DE LA RECHERCHE À L'INDUSTRIE

RobinHood v4 progress report

September 27th 2022

Sebastien Gougeaud

No news, good news?

- The last update at LAD was two years ago
- RobinHood v4 is still in development by the CEA and former members of the team
- Developments can be followed on gerrithub and github (`https://github.com/cea-hpc/librobinhood`)
- Need for having it soon in production on our clusters

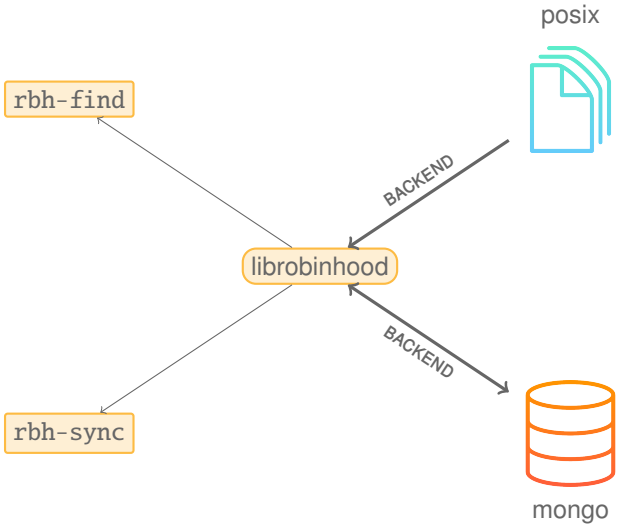
Focus of RobinHood

Providing efficient and easy to use means to replicate and query any filesystem's metadata

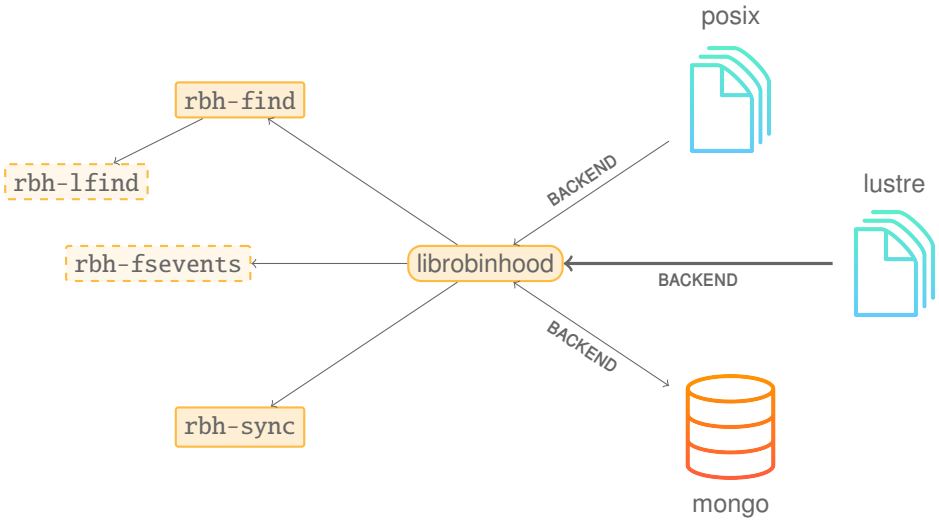
Why a version 4 for RobinHood?

	version 3	version 4
scale up	SQL paradigm *MariaDB*	NoSQL paradigm *MongoDB*
code genericity	software specialised for Lustre filesystems	generic tools calling specific backends
inclusion to Linux repositories	expert system	library of features/applications
code refactoring	heavy code caused by Lustre behavior evolution	clean design to better correspond to current filesystems

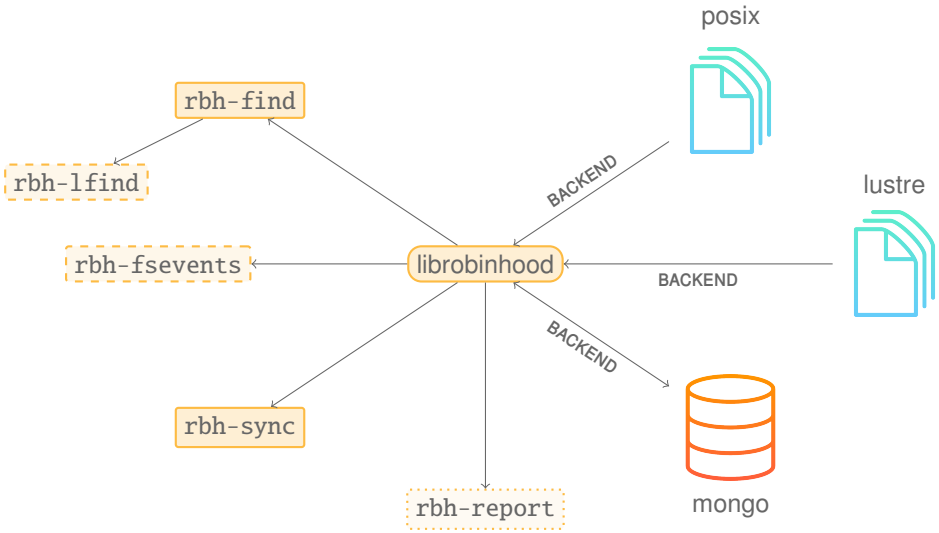
RobinHood components – ready to use



RobinHood components – in development



RobinHood components – future sight



Back to 2020

- Support of custom extended attributes
  - Attach key-value attributes to entries
  - For instance: the path, the name, Lustre attributes
- Conversion between RBH FSEvents and YAML
  - FSEvent stands for File System Event: file creation, metadata change, etc.
  - Handling a new backend only need to convert specific log events into FSEvents

Additions to `rbh-find`

- Filter options are added as long as they may be needed to fetch metadata or define data placement policies:
  - Size – `-size`
  - Permissions – `-perm`
  - User-defined extended attributes – `-xattr`
- Entries may be sorted using `-sort`/`-rsort` options
- Actions on entries are added:
  - Listing – `-ls`
  - Printing to files – `-fprint`/`-fprint0`/`-fls`

Additions to `rbh-find` – demonstration

```
# rbh-sync rbh:posix:/mnt/lustre rbh:mongo:test
# rbh-find rbh:mongo:test -type f -sort name
/big
/huge
/small

# rbh-find rbh:mongo:test -type f -sort size
/small
/big
/huge
```

```
# rbh-find rbh:mongo:test -type f -ls
144...292      4 -rw-r--r-- 1 root root        17 Sep 13 14:00 /small
144...293   1024 -rw-r--r-- 1 root root   1048576 Sep 13 14:10 /big
144...294 284368 -rw-r--r-- 1 root root 291188736 Sep 13 14:11 /huge

# rbh-find rbh:mongo:test -type f -rsort size -ls
144...294 284368 -rw-r--r-- 1 root root 291188736 Sep 13 14:11 /huge
144...293   1024 -rw-r--r-- 1 root root   1048576 Sep 13 14:10 /big
144...292      4 -rw-r--r-- 1 root root        17 Sep 13 14:00 /small

# rbh-find rbh:mongo:test -type f -sort size -ls
144...292      4 -rw-r--r-- 1 root root        17 Sep 13 14:00 /small
144...293   1024 -rw-r--r-- 1 root root   1048576 Sep 13 14:10 /big
144...294 284368 -rw-r--r-- 1 root root 291188736 Sep 13 14:11 /huge

# rbh-find rbh:mongo:test -type f -sort name -ls
144...293   1024 -rw-r--r-- 1 root root   1048576 Sep 13 14:10 /big
144...294 284368 -rw-r--r-- 1 root root 291188736 Sep 13 14:11 /huge
144...292      4 -rw-r--r-- 1 root root        17 Sep 13 14:00 /small
```

Lustre backend

- Lustre is the file system used in our clusters
- Storing lustre relevant information in the mirror backend, such as:
  - FID
  - File layout
  - OST
  - HSM state
- Those information are retrieved during the synchronization and converted to namespace attributes of the entry stored in the database backend
- Filtering entries using those information
- Overload of `rbh-find` tool to add lustre-related options

Lustre backend – demonstration (1/2)

```
# mkdir /mnt/lustre/dir-{0..15}
# touch /mnt/lustre/dir-{0..15}/file-{0..15}
# for i in {0..5}; do
> lfs hsm_set --archived /mnt/lustre/dir-${i}/file-12
> lfs hsm_set --norelease /mnt/lustre/dir-${i}/file-11
> done
# rbh-sync rbh:lustre:/mnt/lustre rbh:mongo:test
# rbh-find rbh:mongo:test | head
/dir-14/file-13
/dir-14/file-14
/dir-14/file-15
/dir-0
/dir-15/file-0
/dir-1
/dir-15/file-1
/dir-2
/dir-15/file-2
/dir-3
# mongo --quiet --eval "db.entries.find({\"statx.type\": 32768})" | head -n 1
{ "_id" : BinData(0,"lwA...AAA"), "statx" : { "blksize" : 4194304, "nlink" : 1, "uid" : 0, "gid" : 0, "type" : 32768,
  "mode" : 420, "ino" : NumberLong("14411205272502451"), "size" : NumberLong(0), "blocks" : NumberLong(0),
  "attributes" : { "immutable" : false, "append" : false }, "atime" : { "sec" : NumberLong(1663145146), "nsec" : 0 },
  "btime" : { "sec" : NumberLong(1663145146), "nsec" : 0 }, "ctime" : { "sec" : NumberLong(1663145146), "nsec" : 0 },
  "mtime" : { "sec" : NumberLong(1663145146), "nsec" : 0 }, "rdev" : { "major" : 0, "minor" : 0 }, "dev" : { "major" : 1273,
  "minor" : 181606 } }, "xattrs" : { "lustre" : { "lov" : BinData(0,"0Av...AAA=") }, "trusted" : { "link" : BinData(0,"3/H...S0w"),
  "lma" : BinData(0,"AAA...AAA"), "lov" : BinData(0,"0Av...AA="), "som" : BinData(0,"BAA...AAA") } },
  "ns" : [ { "parent" : BinData(0,"lwA...AAA"), "name" : "file-0", "xattrs" : { "path" : "/dir-10/file-0",
  "fid" : BinData(0,"AQQ...A=="), "hsm_state" : 0, "hsm_archive_id" : 0, "flags" : 0, "magic" : "LOV_USER_MAGIC_V1", "gen" : 0,
  "stripe_count" : [ ], "stripe_size" : [ ], "pattern" : [ ], "comp_flags" : [ ], "pool" : [ ],
  "ost" : [ NumberLong(0) ], "mdt_index" : 0 } } ] }
```

Lustre backend – demonstration (2/2)

```
# rbh-lfind rbh:mongo:test -sort name -ost 0 | head
/dir-15/file-0
/dir-0/file-0
/dir-1/file-0
/dir-2/file-0
/dir-3/file-0
/dir-4/file-0
/dir-5/file-0
/dir-6/file-0
/dir-7/file-0
/dir-8/file-0

# rbh-lfind rbh:mongo:test -sort name -ost 1 | head
/dir-15/file-1
/dir-0/file-1
/dir-1/file-1
/dir-2/file-1
/dir-3/file-1
/dir-4/file-1
/dir-5/file-1
/dir-6/file-1
/dir-7/file-1
/dir-8/file-1
```

```
# lfs path2fid /mnt/lustre/dir-0/file-0
[0x200000401:0x13:0x0]
# rbh-lfind rbh:mongo:test -fid [0x200000401:0x13:0x0]
/dir-0/file-0

# rbh-lfind rbh:mongo:test -hsm-state archived
/dir-0/file-12
/dir-1/file-12
/dir-2/file-12
/dir-3/file-12
/dir-4/file-12
/dir-5/file-12
# rbh-lfind rbh:mongo:test -hsm-state norelease
/dir-0/file-11
/dir-1/file-11
/dir-2/file-11
/dir-3/file-11
/dir-4/file-11
/dir-5/file-11
```

Lustre changelog reader

- First events are processed:
  - `CREAT`
  - `ATIME/CTIME/MTIME`
  - `CLOSE`
- Others are following
- Still thinking about the right moment to acknowledge the events:
  - When the changelog events are converted to FSEvents?
  - When the events are recorded in the backend?

Lustre changelog reader – demonstration

```
# touch /mnt/lustre/old
# lctl --device lustre-MDT0000 changelog_register
# touch -a /mnt/lustre/old
# touch /mnt/lustre/new
# lfs changelog lustre-MDT0000
1 18CTIME 10:45:21.574804672 2022.09.13 0x5
 t=[0x200000401:0x12:0x0] j=touch.0 ef=0xf u=0:0 nid=0@lo
2 11CLOSE 10:45:21.588901555 2022.09.13 0x42
 t=[0x200000401:0x12:0x0] j=touch.0 ef=0xf u=0:0 nid=0@lo
3 01CREAT 10:45:25.915058526 2022.09.13 0x0
 t=[0x200000401:0x13:0x0] j=touch.0 ef=0xf u=0:0 nid=0@lo
 p=[0x200000007:0x1:0x0] new
4 11CLOSE 10:45:25.922082667 2022.09.13 0x42
 t=[0x200000401:0x13:0x0] j=touch.0 ef=0xf u=0:0 nid=0@lo
# rbh-fsevents --enrich /mnt/lustre lustre-MDT0000 -
--- !upsert
"id": !!binary lwAAAAEEAAACAAAAEgAAAAAAAAAAAAAAAAAAAAAA
"xattrs": {}
"statx":
  "atime":
    "sec": 1663065921
    "nsec": 0
  "ctime":
    "sec": 1663065921
    "nsec": 0
...
--- !upsert
"id": !!binary lwAAAAEEAAACAAAAEgAAAAAAAAAAAAAAAAAAAAAA
"xattrs": {}
"statx":
  "atime":
    "sec": 1663065921
    "nsec": 0
...
```

```
--- !upsert
"id": !!binary lwAAAAEEAAACAAAAEwAAAAAAAAAAAAAAAAAAAAAA
"xattrs":
  "ns":
  - "parent": !!binary lwAAAAcAAAACAAAAAQAAAAAAAAAAAAAAAAAAAAAA
    "name": "new"
    "xattrs":
      "path": "new"
      "fid": !!binary AQQAAAIAAAATAAAAAAAAAA==
"statx":
  "type": "file"
  "uid": 0
  "gid": 0
  "atime":
    "sec": 1663065925
    "nsec": 0
  [...]
  "blksize": 4194304
  "attributes":
    "immutable": n
    "append": n
  "rdev":
    "major": 0
    "minor": 0
  "dev":
    "major": 1273
    "minor": 181606
...
--- !upsert
"id": !!binary lwAAAAEEAAACAAAAEwAAAAAAAAAAAAAAAAAAAAAA
"xattrs": {}
"statx":
  "atime":
    "sec": 1663065925
    "nsec": 0
...
```

What's next?

- Development of a non-lustre related backend, for the IO-SEA european project, using CORTX-Motr and Phobos as object store
- FSEvent deduplication
- Pre-production tests on our systems for a production during 2023

Repositories

project	repository	
librobinhood	https://github.com/cea-hpc/	librobinhood
rbh-find		rbh-find
rbh-find-lustre		rbh-find-lustre
rbh-sync		rbh-sync
rbh-fsevents		rbh-fsevents

Thanks for your attention

RobinHood v4 progress report
Sebastien Gougeaud

Commissariat à l'énergie atomique et aux énergies alternatives
Centre de Bruyères-le-Châtel | 91297 Arpajon Cedex
T. +33 (0)1 69 26 40 00 | F. +33 (0)1 69 26 40 00
Établissement public à caractère industriel et commercial – RCS Paris B 775 685 019