN '1m'
        WHEN log_age < ROUND(LOG(10,15552000),5) THEN '6m'
        WHEN log_age < ROUND(LOG(10,31104000),5) THEN '1y'
        ELSE 'over1y'
    END
    AS age FROM (
        SELECT  ROUND(LOG(10,UNIX_TIMESTAMP(NOW())-last_mod),5)  AS log_age,
                COUNT(*) AS c,
                IFNULL(SUM(size),0) AS v
        FROM ENTRIES GROUP BY log_age)
    AS ps )
AS stats GROUP BY age
```

Grafana/Graphite: memory usage

- Using collectd

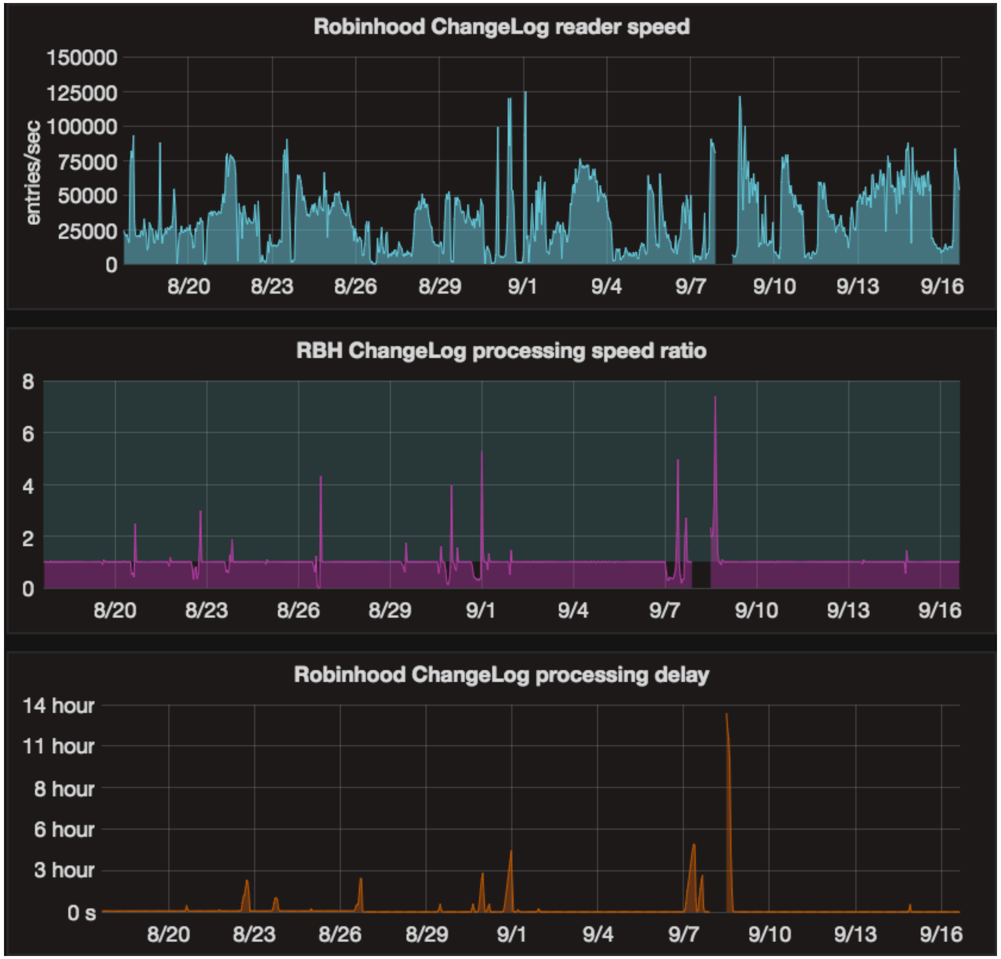




Stanford University

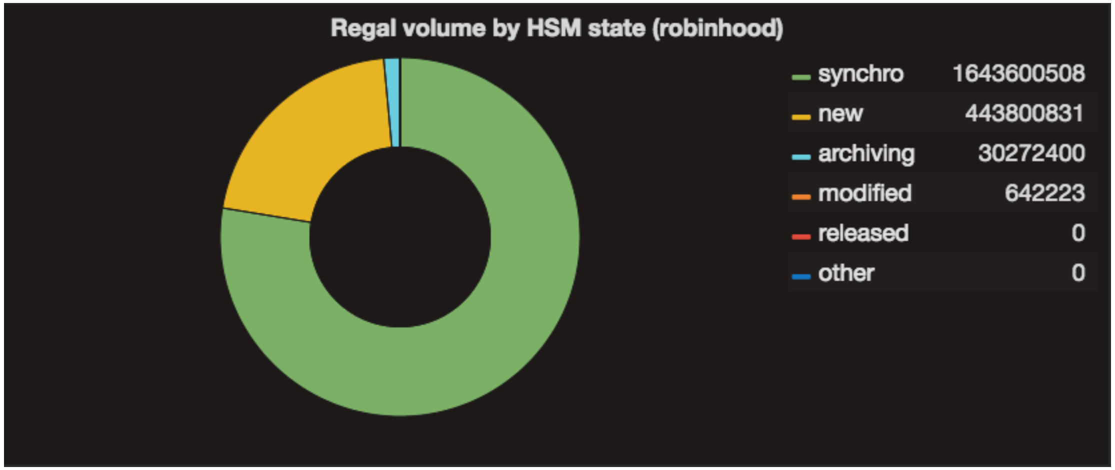
Grafana/Graphite: ChangeLog processing metrics

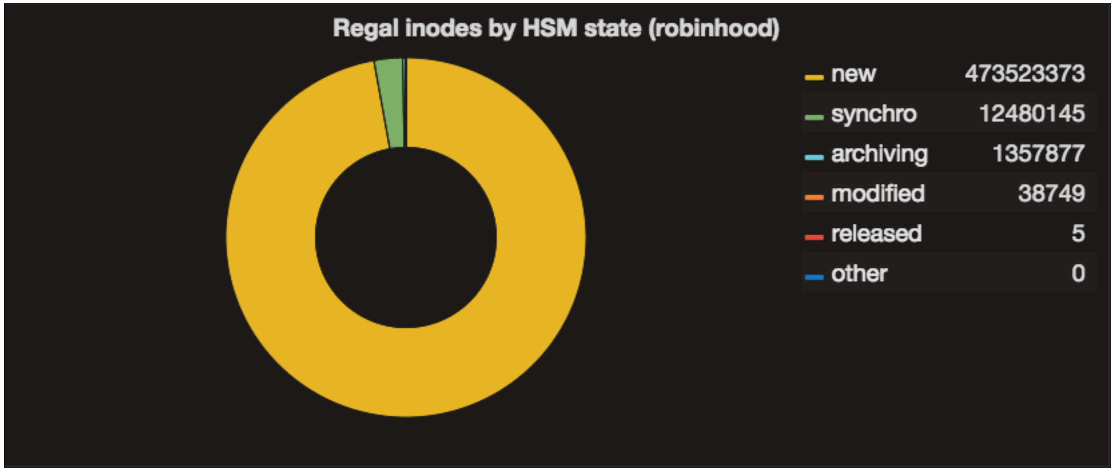
- Parsed from robinhood log's "STATS lines"



Grafana/Graphite: files and volume by state

- Grafana v3 supports Pie charts





Other feedback for discussion

- “+” operator for fileclass is confusing
- We heavily parse rbh-report
  - please don’t change rbh-report’s csv-based output between versions
  - a API would be convenient for many scripts
  - could the new REST web API answer this need?

Questions?

sthiell@stanford.edu