FROM RESEARCH TO INDUSTRY

RobinHood Project Update

Robinhood User Group 2016

Thomas Leibovici <thomas.leibovici@cea.fr>

SEPTEMBER, 19th 2016

www.cea.fr

Project update

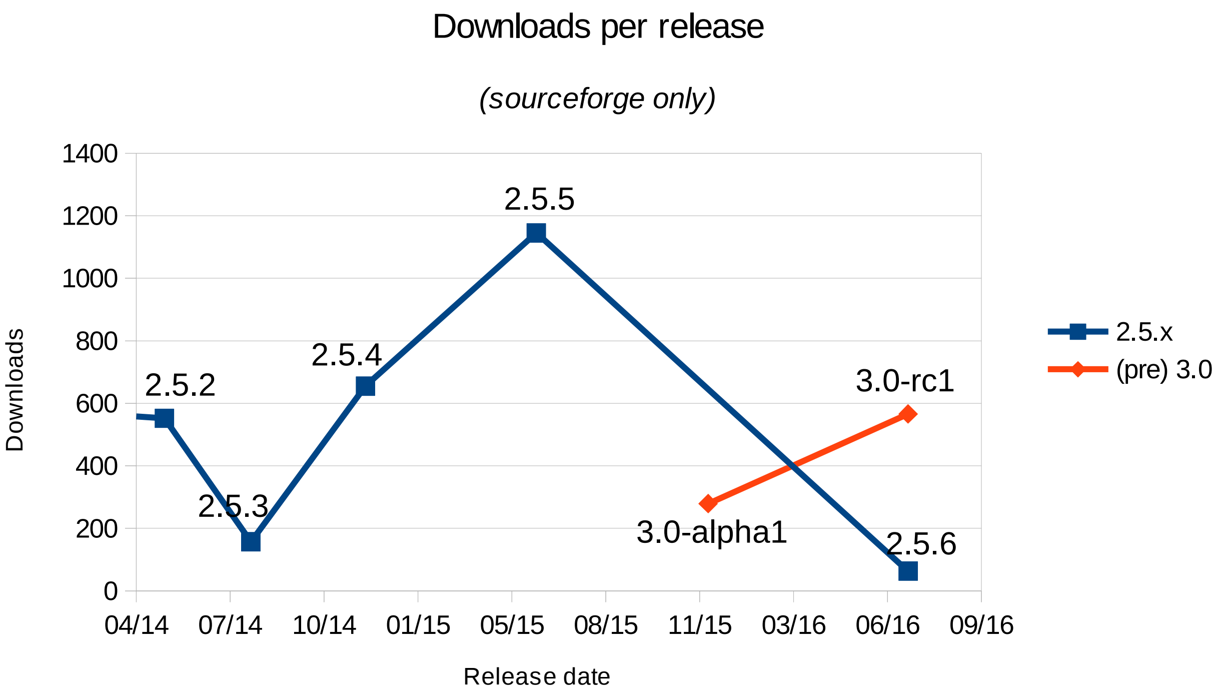
Latest Releases

- Robinhood 2.5.6 (july 2016)
  - Update of 2.5.5 with a few patches that were standing in the git repository
  - Support:
    - RHEL 5, 6, 7
    - Lustre 1.8 to 2.8

- Robinhood 3.0-alpha1 (Dec 2015)
- Robinhood 3.0-alpha2 (March 2016)
- Robinhood 3.0-rc1 (July 2016)
- Robinhood 3.0 final: Just released! (Sept 2016)
  - Support:
    - RHEL 6, 7
    - Lustre 2.1 to 2.8

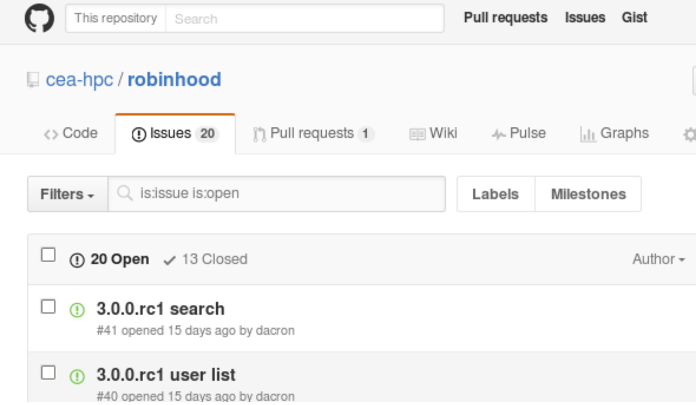
Release Stats

- Version 2.5.5: the most dowloaded robinhood release
  - Nearly 1200 downloads
- Version 3 pre-releases already have a significant base of users



Community Resources

- **Github** and **Gerrithub** are pillars of robinhood project
- Github:
  - Main git repository: https://github.com/cea-hpc/robinhood.git
  - Wiki (project page, online documentation, …)
  - Issue reporting and tracking
    - as discussed at LUG'16

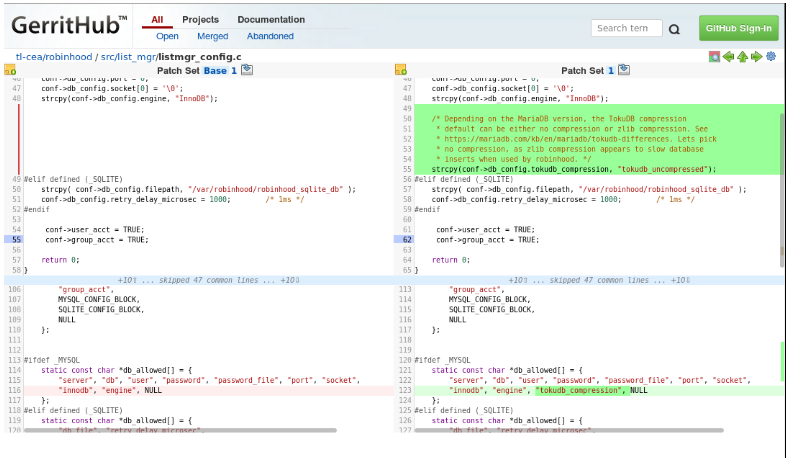


□ 20 Open ✓ 13 Closed

□ **3.0.0.rc1 search**
#41 opened 15 days ago by dacron

□ **3.0.0.rc1 user list**
#40 opened 15 days ago by dacron

- Gerrithub (code review):
  - https://review.gerrithub.io
  - Project: cea-hpc/robinhood
  - **All new code goes through it**
    (please, no “pull requests” on github)



tl-cea/robinhood / src/list_mgr/listmgr_config.c

Patch Set Base 1

```
    conf->db_config.port = 0;
    conf->db_config.socket[0] = '\0';
    strcpy(conf->db_config.engine, "InnoDB");
#elif defined (_SQLITE)
    strcpy( conf->db_config.filepath, "/var/robinhood/robinhood_sqlite_db" );
    conf->db_config.retry_delay_microsec = 1000;        /* 1ms */
#endif

     conf->user_acct = TRUE;
     conf->group_acct = TRUE;

    return 0;
}
                 +10⇧ ... skipped 47 common lines ... +10⇩
        "group_acct",
        MYSQL_CONFIG_BLOCK,
        SQLITE_CONFIG_BLOCK,
        NULL
    };


#ifdef _MYSQL
    static const char *db_allowed[] = {
        "server", "db", "user", "password", "password_file", "port", "socket",
        "innodb", "engine", NULL
    };
#elif defined (_SQLITE)
    static const char *db_allowed[] = {
        "db_file", "retry_delay_microsec",
```

Patch Set 1

```
    conf->db_config.port = 0;
    conf->db_config.socket[0] = '\0';
    strcpy(conf->db_config.engine, "InnoDB");

    /* Depending on the MariaDB version, the TokuDB compression
     * default can be either no compression or zlib compression. See
     * https://mariadb.com/kb/en/mariadb/tokudb-differences. Lets pick
     * no compression, as zlib compression appears to slow database
     * inserts when used by robinhood. */
    strcpy(conf->db_config.tokudb_compression, "tokudb_uncompressed");
#elif defined (_SQLITE)
    strcpy( conf->db_config.filepath, "/var/robinhood/robinhood_sqlite_db" );
    conf->db_config.retry_delay_microsec = 1000;        /* 1ms */
#endif

     conf->user_acct = TRUE;
     conf->group_acct = TRUE;

    return 0;
}
                 +10⇧ ... skipped 47 common lines ... +10⇩
        "group_acct",
        MYSQL_CONFIG_BLOCK,
        SQLITE_CONFIG_BLOCK,
        NULL
    };


#ifdef _MYSQL
    static const char *db_allowed[] = {
        "server", "db", "user", "password", "password_file", "port", "socket",
        "innodb", "engine", "tokudb_compression", NULL
    };
#elif defined (_SQLITE)
    static const char *db_allowed[] = {
        "db_file", "retry_delay_microsec",
```

- Still managed on sourceforge:
  - Mailing lists
  - Short URL: http://robinhood.sf.net (points to github's wiki)
  - Download center

Recent changes about git branches and releases

Adopted a more standard git workflow

- New developments go to “master”
- Older versions are branched to “b_x.y”

Example, today:

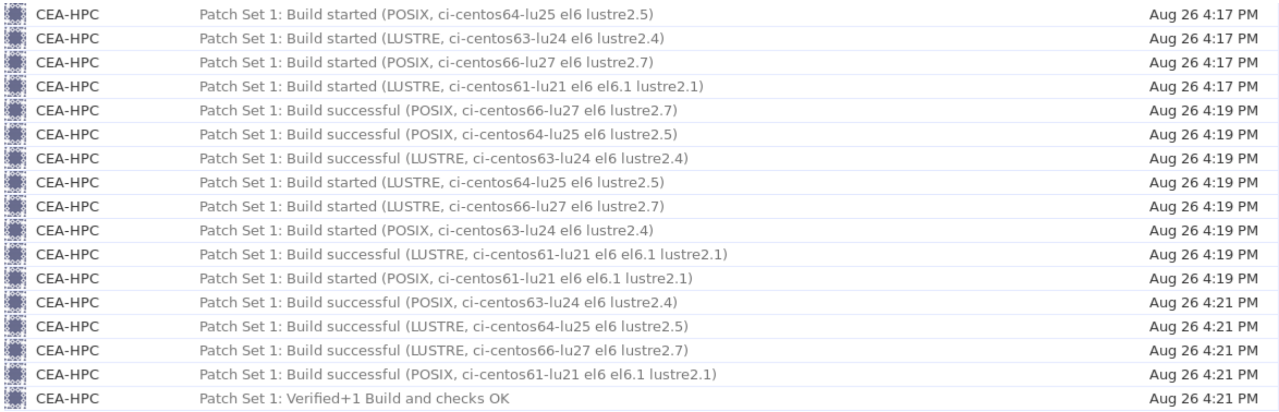
- “master” is v3.0 (and will become v3.1)
- “b_2.5” contains last 2.5.x (2.5.6)

Shorter release cycle

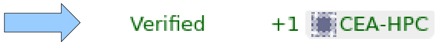
- (Preferred) next release is 3.1, not 3.0.1
- 3.x.x only for emergency fixes

On Gerrithub: Automated Built Test

- An automatic build test has been bound to gerrithub:
  - Run on various Lustre versions from 2.1 to 2.8 + POSIX FS
  - For security reasons: it is run only when a patch is accepted (+2)

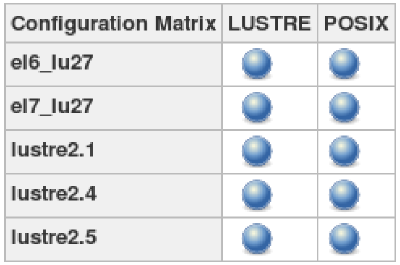


```
CEA-HPC    Patch Set 1: Build started (POSIX, ci-centos64-lu25 el6 lustre2.5)                 Aug 26 4:17 PM
CEA-HPC    Patch Set 1: Build started (LUSTRE, ci-centos63-lu24 el6 lustre2.4)                Aug 26 4:17 PM
CEA-HPC    Patch Set 1: Build started (POSIX, ci-centos66-lu27 el6 lustre2.7)                 Aug 26 4:17 PM
CEA-HPC    Patch Set 1: Build started (LUSTRE, ci-centos61-lu21 el6 el6.1 lustre2.1)          Aug 26 4:17 PM
CEA-HPC    Patch Set 1: Build successful (POSIX, ci-centos66-lu27 el6 lustre2.7)              Aug 26 4:19 PM
CEA-HPC    Patch Set 1: Build successful (POSIX, ci-centos64-lu25 el6 lustre2.5)              Aug 26 4:19 PM
CEA-HPC    Patch Set 1: Build successful (LUSTRE, ci-centos63-lu24 el6 lustre2.4)             Aug 26 4:19 PM
CEA-HPC    Patch Set 1: Build started (LUSTRE, ci-centos64-lu25 el6 lustre2.5)                Aug 26 4:19 PM
CEA-HPC    Patch Set 1: Build started (LUSTRE, ci-centos66-lu27 el6 lustre2.7)                Aug 26 4:19 PM
CEA-HPC    Patch Set 1: Build started (POSIX, ci-centos63-lu24 el6 lustre2.4)                 Aug 26 4:19 PM
CEA-HPC    Patch Set 1: Build successful (LUSTRE, ci-centos61-lu21 el6 el6.1 lustre2.1)       Aug 26 4:19 PM
CEA-HPC    Patch Set 1: Build started (POSIX, ci-centos61-lu21 el6 el6.1 lustre2.1)           Aug 26 4:19 PM
CEA-HPC    Patch Set 1: Build successful (POSIX, ci-centos63-lu24 el6 lustre2.4)              Aug 26 4:21 PM
CEA-HPC    Patch Set 1: Build successful (LUSTRE, ci-centos64-lu25 el6 lustre2.5)             Aug 26 4:21 PM
CEA-HPC    Patch Set 1: Build successful (LUSTRE, ci-centos66-lu27 el6 lustre2.7)             Aug 26 4:21 PM
CEA-HPC    Patch Set 1: Build successful (POSIX, ci-centos61-lu21 el6 el6.1 lustre2.1)        Aug 26 4:21 PM
CEA-HPC    Patch Set 1: Verified+1 Build and checks OK                                        Aug 26 4:21 PM
```



Verified   +1 CEA-HPC

- Full test-suite is still triggered manually (in a private jenkins)
  - Distributed as robinhood-tests RPM



Configuration Matrix	LUSTRE	POSIX
el6_lu27	●	●
el7_lu27	●	●
lustre2.1	●	●
lustre2.4	●	●
lustre2.5	●	●

What's new in Robinhood?

Robinhood 3 in a Nutshell (recall)

V2 “flavors” and their commands

robinhood-tmpfs	robinhood-lhsm	robinood-backup
robinhood	rbh-lhsm	rbh-backup
rbh-diff	rbh-lhsm-diff	rbh-backup-diff
rbh-report	rbh-lhsm-report	rbh-backup-report
rbh-du	rbh-lhsm-du	rbh-backup-du
rbh-find	rbh-lhsm-find	rbh-backup-find
		...

→ A static set of available policies per flavor

V3: a single instance to manage all “legacy” policies ...and much more!

robinhood
robinhood
rbh-diff
rbh-report
rbh-undelete
rbh-du
rbh-find

→ Policies declared in configuration

*Robinhood for Lustre vs. POSIX FS are still distributed as distinct RPMs: `robinhood-lustre` and `robinhood-posix`

V3 Generic policies

All policies declared by configuration

- Example: Old tmpfs purge policy (deleting old files)
- In v3: “include” the related template:

```
%include “includes/tmpfs.inc”
```

```
declare_policy cleanup {
    scope { type != directory }
    default_action = common.unlink;
    status_manager = none;
    default_lru_sort_attr = last_access;
}
```

- … or define your own policy!

- Then specify your policy rules as usual*:

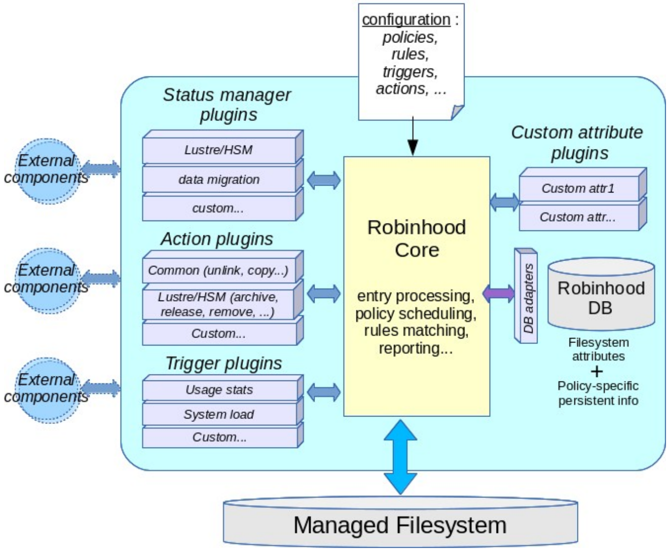
```
cleanup_rules {
    ignore_fileclass = keep_it;
    ignore_fileclass = keep_that_too;

    rule purge30d {
        target_fileclass = user_data1;
        target_fileclass = user_data2;
        condition { last_access > 30d}
    }
    …
}
```

** “purge” renamed to “cleanup” in v3 templates for clarification*

Other new features in robinhood v3 (recall)

Presented last year

- Plugin-based architecture:
  - Action plugins
  - Status managers
  - Custom plugins
- Configurable actions, action parameters
- New fileclass implementation and reporting



- See http://robinhood.sourceforge.net/rug15/rug_robinhood_v3.pdf for more details

Nice new features since last year (1/7)

rbh-find -printf (by Cray)

- Makes it possible to build highly customized reports:
  - Filter using rbh-find options
  - Customized output using “-printf” option
- Most standard arguments of 'find -printf' are supported
  - “`%p`” for path, “`%M`” or “`%m`” for mode, “`%s`” for size...
- + all robinhood specific information (prefix: `%R`)
  - “`%Rf`” for Lustre FID, “`%Ro`” for OSTs, “`%Rc`” for fileclasses
  - “`%Rm{`*<module>*`.`*<attribute>*`}`” for module-specific attributes
- Example:

```
rbh-find -status lhsm:released -printf "%p %Rm{lhsm.archive_id}\n"
```

- See “man rbh-find” for full description

Nice new features since last year (2/7)

Automatic DB schema conversion

- No longer need to drop the DB in case of schema change
  - This can now be needed after declaring/removing policy definitions
- DB schema can now be fixed automatically
  - Insert/drop/rename fields, change field type, default...
- It's under control:
  - If a change is detected, robinhood reports the detected changes in its log but keeps your DB unchanged.

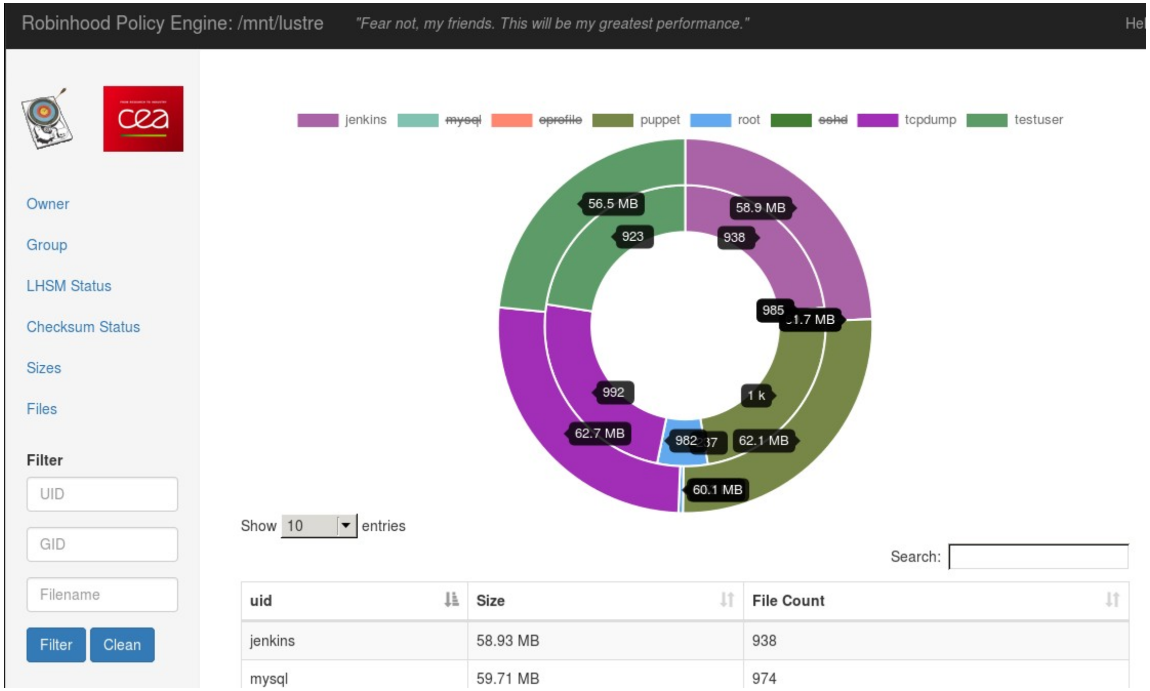
```
2016/09/02 15:15:20 [11664/1] ListMgr | DB schema change detected: field 'ENTRIES.lhsm_uuid' must be
added  => Run 'robinhood --alter-db' to apply this change.
2016/09/02 15:15:20 [11664/1] ListMgr | DB schema change detected: type of field 'ANNEX_INFO.link' must
be changed  => Run 'robinhood --alter-db' to confirm this change.
```

  - At this point, you can backup your DB if you wish :-)
  - Then explicitly run “`robinhood --alter-db`” to apply the changes.

Nice new features since last year (3/7)

New web interface (in final 3.0)

- New WebUI, compatible with robinhood 3 DB schema
- Upgrade of the old web UI, with the state of the art in terms of web interfaces and charts
- Fine-grained authentication
- Compatibility with newer MySQL versions

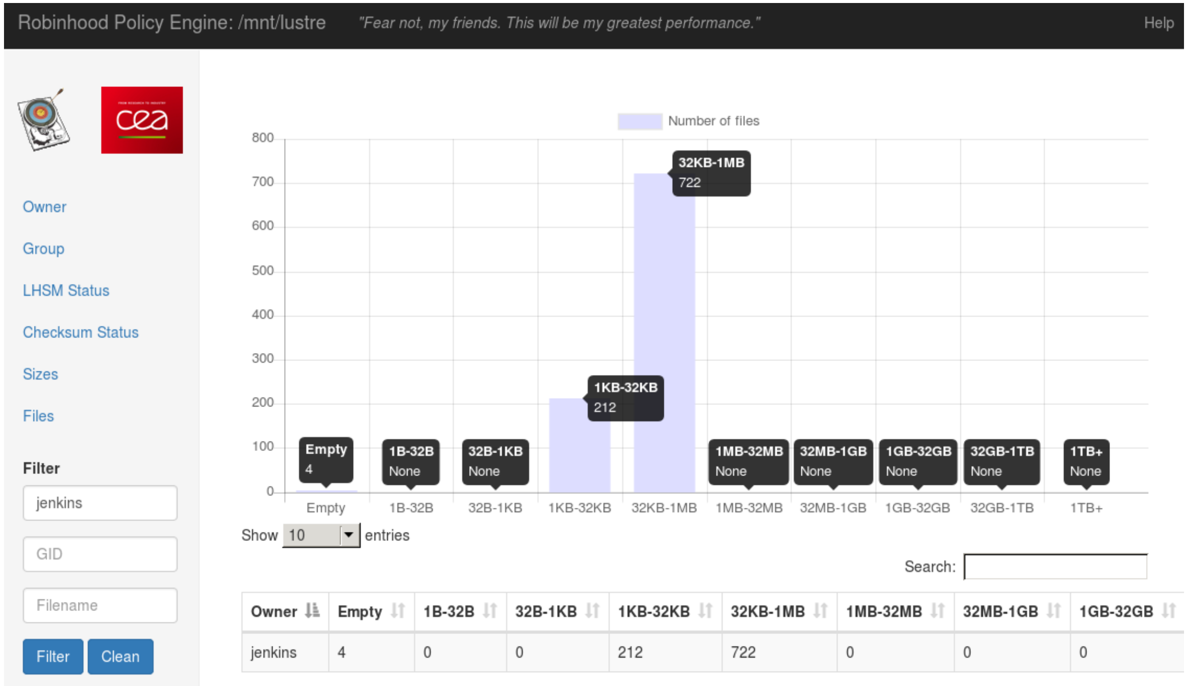


uid	Size	File Count
jenkins	58.93 MB	938
mysql	59.71 MB	974

Nice new features since last year (4/7)

New web interface (in final 3.0)

- New WebUI, compatible with robinhood 3 DB schema
- Upgrade of the old web UI, with the state of the art in terms of web interfaces and charts
- Fine-grained authentication
- Compatibility with newer MySQL versions

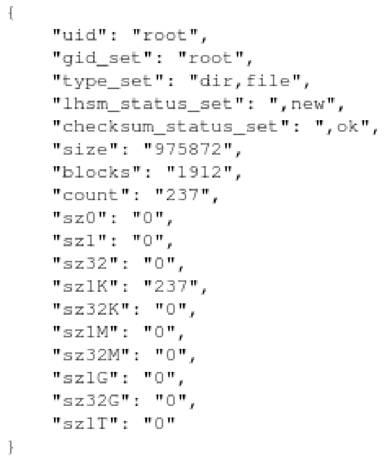


Owner	Empty	1B-32B	32B-1KB	1KB-32KB	32KB-1MB	1MB-32MB	32MB-1GB	1GB-32GB
jenkins	4	0	0	212	722	0	0	0

Nice new features since last year (5/7)

REST interface (in final 3.0)

- Makes it possible to query robinhood DB through a standard protocol (HTTP)
- 3 possible output format:
  - Classic JSON (key-value) : http://server/api/**native**/...
  - Datatables.js: http://server/api/**data**/...
  - GraphJS: http://server/api/**graph**/...



```
{
        "uid": "root",
        "gid_set": "root",
        "type_set": "dir,file",
        "lhsm_status_set": ",new",
        "checksum_status_set": ",ok",
        "size": "975872",
        "blocks": "1912",
        "count": "237",
        "sz0": "0",
        "sz1": "0",
        "sz32": "0",
        "sz1K": "237",
        "sz32K": "0",
        "sz1M": "0",
        "sz32M": "0",
        "sz1G": "0",
        "sz32G": "0",
        "sz1T": "0"
}
```

- Simple and convenient query language:

  > Returns usage stats about all users and status (as JSON)
  http://rbh/api/native/acct/...

  > Returns usage stats about a given user (as JSON)
  http://rbh/api/native/acct/**uid.filter/foo**

  Advanced querying. Example: split user's info by gid
  http://rbh/api/native/acct/**uid.filter/foo/gid.group**

- Allow querying robinhood stats from scripts, dashboards, …
  - E.g: take usage stats into account for job scheduling

Nice new features since last year (6/7)

Features for Lustre/HSM [teaser]

- Enriched “lhsm_remove” policy
- Undelete
- UUID support
- archive_id support
- Passing custom parameters to copytools
- Generic command-based copytool

=> Details in Henri's talk

Nice new features since last year (7/7)

Other minor features (unsorted)

- “robinhood” with no option => run nothing!
- RHEL7 systemd service: `robinhood@<fsname>`
  - Related parameters in `/etc/sysconfig/robinhood.<fsname>`
- Option to store users/groups as the numeric uid/gid
  - `general::uid_gid_as_numbers = yes;`
  - Also reported as numbers in reports.
  - Users/groups must be matched as numeric values in policies e.g. `owner == 1234` (This won't work: `owner == “foo*”`)
- Store ctime in the db (new policy criteria: last_mdchange)
- Option to store last_access as real posix *atime* instead of max(*atime*, *mtime*)
  - `general::last_access_only_atime = yes;`
- New rbh-find criteria: `-ctime`, `-iname`, `-class`

Future plans

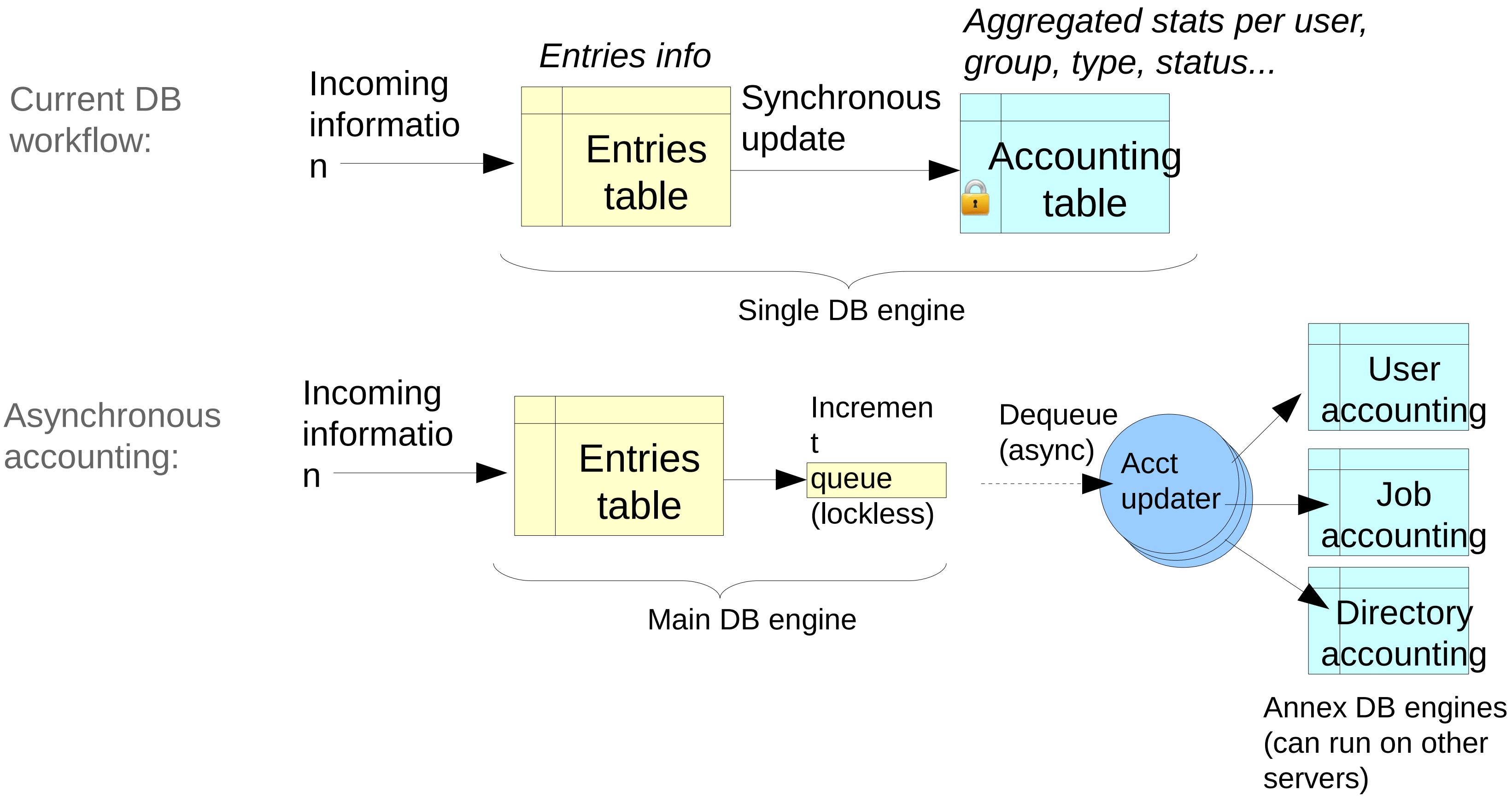
Next Plans

Next candidate features (3.1 to 3.x...)

- Improved Lustre/HSM workload
  - Cf. Henri's talk
- New policy templates:
  - Pool-to-pool migration
  - Managing a “trash” directory
  - Yours are welcome!
- Plugin'ify everything (triggers, DB, ...)
- Support Postgresql DB
  - To take advantage of its sharding features for scalability
- Performance improvements
  - V3.0 mainly focused on new features.
    There may be some room for optimization.
- POSIX: use VFS handles
- Support some object stores

Next Plans: Asynchronous Accounting

- Asynchronous accounting
  - Goal: reduce the impact of accounting on ingest rate.
  - Make it possible to distribute the accounting processing and its DB.



Current DB workflow:

Incoming information

Entries info

Entries table

Synchronous update

Aggregated stats per user, group, type, status...

Accounting table

Single DB engine

Asynchronous accounting:

Incoming information

Entries table

Increment queue (lockless)

Main DB engine

Dequeue (async)

Acct updater

User accounting

Job accounting

Directory accounting

Annex DB engines (can run on other servers)

Next Plans (cont'd)

...this will make possible:

- New aggregated stats:
  - Stats per job (based on JOBID)
  - Metadata accounting
  - Overall stats per top-level directory

WIP in Lustre

Bulk MDT scans

- Faster than POSIX namespace scan

New Kernel-Userland communications

- Optimize changelog streaming

Thanks for your attention!

Questions?

Commissariat à l'énergie atomique et aux énergies alternatives
CEA / DAM Ile-de-France| Bruyères-le-Châtel - 91297 Arpajon Cedex
T. +33 (0)1 69 26 40 00

DAM Île-de-France

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019