RADOS - Bug #21388

inconsistent pg but repair does nothing reporting head data_digest != data_digest from auth oi / hopefully data seems ok

09/14/2017 01:37 PM - Laurent GUERBY

Status:	Duplicate	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:	David Zafman		
Category:			
Target version:			
Source:	Community (user)	**Affected Versions:**	v10.2.9
Tags:		**ceph-qa-suite:**	
Backport:	jewel, luminous	**Component(RADOS):**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	
Reviewed:			

Description

ceph pg repair is currently not fixing three "inconsistent" objects
on one of our pg on a replica 3 pool.

For all three objects the 3 replica data objets are identical
(we checked them on disk on the 3 OSD),
the error says "head data_digest != data_digest from auth oi",
see below.

The data in question are used on rbd volumes from KVM and we did a read
from /dev/sdX at the right place on the VM and got a good looking
result : text file, uncorrupted according to our user, so data
currently returned by ceph and replicated 3 times seems fine.

Now the question is how to tell ceph that the replica data is correct
so that the inconsistent message disappears?

We're thinking of doing manually rados get/put but may be this is not
a good idea or there is another way.

May be ceph should handle this via a "force-repair-identical-replica" option
or something similar if automatic "background" repair seems risky in this
case.

1. ceph --version
   ceph version 10.2.9 (2ee413f77150c0f375ff6f10edd6c8f9c7d060d0)
2. ceph health detail
   HEALTH_ERR 1 pgs inconsistent; 9 scrub errors;
   pg 58.6c1 is active+clean+inconsistent, acting [46,44,19]
   ...
3. rados list-inconsistent-obj 58.6c1 --format=json-pretty {
   "epoch": 277681,
   "inconsistents": []
   }
4. ceph pg repair 58.6c1

... osd 46 /var/log :

shard 19: soid 58:83772424:::rbd_data.30fce9e39dad7a6.000000000007f027:head data_digest 0x783cd2c5 != data_digest 0x501f846c from auth oi 58:83772424:::rbd_data.30fce9e39dad7a6.000000000007f027:head(252707'3100507 osd.2.0:710926 dirty|data_digest|omap_digest s 4194304 uv 3481755 dd 501f846c od ffffffff alloc_hint [0 0])
shard 44: soid 58:83772424:::rbd_data.30fce9e39dad7a6.000000000007f027:head data_digest 0x783cd2c5 != data_digest

0x501f846c from auth oi 58:83772424:::rbd_data.30fce9e39dad7a6.000000000007f027:head(252707'3100507 osd.2.0:710926 dirty|data_digest|omap_digest s 4194304 uv 3481755 dd 501f846c od ffffffff alloc_hint [0 0])
shard 46: soid 58:83772424:::rbd_data.30fce9e39dad7a6.000000000007f027:head data_digest 0x783cd2c5 != data_digest 0x501f846c from auth oi 58:83772424:::rbd_data.30fce9e39dad7a6.000000000007f027:head(252707'3100507 osd.2.0:710926 dirty|data_digest|omap_digest s 4194304 uv 3481755 dd 501f846c od ffffffff alloc_hint [0 0])
soid 58:83772424:::rbd_data.30fce9e39dad7a6.000000000007f027:head: failed to pick suitable auth object
shard 19: soid 58:83772d9e:::rbd_data.68cb7f74b0dc51.000000000000181e:head data_digest 0xd8f6895a != data_digest 0x4edc70a3 from auth oi 58:83772d9e:::rbd_data.68cb7f74b0dc51.000000000000181e:head(77394'2047065 osd.16.0:4500125 dirty|data_digest|omap_digest s 4194304 uv 1895034 dd 4edc70a3 od ffffffff alloc_hint [0 0])
shard 44: soid 58:83772d9e:::rbd_data.68cb7f74b0dc51.000000000000181e:head data_digest 0xd8f6895a != data_digest 0x4edc70a3 from auth oi 58:83772d9e:::rbd_data.68cb7f74b0dc51.000000000000181e:head(77394'2047065 osd.16.0:4500125 dirty|data_digest|omap_digest s 4194304 uv 1895034 dd 4edc70a3 od ffffffff alloc_hint [0 0])
shard 46: soid 58:83772d9e:::rbd_data.68cb7f74b0dc51.000000000000181e:head data_digest 0xd8f6895a != data_digest 0x4edc70a3 from auth oi 58:83772d9e:::rbd_data.68cb7f74b0dc51.000000000000181e:head(77394'2047065 osd.16.0:4500125 dirty|data_digest|omap_digest s 4194304 uv 1895034 dd 4edc70a3 od ffffffff alloc_hint [0 0])
soid 58:83772d9e:::rbd_data.68cb7f74b0dc51.000000000000181e:head: failed to pick suitable auth object
shard 19: soid 58:8377bf9a:::rbd_data.2ef7e1a528b30ea.00000000000254f6:head data_digest 0xdf8916bf != data_digest 0x47b79db8 from auth oi 58:8377bf9a:::rbd_data.2ef7e1a528b30ea.00000000000254f6:head(252707'3100535 osd.2.0:710954 dirty|data_digest|omap_digest s 4194304 uv 3298154 dd 47b79db8 od ffffffff alloc_hint [0 0])
shard 44: soid 58:8377bf9a:::rbd_data.2ef7e1a528b30ea.00000000000254f6:head data_digest 0xdf8916bf != data_digest 0x47b79db8 from auth oi 58:8377bf9a:::rbd_data.2ef7e1a528b30ea.00000000000254f6:head(252707'3100535 osd.2.0:710954 dirty|data_digest|omap_digest s 4194304 uv 3298154 dd 47b79db8 od ffffffff alloc_hint [0 0])
shard 46: soid 58:8377bf9a:::rbd_data.2ef7e1a528b30ea.00000000000254f6:head data_digest 0xdf8916bf != data_digest 0x47b79db8 from auth oi 58:8377bf9a:::rbd_data.2ef7e1a528b30ea.00000000000254f6:head(252707'3100535 osd.2.0:710954 dirty|data_digest|omap_digest s 4194304 uv 3298154 dd 47b79db8 od ffffffff alloc_hint [0 0])
soid 58:8377bf9a:::rbd_data.2ef7e1a528b30ea.00000000000254f6:head: failed to pick suitable auth object
repair 9 errors, 0 fixed

Related issues:

Duplicates RADOS - Feature #25085: Allow repair of an object with a bad data_...	**Resolved**	**07/24/2018**

History

#1 - 09/14/2017 03:35 PM - David Zafman

- Status changed from New to 12

- Assignee set to David Zafman

- Backport set to jewel, luminous

Doing the following should produce list-inconsistent-obj information:

$ ceph pg deep-scrub 58.6c1
(Wait for scrub to finish)
$ rados list-inconsistent-obj 58.6c1 --format=json-pretty

This is a very particular scenario which if identified should be safe to repair. In this case the automatic repair rejects all copies because none match the selected_object_info thus setting data_digest_mismatch_oi on all shards.

Requirements:

1. data_digest_mismatch_oi is set on all shards make it unrepairable
2. union_shard_errors has only data_digest_mismatch_oi listed, no other issues involved
3. Object "errors" is empty { "inconsistent": [{ ..."errors": []....}] } which means the data_digest value is the same on all shards (0x2d4a11c2 example below)
4. No down OSDs which might have different/correct data

Rados get/put is the solution to fix these objects, followed by a deep-scrub to clear the "inconsistent" pg state.

Here is an example of what this scenario looks like:


```
{
  "inconsistents": [
    {
      "shards": [
        {
          "data_digest": "0x2d4a11c2",
          "omap_digest": "0xf5fba2c6",
          "size": 143456,
          "errors": [
```

```
            "data_digest_mismatch_oi"
          ],
          "osd": 0,
          "primary": true
        },
        {
          "data_digest": "0x2d4a11c2",
          "omap_digest": "0xf5fba2c6",
          "size": 143456,
          "errors": [
            "data_digest_mismatch_oi"
          ],
          "osd": 1,
          "primary": false
        },
        {
          "data_digest": "0x2d4a11c2",
          "omap_digest": "0xf5fba2c6",
          "size": 143456,
          "errors": [
            "data_digest_mismatch_oi"
          ],
          "osd": 2,
          "primary": false
        }
      ],
      "selected_object_info": "3:ce3f1d6a::: mytestobject:head(47'54 osd.0.0:53 dirty|omap|data_digest|omap_di
gest s 143456 uv 3 dd 2ddbf8f5 od f5fba2c6 alloc_hint [0 0 0])",
      "union_shard_errors": [
        "data_digest_mismatch_oi"
      ],
      "errors": [
      ],
      "object": {
        "version": 3,
        "snap": "head",
        "locator": "",
        "nspace": "",
        "name": "mytestobject"
      }
    }
  ],
  "epoch": 103443
}
```

#2 - 09/14/2017 10:18 PM - Mehdi Abaakouk

After the deep-scrub, "rados get" give us some errors:

1. rados list-inconsistent-obj 58.6c1 --format=json-pretty > inconsistent
2. for i in $(cat inconsistent | jq .inconsistents[].object.name -r); do rados -p disks get $i $i ; done
   error getting disks/rbd_data.2ef7e1a528b30ea.00000000000254f6: (5) Input/output error
   error getting disks/rbd_data.30fce9e39dad7a6.000000000007f027: (5) Input/output error
   error getting disks/rbd_data.68cb7f74b0dc51.000000000000181e: (5) Input/output error

So I directly take the objects on the osd itself.

And put it back with rados.
The I issue "ceph pg deep-scrub 58.6c1" again

The inconsistent status have been clearer.

Thanks for the help

For reference an example of object state after the first deep-scrub


```
        {
            "object": {
                "name": "rbd_data.68cb7f74b0dc51.000000000000181e",
                "nspace": "",
                "locator": "",
                "snap": "head",
                "version": 1895034
            },
            "errors": [],
            "union_shard_errors": [
                "data_digest_mismatch_oi"
            ],
            "selected_object_info": "58:83772d9e:::rbd_data.68cb7f74b0dc51.000000000000181e:head(77394'2047065
 osd.16.0:4500125 dirty|data_digest|omap_digest s 4194304 uv 1895034 dd 4edc70a3 od ffffffff alloc_hint [0 0])
",
            "shards": [
                {
                    "osd": 19,
                    "errors": [
                        "data_digest_mismatch_oi"
                    ],
                    "size": 4194304,
                    "omap_digest": "0xffffffff",
                    "data_digest": "0xd8f6895a"
                },
                {
                    "osd": 44,
                    "errors": [
                        "data_digest_mismatch_oi"
                    ],
                    "size": 4194304,
                    "omap_digest": "0xffffffff",
                    "data_digest": "0xd8f6895a"
                },
                {
                    "osd": 46,
                    "errors": [
                        "data_digest_mismatch_oi"
                    ],
                    "size": 4194304,
                    "omap_digest": "0xffffffff",
                    "data_digest": "0xd8f6895a"
                }
            ]
        }
```

#3 - 09/14/2017 10:20 PM - David Zafman

I forgot the mention that on get you have to prevent the code from checking the digest by doing reads that are smaller than the total size of the object. That would have avoided the EIO errors.

#4 - 09/15/2017 08:11 AM - Laurent GUERBY

We did an scp of the OSD object file then a rados put and deep-scrub got it back to OK.

Thanks for your quick answer!

#5 - 10/25/2017 10:19 AM - wei jin

I also ran into the same issue and fixed it by rados get/put command.
The inconsistent info is not exactly the same but similar:


```
{
    "epoch": 2373,
    "inconsistents": [
        {
            "object": {
                "name": "1000003528d.00000058",
                "nspace": "fsvolumens_87c46348-9869-11e7-8525-3497f65a8415",
                "locator": "",
                "snap": "head",
                "version": 147490
            },
            "errors": [],
            "union_shard_errors": [
                "size_mismatch_oi"
            ],
            "selected_object_info": "
1:b3f61048:fsvolumens_87c46348-9869-11e7-8525-3497f65a8415::1000003528d.00000058:head(2401'147490 client.90154
9.1:33749 dirty|omap_digest s 3461120 uv 147490 od ffffffff alloc_hint [0 0])",
            "shards": [
                {
                    "osd": 27,
                    "errors": [
                        "size_mismatch_oi"
                    ],
                    "size": 0,
                    "omap_digest": "0xffffffff",
                    "data_digest": "0xffffffff"
                },
                {
                    "osd": 62,
                    "errors": [
                        "size_mismatch_oi"
                    ],
                    "size": 0,
                    "omap_digest": "0xffffffff",
                    "data_digest": "0xffffffff"
                },
                {
                    "osd": 133,
                    "errors": [
                        "size_mismatch_oi"
                    ],
                    "size": 0,
                    "omap_digest": "0xffffffff",
                    "data_digest": "0xffffffff"
                }
            ]
        }
    ]
}
```

#6 - 10/25/2017 06:22 PM - David Zafman

Wei Jin,

According to the inconsistent output, the object info said the object size is 3461120, but all the shards are 0 length.

The osd log on the primary must say "failed to pick suitable auth object" for this object. Unless you can find this object on some other disk like on a down OSD, you can remove the object or write 0 bytes to create it as an empty object. Either will make the scrub error go away.

I different bug would have to be filed about the corruption itself. If you have a production system with default logging, there may not be much to go on. If this happens again you could turn up the osd and fs log levels and create a tracker with logs attached.

#7 - 11/07/2017 09:11 PM - David Zafman

In a case where an objects snapshot is seeing an error, we can't use rados to get and put. Use the following procedure to repair the objects snapshot.


```
{
  "inconsistents": [
    {
      "shards": [
        {
          "data_digest": "0x2d4a11c2",
          "omap_digest": "0xf5fba2c6",
          "size": 143456,
          "errors": [
            "data_digest_mismatch_oi"
          ],
          "osd": 0,
          "primary": true
        },
        {
          "data_digest": "0x2d4a11c2",
          "omap_digest": "0xf5fba2c6",
          "size": 143456,
          "errors": [
            "data_digest_mismatch_oi"
          ],
          "osd": 1,
          "primary": false
        },
        {
          "data_digest": "0x2d4a11c2",
          "omap_digest": "0xf5fba2c6",
          "size": 143456,
          "errors": [
            "data_digest_mismatch_oi"
          ],
          "osd": 2,
          "primary": false
        }
      ],
      "selected_object_info": "3:ce3f1d6a::: mytestobject:head(47'54 osd.0.0:53 dirty|omap|data_digest|omap_di
gest s 143456 uv 3 dd 2ddbf8f5 od f5fba2c6 alloc_hint [0 0 0])",
      "union_shard_errors": [
        "data_digest_mismatch_oi"
      ],
      "errors": [
      ],
      "object": {
        "version": 3,
        "snap": "1",
        "locator": "",
        "nspace": "",
        "name": "mytestobject"
      }
    }
```

```
  ],
  "epoch": 103443
}

# Install binary editor
sudo apt-get install ghex

ceph osd set noout
stop osd.0, osd.1, osd.2

ceph-objectstore-tool --data-path ....osd-0 --op list  --pgid 1.0 mytestobject
["1.0",{"oid":"mytestobject","key":"","snapid":1,"hash":3016456255,"max":0,"pool":1,"namespace":"","max":0}]
["1.0",{"oid":"mytestobject","key":"","snapid":-2,"hash":3016456255,"max":0,"pool":1,"namespace":"","max":0}]
        We use the corresponding snapid that has the problem in this case snapid 1 in the commands below

ceph-objectstore-tool --data-path ….osd-0 '["1.0",{"oid":"mytestobject","key":"","snapid": 1,"hash":30164562
55,"max":0,"pool":1,"namespace":"","max":0}]' get-attr _ > info.data
ceph-dencoder import info.data type object_info_t decode dump_json
    Should see "data_digest": 769390837 in the json output (decimal for 0x2ddbf8f5)
ghex info.data
     Change F5 F8 DB 2D to C2 11 4A 2D
    Save file
ceph-dencoder import info.data type object_info_t decode dump_json
    Should now see "data_digest": 759828930  in the json output (decimal for 0x2D4A11C2)

ceph-objectstore-tool --data-path ….osd-0 '["1.0",{"oid":"mytestobject","key":"","snapid": 1,"hash":3016456255
,"max":0,"pool":1,"namespace":"","max":0}]' set-attr _ info.data
ceph-objectstore-tool --data-path ….osd-1 '["1.0",{"oid":"mytestobject","key":"","snapid": 1,"hash":3016456255
,"max":0,"pool":1,"namespace":"","max":0}]' set-attr _ info.data
ceph-objectstore-tool --data-path ….osd-2 '["1.0",{"oid":"mytestobject","key":"","snapid": 1,"hash":3016456255
,"max":0,"pool":1,"namespace":"","max":0}]' set-attr _  info.data

restart osd.0, osd.1, osd.2
ceph pg deep-scrub 1.0
    Should see pg go active+clean
ceph osd unset noout
```

#8 - 11/07/2017 09:21 PM - David Zafman

We should add code to be_select_auth_object() that checks that data_digest and omap_digest. If we do that then if any 1 object_info_t matches its data, it will be selected and repair will work. In these cases all object_info_t are broken so that change won't help.

#9 - 12/22/2017 05:33 AM - Ryan Anstey

I'm also having this issue. I'm getting new scrub errors every few days. No idea what's going on. This is something new since my upgrade from Jewel. How do I prevent the scrub errors from piling up?

#10 - 12/31/2017 11:13 PM - Ryan Anstey

I'm working on fixing all my inconsistent pgs but I'm having issues with rados get... hopefully I'm just doing the command wrong somehow?

```
# bad_object=rb.0.854e.238e1f29.000000281376
# rados -p rbd stat ${bad_object}
rbd/rb.0.854e.238e1f29.000000281376 mtime 2017-07-31 03:48:39.000000, size 4194304
# rados -p rbd get ${bad_object} /tmp/tmpobj
error getting rbd/rb.0.854e.238e1f29.000000281376: (5) Input/output error
```

#11 - 01/17/2018 10:26 PM - Brian Andrus

Ryan Anstey wrote:

> I'm working on fixing all my inconsistent pgs but I'm having issues with rados get... hopefully I'm just doing the command wrong somehow?
>
> [...]

We had this same issue with I/O errors. I used ceph-objectstore-tool get-bytes to get the object directly from a stopped OSD and rados put that output. YMMV, I just emailed the mailing list to welcome discussion on the viability of that as a fix.

#12 - 02/22/2018 01:14 PM - Yoann Moulin

Hello,

I'm also having this issue

```
$ ceph health detail
HEALTH_ERR 3 scrub errors; Possible data damage: 1 pg inconsistent
OSD_SCRUB_ERRORS 3 scrub errors
PG_DAMAGED Possible data damage: 1 pg inconsistent
    pg 11.5f is active+clean+inconsistent, acting [78,154,170]
```

```
$ ceph -s
  cluster:
    id:     f9dfd27f-c704-4d53-9aa0-4a23d655c7c4
    health: HEALTH_ERR
```

```
            3 scrub errors
            Possible data damage: 1 pg inconsistent

  services:
    mon: 3 daemons, quorum iccluster002.iccluster.epfl.ch,iccluster010.iccluster.epfl.ch,iccluster018.iccluste
r.epfl.ch
    mgr: iccluster001(active), standbys: iccluster009, iccluster017
    mds: cephfs-3/3/3 up  {0=iccluster022.iccluster.epfl.ch=up:active,1=iccluster006.iccluster.epfl.ch=up:acti
ve,2=iccluster014.iccluster.epfl.ch=up:active}
    osd: 180 osds: 180 up, 180 in
    rgw: 6 daemons active

  data:
    pools:   29 pools, 10432 pgs
    objects: 82862k objects, 171 TB
    usage:   515 TB used, 465 TB / 980 TB avail
    pgs:     10425 active+clean
             6     active+clean+scrubbing+deep
             1     active+clean+inconsistent

  io:
    client:   21538 B/s wr, 0 op/s rd, 33 op/s wr



ceph version 12.2.2 (cf0baeeeeba3b47f9427c6c97e2144b094b7e5ba) luminous (stable)
```

Short log :

```
2018-02-21 09:08:33.408396 7fb7b8222700  0 log_channel(cluster) log [DBG] : 11.5f repair starts
2018-02-21 09:08:33.727277 7fb7b8222700 -1 log_channel(cluster) log [ERR] : 11.5f shard 78: soid 11:fb71fe10::
:.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19:head omap_digest 0x29fdd712 != omap_digest 0xd46bb5a1 fro
m auth oi 11:fb71fe10:::.dir.c9724aff-5fa0-4dd9- b494-57bdb48fab4e.314528.19:head(98394'20014544 osd.78.0:1623
704 dirty|omap|data_digest|omap_digest s 0 uv 20014543 dd ffffffff od d46bb5a1 alloc_hint [0 0 0])
2018-02-21 09:08:33.727290 7fb7b8222700 -1 log_channel(cluster) log [ERR] : 11.5f shard 154: soid 11:fb71fe10:
::.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19:head omap_digest 0x29fdd712 != omap_digest 0xd46bb5a1 fr
om auth oi 11:fb71fe10:::.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19:head(98394'20014544 osd.78.0:1623
704 dirty|omap|data_digest|omap_digest s 0 uv 20014543 dd ffffffff od d46bb5a1 alloc_hint [0 0 0])
2018-02-21 09:08:33.727293 7fb7b8222700 -1 log_channel(cluster) log [ERR] : 11.5f shard 170: soid 11:fb71fe10:
::.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19:head omap_digest 0x29fdd712 != omap_digest 0xd46bb5a1 fr
om auth oi 11:fb71fe10:::.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19:head(98394'20014544 osd.78.0:1623
704 dirty|omap|data_digest|omap_digest s 0 uv 20014543 dd ffffffff od d46bb5a1 alloc_hint [0 0 0])
2018-02-21 09:08:33.727295 7fb7b8222700 -1 log_channel(cluster) log [ERR] : 11.5f soid 11:fb71fe10:::.dir.c972
4aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19:head: failed to pick suitable auth object
2018-02-21 09:08:33.727333 7fb7b8222700 -1 log_channel(cluster) log [ERR] : 11.5f repair 3 errors, 0 fixed
```

I set "debug_osd 20/20" on osd.78 and start the repair again, the log file is here :

ceph-post-file: 1ccac8ea-0947-4fe4-90b1-32d1048548f1

```
rados list-inconsistent-obj 11.5f --format=json-pretty
{
    "epoch": 118997,
    "inconsistents": [
        {
            "object": {
                "name": ".dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19",
                "nspace": "",
                "locator": "",
                "snap": "head",
                "version": 20014543
            },
            "errors": [],
            "union_shard_errors": [
                "omap_digest_mismatch_oi"
            ],
            "selected_object_info": "11:fb71fe10:::.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19:head(98
```

```
394'20014544 osd.78.0:1623704 dirty|omap|data_digest|omap_digest s 0 uv 20014543 dd ffffffff od d46bb5a1 alloc
_hint [0 0 0])",
            "shards": [
                {
                    "osd": 78,
                    "primary": true,
                    "errors": [
                        "omap_digest_mismatch_oi"
                    ],
                    "size": 0,
                    "omap_digest": "0x29fdd712",
                    "data_digest": "0xffffffff"
                },
                {
                    "osd": 154,
                    "primary": false,
                    "errors": [
                        "omap_digest_mismatch_oi"
                    ],
                    "size": 0,
                    "omap_digest": "0x29fdd712",
                    "data_digest": "0xffffffff"
                },
                {
                    "osd": 170,
                    "primary": false,
                    "errors": [
                        "omap_digest_mismatch_oi"
                    ],
                    "size": 0,
                    "omap_digest": "0x29fdd712",
                    "data_digest": "0xffffffff"
                }
            ]
        }
    ]
}
```


unfortunatly, I cannot get that object.

```
$ rados -p disks get  .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19 .dir.c9724aff-5fa0-4dd9-b494-57bdb48
fab4e.314528.19
error opening pool disks: (2) No such file or directory
```

even with the 3 OSDs stopped, ceph-objectstore-tool return nothing

```
$ ceph-objectstore-tool --cluster cephprod --data-path /var/lib/ceph/osd/cephprod-78 --op list --pgid 1.0 .dir
.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19
$
```

Am I wrong somewhere ?

Best regards,

Yoann Moulin

#13 - 02/23/2018 03:42 AM - Brad Hubbard

Does "rados -p disks ls" list the object? Can you find the actual storage for this object on the disks used for these osds (or any others)?

That pgid of "1.0" doesn't look right, I'd expect "11.5f". Is "disks" pool 1 or 11?

#14 - 02/23/2018 08:33 AM - Yoann Moulin

Hello Brad,

Sorry I have been too fast,

the rados get with the good pool return a file with size=0

```
$ rados -p default.rgw.buckets.index  get  .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19 .dir.c9724aff-5
fa0-4dd9-b494-57bdb48fab4e.314528.19
$
```

```
$ du .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19
0    .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19
```

and ceph-objectstore-tool with the good pgid return nothing too

```
$ ceph-objectstore-tool --cluster cephprod --data-path /var/lib/ceph/osd/cephprod-78 --journal-path /var/lib/c
eph/osd/cephprod-78/journal --op list --pgid 11.5f  .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19
$
```

Thanks

#15 - 02/26/2018 04:24 AM - Brad Hubbard

Yes, size 0 object is expected since all copies report '"size": 0'.

The discrepancy appears to be in the omap data so I would suggest comparing the omap header and key-value pairs on each of the OSDs using the ceph-objectstore-tool.

#16 - 02/28/2018 08:17 AM - Yoann Moulin

Here the result of the 3 commands for each replicate of the PG, osd.78 on iccluster020 is the one with the error :

iccluster020

```
================================================================================================================
==============
ceph-objectstore-tool --cluster cephprod --data-path /var/lib/ceph/osd/cephprod-78 --pgid 11.5f --journal /var
/lib/ceph/osd/cephprod-78/journal --debug .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8EDF__
b get-omaphdr --op list

2018-02-28 08:10:48.203910 7f22273d2400  0 filestore(/var/lib/ceph/osd/cephprod-78) backend xfs (magic 0x58465
342)
2018-02-28 08:10:48.204429 7f22273d2400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-78) detect_featu
res: FIEMAP ioctl is disabled via 'filestore fiemap' config option
2018-02-28 08:10:48.204446 7f22273d2400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-78) detect_featu
res: SEEK_DATA/SEEK_HOLE is disabled via 'filestore seek data hole' config option
2018-02-28 08:10:48.204448 7f22273d2400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-78) detect_featu
res: splice() is disabled via 'filestore splice' config option
2018-02-28 08:10:48.224226 7f22273d2400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-78) detect_featu
res: syncfs(2) syscall fully supported (by glibc and kernel)
2018-02-28 08:10:48.224334 7f22273d2400  0 xfsfilestorebackend(/var/lib/ceph/osd/cephprod-78) detect_feature:
extsize is disabled by conf
2018-02-28 08:10:48.225208 7f22273d2400  0 filestore(/var/lib/ceph/osd/cephprod-78) start omap initiation
2018-02-28 08:10:48.225302 7f22273d2400  0  set rocksdb option compaction_readahead_size = 2097152
2018-02-28 08:10:48.225319 7f22273d2400  0  set rocksdb option compression = kNoCompression
2018-02-28 08:10:48.225328 7f22273d2400  0  set rocksdb option max_background_compactions = 8
2018-02-28 08:10:48.225374 7f22273d2400  0  set rocksdb option compaction_readahead_size = 2097152
2018-02-28 08:10:48.225384 7f22273d2400  0  set rocksdb option compression = kNoCompression
2018-02-28 08:10:48.225390 7f22273d2400  0  set rocksdb option max_background_compactions = 8
2018-02-28 08:10:48.226247 7f22273d2400  4 rocksdb: RocksDB version: 5.4.0

2018-02-28 08:10:48.226263 7f22273d2400  4 rocksdb: Git sha rocksdb_build_git_sha:@0@
2018-02-28 08:10:48.226265 7f22273d2400  4 rocksdb: Compile date Nov 30 2017
2018-02-28 08:10:48.226267 7f22273d2400  4 rocksdb: DB SUMMARY

2018-02-28 08:10:48.226338 7f22273d2400  4 rocksdb: CURRENT file:  CURRENT

2018-02-28 08:10:48.226341 7f22273d2400  4 rocksdb: IDENTITY file:  IDENTITY

2018-02-28 08:10:48.226347 7f22273d2400  4 rocksdb: MANIFEST file:  MANIFEST-000368 size: 2020 Bytes

2018-02-28 08:10:48.226352 7f22273d2400  4 rocksdb: SST files in /var/lib/ceph/osd/cephprod-78/current/omap di
r, Total Num: 10, files: 000158.sst 000304.sst 000316.sst 000317.sst 000326.sst 000327.sst 000328.sst 000330.s
st 000332.sst

2018-02-28 08:10:48.226356 7f22273d2400  4 rocksdb: Write Ahead Log file in /var/lib/ceph/osd/cephprod-78/curr
ent/omap: 000369.log size: 0 ;

2018-02-28 08:10:48.226358 7f22273d2400  4 rocksdb:                         Options.error_if_exists: 0
2018-02-28 08:10:48.226360 7f22273d2400  4 rocksdb:                       Options.create_if_missing: 1
2018-02-28 08:10:48.226362 7f22273d2400  4 rocksdb:                         Options.paranoid_checks: 1
2018-02-28 08:10:48.226363 7f22273d2400  4 rocksdb:                                     Options.env: 0x55a9b74
8bc80
2018-02-28 08:10:48.226365 7f22273d2400  4 rocksdb:                                Options.info_log: 0x55a9b96
1a500
2018-02-28 08:10:48.226366 7f22273d2400  4 rocksdb:                          Options.max_open_files: -1
2018-02-28 08:10:48.226370 7f22273d2400  4 rocksdb:                Options.max_file_opening_threads: 16
2018-02-28 08:10:48.226371 7f22273d2400  4 rocksdb:                               Options.use_fsync: 0
2018-02-28 08:10:48.226373 7f22273d2400  4 rocksdb:                       Options.max_log_file_size: 0
2018-02-28 08:10:48.226375 7f22273d2400  4 rocksdb:                  Options.max_manifest_file_size: 184467440
73709551615
2018-02-28 08:10:48.226377 7f22273d2400  4 rocksdb:                   Options.log_file_time_to_roll: 0
2018-02-28 08:10:48.226378 7f22273d2400  4 rocksdb:                       Options.keep_log_file_num: 1000
2018-02-28 08:10:48.226379 7f22273d2400  4 rocksdb:                    Options.recycle_log_file_num: 0
2018-02-28 08:10:48.226381 7f22273d2400  4 rocksdb:                         Options.allow_fallocate: 1
2018-02-28 08:10:48.226383 7f22273d2400  4 rocksdb:                        Options.allow_mmap_reads: 0
2018-02-28 08:10:48.226384 7f22273d2400  4 rocksdb:                       Options.allow_mmap_writes: 0
2018-02-28 08:10:48.226385 7f22273d2400  4 rocksdb:                        Options.use_direct_reads: 0
2018-02-28 08:10:48.226387 7f22273d2400  4 rocksdb:                        Options.use_direct_io_for_flush_and
_compaction: 0
2018-02-28 08:10:48.226388 7f22273d2400  4 rocksdb:          Options.create_missing_column_families: 0
2018-02-28 08:10:48.226389 7f22273d2400  4 rocksdb:                                Options.db_log_dir:
2018-02-28 08:10:48.226391 7f22273d2400  4 rocksdb:                                   Options.wal_dir: /var/lib/
```

```
ceph/osd/cephprod-78/current/omap
2018-02-28 08:10:48.226393 7f22273d2400  4 rocksdb:                Options.table_cache_numshardbits: 6
2018-02-28 08:10:48.226394 7f22273d2400  4 rocksdb:                      Options.max_subcompactions: 1
2018-02-28 08:10:48.226395 7f22273d2400  4 rocksdb:                  Options.max_background_flushes: 1
2018-02-28 08:10:48.226408 7f22273d2400  4 rocksdb:                         Options.WAL_ttl_seconds: 0
2018-02-28 08:10:48.226410 7f22273d2400  4 rocksdb:                       Options.WAL_size_limit_MB: 0
2018-02-28 08:10:48.226411 7f22273d2400  4 rocksdb:             Options.manifest_preallocation_size: 4194304
2018-02-28 08:10:48.226412 7f22273d2400  4 rocksdb:                     Options.is_fd_close_on_exec: 1
2018-02-28 08:10:48.226414 7f22273d2400  4 rocksdb:                   Options.advise_random_on_open: 1
2018-02-28 08:10:48.226415 7f22273d2400  4 rocksdb:                    Options.db_write_buffer_size: 0
2018-02-28 08:10:48.226416 7f22273d2400  4 rocksdb:         Options.access_hint_on_compaction_start: 1
2018-02-28 08:10:48.226417 7f22273d2400  4 rocksdb:  Options.new_table_reader_for_compaction_inputs: 1
2018-02-28 08:10:48.226418 7f22273d2400  4 rocksdb:              Options.compaction_readahead_size: 2097152
2018-02-28 08:10:48.226420 7f22273d2400  4 rocksdb:           Options.random_access_max_buffer_size: 1048576
2018-02-28 08:10:48.226421 7f22273d2400  4 rocksdb:           Options.writable_file_max_buffer_size: 1048576
2018-02-28 08:10:48.226422 7f22273d2400  4 rocksdb:                       Options.use_adaptive_mutex: 0
2018-02-28 08:10:48.226424 7f22273d2400  4 rocksdb:                             Options.rate_limiter: (nil)
2018-02-28 08:10:48.226425 7f22273d2400  4 rocksdb:     Options.sst_file_manager.rate_bytes_per_sec: 0
2018-02-28 08:10:48.226426 7f22273d2400  4 rocksdb:                        Options.bytes_per_sync: 0
2018-02-28 08:10:48.226428 7f22273d2400  4 rocksdb:                      Options.wal_bytes_per_sync: 0
2018-02-28 08:10:48.226429 7f22273d2400  4 rocksdb:                       Options.wal_recovery_mode: 2
2018-02-28 08:10:48.226431 7f22273d2400  4 rocksdb:                  Options.enable_thread_tracking: 0
2018-02-28 08:10:48.226432 7f22273d2400  4 rocksdb:         Options.allow_concurrent_memtable_write: 1
2018-02-28 08:10:48.226433 7f22273d2400  4 rocksdb:      Options.enable_write_thread_adaptive_yield: 1
2018-02-28 08:10:48.226435 7f22273d2400  4 rocksdb:             Options.write_thread_max_yield_usec: 100
2018-02-28 08:10:48.226437 7f22273d2400  4 rocksdb:            Options.write_thread_slow_yield_usec: 3
2018-02-28 08:10:48.226438 7f22273d2400  4 rocksdb:                                 Options.row_cache: None
2018-02-28 08:10:48.226439 7f22273d2400  4 rocksdb:                                Options.wal_filter: None
2018-02-28 08:10:48.226441 7f22273d2400  4 rocksdb:             Options.avoid_flush_during_recovery: 0
2018-02-28 08:10:48.226443 7f22273d2400  4 rocksdb:             Options.base_background_compactions: 1
2018-02-28 08:10:48.226444 7f22273d2400  4 rocksdb:             Options.max_background_compactions: 8
2018-02-28 08:10:48.226445 7f22273d2400  4 rocksdb:             Options.avoid_flush_during_shutdown: 0
2018-02-28 08:10:48.226446 7f22273d2400  4 rocksdb:             Options.delayed_write_rate : 16777216
2018-02-28 08:10:48.226448 7f22273d2400  4 rocksdb:             Options.max_total_wal_size: 0
2018-02-28 08:10:48.226449 7f22273d2400  4 rocksdb:             Options.delete_obsolete_files_period_micros: 2
1600000000
2018-02-28 08:10:48.226451 7f22273d2400  4 rocksdb:                   Options.stats_dump_period_sec: 600
2018-02-28 08:10:48.226452 7f22273d2400  4 rocksdb: Compression algorithms supported:
2018-02-28 08:10:48.226453 7f22273d2400  4 rocksdb:     Snappy supported: 0
2018-02-28 08:10:48.226455 7f22273d2400  4 rocksdb:     Zlib supported: 0
2018-02-28 08:10:48.226456 7f22273d2400  4 rocksdb:     Bzip supported: 0
2018-02-28 08:10:48.226457 7f22273d2400  4 rocksdb:     LZ4 supported: 0
2018-02-28 08:10:48.226459 7f22273d2400  4 rocksdb:     ZSTD supported: 0
2018-02-28 08:10:48.226460 7f22273d2400  4 rocksdb: Fast CRC32 supported: 0
2018-02-28 08:10:48.226949 7f22273d2400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/version_set.cc:2609] Re
covering from manifest file: MANIFEST-000368

2018-02-28 08:10:48.227183 7f22273d2400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/column_family.cc:407] -
---------------- Options for column family [default]:

2018-02-28 08:10:48.227195 7f22273d2400  4 rocksdb:                Options.comparator: leveldb.BytewiseComparat
or
2018-02-28 08:10:48.227198 7f22273d2400  4 rocksdb:           Options.merge_operator:
2018-02-28 08:10:48.227199 7f22273d2400  4 rocksdb:        Options.compaction_filter: None
2018-02-28 08:10:48.227201 7f22273d2400  4 rocksdb:        Options.compaction_filter_factory: None
2018-02-28 08:10:48.227202 7f22273d2400  4 rocksdb:         Options.memtable_factory: SkipListFactory
2018-02-28 08:10:48.227204 7f22273d2400  4 rocksdb:            Options.table_factory: BlockBasedTable
2018-02-28 08:10:48.227239 7f22273d2400  4 rocksdb:            table_factory options:   flush_block_policy_fac
tory: FlushBlockBySizePolicyFactory (0x55a9b93680f8)
  cache_index_and_filter_blocks: 1
  cache_index_and_filter_blocks_with_high_priority: 1
  pin_l0_filter_and_index_blocks_in_cache: 1
  index_type: 0
  hash_index_allow_collision: 1
  checksum: 1
  no_block_cache: 0
  block_cache: 0x55a9b960b500
  block_cache_name: LRUCache
  block_cache_options:
    capacity : 134217728
    num_shard_bits : 4
    strict_capacity_limit : 0
    high_pri_pool_ratio: 0.000
  block_cache_compressed: (nil)
  persistent_cache: (nil)
```

```
  block_size: 4096
  block_size_deviation: 10
  block_restart_interval: 16
  index_block_restart_interval: 1
  filter_policy: rocksdb.BuiltinBloomFilter
  whole_key_filtering: 1
  format_version: 2

2018-02-28 08:10:48.227245 7f22273d2400  4 rocksdb:        Options.write_buffer_size: 67108864
2018-02-28 08:10:48.227247 7f22273d2400  4 rocksdb:  Options.max_write_buffer_number: 2
2018-02-28 08:10:48.227250 7f22273d2400  4 rocksdb:          Options.compression: NoCompression
2018-02-28 08:10:48.227251 7f22273d2400  4 rocksdb:                 Options.bottommost_compression: Disabled
2018-02-28 08:10:48.227252 7f22273d2400  4 rocksdb:       Options.prefix_extractor: nullptr
2018-02-28 08:10:48.227254 7f22273d2400  4 rocksdb:   Options.memtable_insert_with_hint_prefix_extractor: null
ptr
2018-02-28 08:10:48.227255 7f22273d2400  4 rocksdb:             Options.num_levels: 7
2018-02-28 08:10:48.227256 7f22273d2400  4 rocksdb:        Options.min_write_buffer_number_to_merge: 1
2018-02-28 08:10:48.227258 7f22273d2400  4 rocksdb:     Options.max_write_buffer_number_to_maintain: 0
2018-02-28 08:10:48.227259 7f22273d2400  4 rocksdb:            Options.compression_opts.window_bits: -14
2018-02-28 08:10:48.227260 7f22273d2400  4 rocksdb:                  Options.compression_opts.level: -1
2018-02-28 08:10:48.227262 7f22273d2400  4 rocksdb:               Options.compression_opts.strategy: 0
2018-02-28 08:10:48.227263 7f22273d2400  4 rocksdb:         Options.compression_opts.max_dict_bytes: 0
2018-02-28 08:10:48.227264 7f22273d2400  4 rocksdb:      Options.level0_file_num_compaction_trigger: 4
2018-02-28 08:10:48.227265 7f22273d2400  4 rocksdb:          Options.level0_slowdown_writes_trigger: 20
2018-02-28 08:10:48.227267 7f22273d2400  4 rocksdb:              Options.level0_stop_writes_trigger: 36
2018-02-28 08:10:48.227268 7f22273d2400  4 rocksdb:                   Options.target_file_size_base: 67108864
2018-02-28 08:10:48.227269 7f22273d2400  4 rocksdb:             Options.target_file_size_multiplier: 1
2018-02-28 08:10:48.227270 7f22273d2400  4 rocksdb:                Options.max_bytes_for_level_base: 268435456
2018-02-28 08:10:48.227272 7f22273d2400  4 rocksdb: Options.level_compaction_dynamic_level_bytes: 0
2018-02-28 08:10:48.227273 7f22273d2400  4 rocksdb:          Options.max_bytes_for_level_multiplier: 10.000000
2018-02-28 08:10:48.227277 7f22273d2400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[0]: 1
2018-02-28 08:10:48.227278 7f22273d2400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[1]: 1
2018-02-28 08:10:48.227279 7f22273d2400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[2]: 1
2018-02-28 08:10:48.227281 7f22273d2400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[3]: 1
2018-02-28 08:10:48.227282 7f22273d2400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[4]: 1
2018-02-28 08:10:48.227283 7f22273d2400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[5]: 1
2018-02-28 08:10:48.227298 7f22273d2400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[6]: 1
2018-02-28 08:10:48.227300 7f22273d2400  4 rocksdb:       Options.max_sequential_skip_in_iterations: 8
2018-02-28 08:10:48.227301 7f22273d2400  4 rocksdb:                    Options.max_compaction_bytes: 167772160
0
2018-02-28 08:10:48.227302 7f22273d2400  4 rocksdb:                        Options.arena_block_size: 8388608
2018-02-28 08:10:48.227303 7f22273d2400  4 rocksdb:   Options.soft_pending_compaction_bytes_limit: 68719476736
2018-02-28 08:10:48.227305 7f22273d2400  4 rocksdb:   Options.hard_pending_compaction_bytes_limit: 27487790694
4
2018-02-28 08:10:48.227306 7f22273d2400  4 rocksdb:       Options.rate_limit_delay_max_milliseconds: 100
2018-02-28 08:10:48.227307 7f22273d2400  4 rocksdb:                Options.disable_auto_compactions: 0
2018-02-28 08:10:48.227309 7f22273d2400  4 rocksdb:                         Options.compaction_style: kCompact
ionStyleLevel
2018-02-28 08:10:48.227311 7f22273d2400  4 rocksdb:                           Options.compaction_pri: kByCompe
nsatedSize
2018-02-28 08:10:48.227312 7f22273d2400  4 rocksdb:  Options.compaction_options_universal.size_ratio: 1
2018-02-28 08:10:48.227313 7f22273d2400  4 rocksdb: Options.compaction_options_universal.min_merge_width: 2
2018-02-28 08:10:48.227314 7f22273d2400  4 rocksdb: Options.compaction_options_universal.max_merge_width: 4294
967295
2018-02-28 08:10:48.227315 7f22273d2400  4 rocksdb: Options.compaction_options_universal.max_size_amplificatio
n_percent: 200
2018-02-28 08:10:48.227317 7f22273d2400  4 rocksdb: Options.compaction_options_universal.compression_size_perc
ent: -1
2018-02-28 08:10:48.227318 7f22273d2400  4 rocksdb: Options.compaction_options_fifo.max_table_files_size: 1073
741824
2018-02-28 08:10:48.227319 7f22273d2400  4 rocksdb:                   Options.table_properties_collectors:
2018-02-28 08:10:48.227321 7f22273d2400  4 rocksdb:                   Options.inplace_update_support: 0
2018-02-28 08:10:48.227323 7f22273d2400  4 rocksdb:                 Options.inplace_update_num_locks: 10000
2018-02-28 08:10:48.227324 7f22273d2400  4 rocksdb:               Options.memtable_prefix_bloom_size_ratio: 0.
000000
2018-02-28 08:10:48.227326 7f22273d2400  4 rocksdb:   Options.memtable_huge_page_size: 0
2018-02-28 08:10:48.227327 7f22273d2400  4 rocksdb:                            Options.bloom_locality: 0
2018-02-28 08:10:48.227329 7f22273d2400  4 rocksdb:                     Options.max_successive_merges: 0
2018-02-28 08:10:48.227330 7f22273d2400  4 rocksdb:                 Options.optimize_filters_for_hits: 0
2018-02-28 08:10:48.227331 7f22273d2400  4 rocksdb:                 Options.paranoid_file_checks: 0
2018-02-28 08:10:48.227333 7f22273d2400  4 rocksdb:                 Options.force_consistency_checks: 0
2018-02-28 08:10:48.227334 7f22273d2400  4 rocksdb:                 Options.report_bg_io_stats: 0
2018-02-28 08:10:48.233940 7f22273d2400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/version_set.cc:2859] Re
covered from manifest file:/var/lib/ceph/osd/cephprod-78/current/omap/MANIFEST-000368 succeeded,manifest_file_
number is 368, next_file_number is 370, last_sequence is 19073448, log_number is 0,prev_log_number is 0,max_co
```

```
lumn_family is 0

2018-02-28 08:10:48.233963 7f22273d2400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/version_set.cc:2867] Co
lumn family [default] (ID 0), log number is 367

2018-02-28 08:10:48.234203 7f22273d2400  4 rocksdb: EVENT_LOG_v1 {"time_micros": 1519801848234192, "job": 1, "
event": "recovery_started", "log_files": [369]}
2018-02-28 08:10:48.234215 7f22273d2400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/db_impl_open.cc:482] Re
covering log #369 mode 2
2018-02-28 08:10:48.234377 7f22273d2400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/version_set.cc:2395] Cr
eating manifest 371

2018-02-28 08:10:48.265509 7f22273d2400  4 rocksdb: EVENT_LOG_v1 {"time_micros": 1519801848265503, "job": 1, "
event": "recovery_finished"}
2018-02-28 08:10:48.308426 7f22273d2400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/db_impl_open.cc:1063] D
B pointer 0x55a9b968c000
2018-02-28 08:10:49.423404 7f22273d2400  0 filestore(/var/lib/ceph/osd/cephprod-78) mount(1757): enabling WRIT
EAHEAD journal mode: checkpoint is not enabled
2018-02-28 08:10:49.426222 7f22273d2400  1 journal _open /var/lib/ceph/osd/cephprod-78/journal fd 21: 10737418
240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:10:49.426729 7f22273d2400 -1 journal do_read_entry(626839552): bad header magic
2018-02-28 08:10:49.426750 7f22273d2400 -1 journal do_read_entry(626839552): bad header magic
2018-02-28 08:10:49.426966 7f22273d2400  1 journal _open /var/lib/ceph/osd/cephprod-78/journal fd 21: 10737418
240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:10:49.428234 7f22273d2400  1 filestore(/var/lib/ceph/osd/cephprod-78) upgrade(1364)
Cluster fsid=f9dfd27f-c704-4d53-9aa0-4a23d655c7c4
Supported features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object loca
tor,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharde
d objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in mis
sing set}
On-disk features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object locato
r,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharded
objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in missi
ng set}
1 pgs to scan
Scanning 11.5f_head, 0/1 completed
Error getting attr on : 11.5f_head,#-13:fa000000:::scrub_11.5f:head#, (61) No data available
2018-02-28 08:10:49.430341 7f22273d2400  1 journal close /var/lib/ceph/osd/cephprod-78/journal
2018-02-28 08:10:49.432967 7f22273d2400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/db_impl.cc:217] Shutdow
n: canceling all background work
2018-02-28 08:10:49.434580 7f22273d2400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/db_impl.cc:343] Shutdow
n complete

================================================================================================================
==============
ceph-objectstore-tool --cluster cephprod --data-path /var/lib/ceph/osd/cephprod-78 --pgid 11.5f --journal /var
/lib/ceph/osd/cephprod-78/journal --debug .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8EDF__
b list-omap --op list

2018-02-28 08:10:49.636007 7f233425a400  0 filestore(/var/lib/ceph/osd/cephprod-78) backend xfs (magic 0x58465
342)
2018-02-28 08:10:49.636563 7f233425a400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-78) detect_featu
res: FIEMAP ioctl is disabled via 'filestore fiemap' config option
2018-02-28 08:10:49.636584 7f233425a400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-78) detect_featu
res: SEEK_DATA/SEEK_HOLE is disabled via 'filestore seek data hole' config option
2018-02-28 08:10:49.636585 7f233425a400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-78) detect_featu
res: splice() is disabled via 'filestore splice' config option
2018-02-28 08:10:49.658723 7f233425a400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-78) detect_featu
res: syncfs(2) syscall fully supported (by glibc and kernel)
2018-02-28 08:10:49.658828 7f233425a400  0 xfsfilestorebackend(/var/lib/ceph/osd/cephprod-78) detect_feature:
extsize is disabled by conf
2018-02-28 08:10:49.659646 7f233425a400  0 filestore(/var/lib/ceph/osd/cephprod-78) start omap initiation
2018-02-28 08:10:49.659736 7f233425a400  0  set rocksdb option compaction_readahead_size = 2097152
2018-02-28 08:10:49.659754 7f233425a400  0  set rocksdb option compression = kNoCompression
2018-02-28 08:10:49.659764 7f233425a400  0  set rocksdb option max_background_compactions = 8
2018-02-28 08:10:49.659808 7f233425a400  0  set rocksdb option compaction_readahead_size = 2097152
2018-02-28 08:10:49.659818 7f233425a400  0  set rocksdb option compression = kNoCompression
2018-02-28 08:10:49.659824 7f233425a400  0  set rocksdb option max_background_compactions = 8
2018-02-28 08:10:49.660719 7f233425a400  4 rocksdb: RocksDB version: 5.4.0

2018-02-28 08:10:49.660737 7f233425a400  4 rocksdb: Git sha rocksdb_build_git_sha:@0@
2018-02-28 08:10:49.660740 7f233425a400  4 rocksdb: Compile date Nov 30 2017
2018-02-28 08:10:49.660743 7f233425a400  4 rocksdb: DB SUMMARY

2018-02-28 08:10:49.660818 7f233425a400  4 rocksdb: CURRENT file:  CURRENT
```

```
2018-02-28 08:10:49.660822 7f233425a400  4 rocksdb: IDENTITY file:  IDENTITY

2018-02-28 08:10:49.660829 7f233425a400  4 rocksdb: MANIFEST file:  MANIFEST-000371 size: 2020 Bytes

2018-02-28 08:10:49.660835 7f233425a400  4 rocksdb: SST files in /var/lib/ceph/osd/cephprod-78/current/omap di
r, Total Num: 10, files: 000158.sst 000304.sst 000316.sst 000317.sst 000326.sst 000327.sst 000328.sst 000330.s
st 000332.sst

2018-02-28 08:10:49.660840 7f233425a400  4 rocksdb: Write Ahead Log file in /var/lib/ceph/osd/cephprod-78/curr
ent/omap: 000372.log size: 0 ;

2018-02-28 08:10:49.660843 7f233425a400  4 rocksdb:                         Options.error_if_exists: 0
2018-02-28 08:10:49.660846 7f233425a400  4 rocksdb:                       Options.create_if_missing: 1
2018-02-28 08:10:49.660848 7f233425a400  4 rocksdb:                         Options.paranoid_checks: 1
2018-02-28 08:10:49.660849 7f233425a400  4 rocksdb:                                     Options.env: 0x5572fde
b9c80
2018-02-28 08:10:49.660851 7f233425a400  4 rocksdb:                                Options.info_log: 0x5572fff
2e500
2018-02-28 08:10:49.660852 7f233425a400  4 rocksdb:                          Options.max_open_files: -1
2018-02-28 08:10:49.660853 7f233425a400  4 rocksdb:                Options.max_file_opening_threads: 16
2018-02-28 08:10:49.660855 7f233425a400  4 rocksdb:                             Options.use_fsync: 0
2018-02-28 08:10:49.660857 7f233425a400  4 rocksdb:                     Options.max_log_file_size: 0
2018-02-28 08:10:49.660859 7f233425a400  4 rocksdb:                  Options.max_manifest_file_size: 184467440
73709551615
2018-02-28 08:10:49.660860 7f233425a400  4 rocksdb:                   Options.log_file_time_to_roll: 0
2018-02-28 08:10:49.660862 7f233425a400  4 rocksdb:                       Options.keep_log_file_num: 1000
2018-02-28 08:10:49.660863 7f233425a400  4 rocksdb:                    Options.recycle_log_file_num: 0
2018-02-28 08:10:49.660865 7f233425a400  4 rocksdb:                         Options.allow_fallocate: 1
2018-02-28 08:10:49.660867 7f233425a400  4 rocksdb:                        Options.allow_mmap_reads: 0
2018-02-28 08:10:49.660868 7f233425a400  4 rocksdb:                       Options.allow_mmap_writes: 0
2018-02-28 08:10:49.660870 7f233425a400  4 rocksdb:                        Options.use_direct_reads: 0
2018-02-28 08:10:49.660871 7f233425a400  4 rocksdb:                        Options.use_direct_io_for_flush_and
_compaction: 0
2018-02-28 08:10:49.660872 7f233425a400  4 rocksdb:          Options.create_missing_column_families: 0
2018-02-28 08:10:49.660874 7f233425a400  4 rocksdb:                                Options.db_log_dir:
2018-02-28 08:10:49.660876 7f233425a400  4 rocksdb:                                   Options.wal_dir: /var/lib/
ceph/osd/cephprod-78/current/omap
2018-02-28 08:10:49.660877 7f233425a400  4 rocksdb:                Options.table_cache_numshardbits: 6
2018-02-28 08:10:49.660879 7f233425a400  4 rocksdb:                      Options.max_subcompactions: 1
2018-02-28 08:10:49.660893 7f233425a400  4 rocksdb:                  Options.max_background_flushes: 1
2018-02-28 08:10:49.660895 7f233425a400  4 rocksdb:                         Options.WAL_ttl_seconds: 0
2018-02-28 08:10:49.660896 7f233425a400  4 rocksdb:                       Options.WAL_size_limit_MB: 0
2018-02-28 08:10:49.660898 7f233425a400  4 rocksdb:            Options.manifest_preallocation_size: 4194304
2018-02-28 08:10:49.660899 7f233425a400  4 rocksdb:                    Options.is_fd_close_on_exec: 1
2018-02-28 08:10:49.660901 7f233425a400  4 rocksdb:                  Options.advise_random_on_open: 1
2018-02-28 08:10:49.660902 7f233425a400  4 rocksdb:                   Options.db_write_buffer_size: 0
2018-02-28 08:10:49.660903 7f233425a400  4 rocksdb:         Options.access_hint_on_compaction_start: 1
2018-02-28 08:10:49.660905 7f233425a400  4 rocksdb:  Options.new_table_reader_for_compaction_inputs: 1
2018-02-28 08:10:49.660906 7f233425a400  4 rocksdb:              Options.compaction_readahead_size: 2097152
2018-02-28 08:10:49.660907 7f233425a400  4 rocksdb:           Options.random_access_max_buffer_size: 1048576
2018-02-28 08:10:49.660909 7f233425a400  4 rocksdb:           Options.writable_file_max_buffer_size: 1048576
2018-02-28 08:10:49.660911 7f233425a400  4 rocksdb:                        Options.use_adaptive_mutex: 0
2018-02-28 08:10:49.660912 7f233425a400  4 rocksdb:                              Options.rate_limiter: (nil)
2018-02-28 08:10:49.660914 7f233425a400  4 rocksdb:    Options.sst_file_manager.rate_bytes_per_sec: 0
2018-02-28 08:10:49.660916 7f233425a400  4 rocksdb:                           Options.bytes_per_sync: 0
2018-02-28 08:10:49.660917 7f233425a400  4 rocksdb:                       Options.wal_bytes_per_sync: 0
2018-02-28 08:10:49.660919 7f233425a400  4 rocksdb:                        Options.wal_recovery_mode: 2
2018-02-28 08:10:49.660920 7f233425a400  4 rocksdb:                  Options.enable_thread_tracking: 0
2018-02-28 08:10:49.660922 7f233425a400  4 rocksdb:         Options.allow_concurrent_memtable_write: 1
2018-02-28 08:10:49.660923 7f233425a400  4 rocksdb:      Options.enable_write_thread_adaptive_yield: 1
2018-02-28 08:10:49.660924 7f233425a400  4 rocksdb:             Options.write_thread_max_yield_usec: 100
2018-02-28 08:10:49.660925 7f233425a400  4 rocksdb:            Options.write_thread_slow_yield_usec: 3
2018-02-28 08:10:49.660927 7f233425a400  4 rocksdb:                                 Options.row_cache: None
2018-02-28 08:10:49.660929 7f233425a400  4 rocksdb:                                Options.wal_filter: None
2018-02-28 08:10:49.660930 7f233425a400  4 rocksdb:             Options.avoid_flush_during_recovery: 0
2018-02-28 08:10:49.660932 7f233425a400  4 rocksdb:             Options.base_background_compactions: 1
2018-02-28 08:10:49.660933 7f233425a400  4 rocksdb:             Options.max_background_compactions: 8
2018-02-28 08:10:49.660935 7f233425a400  4 rocksdb:             Options.avoid_flush_during_shutdown: 0
2018-02-28 08:10:49.660936 7f233425a400  4 rocksdb:             Options.delayed_write_rate : 16777216
2018-02-28 08:10:49.660938 7f233425a400  4 rocksdb:             Options.max_total_wal_size: 0
2018-02-28 08:10:49.660939 7f233425a400  4 rocksdb:             Options.delete_obsolete_files_period_micros: 2
1600000000
2018-02-28 08:10:49.660941 7f233425a400  4 rocksdb:                   Options.stats_dump_period_sec: 600
2018-02-28 08:10:49.660943 7f233425a400  4 rocksdb: Compression algorithms supported:
```

```
2018-02-28 08:10:49.660945 7f233425a400  4 rocksdb:     Snappy supported: 0
2018-02-28 08:10:49.660947 7f233425a400  4 rocksdb:     Zlib supported: 0
2018-02-28 08:10:49.660948 7f233425a400  4 rocksdb:     Bzip supported: 0
2018-02-28 08:10:49.660949 7f233425a400  4 rocksdb:     LZ4 supported: 0
2018-02-28 08:10:49.660951 7f233425a400  4 rocksdb:     ZSTD supported: 0
2018-02-28 08:10:49.660953 7f233425a400  4 rocksdb: Fast CRC32 supported: 0
2018-02-28 08:10:49.661492 7f233425a400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/version_set.cc:2609] Re
covering from manifest file: MANIFEST-000371

2018-02-28 08:10:49.661700 7f233425a400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/column_family.cc:407] -
---------------- Options for column family [default]:

2018-02-28 08:10:49.661711 7f233425a400  4 rocksdb:               Options.comparator: leveldb.BytewiseComparat
or
2018-02-28 08:10:49.661713 7f233425a400  4 rocksdb:           Options.merge_operator:
2018-02-28 08:10:49.661715 7f233425a400  4 rocksdb:        Options.compaction_filter: None
2018-02-28 08:10:49.661717 7f233425a400  4 rocksdb:        Options.compaction_filter_factory: None
2018-02-28 08:10:49.661719 7f233425a400  4 rocksdb:         Options.memtable_factory: SkipListFactory
2018-02-28 08:10:49.661721 7f233425a400  4 rocksdb:            Options.table_factory: BlockBasedTable
2018-02-28 08:10:49.661757 7f233425a400  4 rocksdb:            table_factory options:   flush_block_policy_fac
tory: FlushBlockBySizePolicyFactory (0x5572ffc7c0f8)
  cache_index_and_filter_blocks: 1
  cache_index_and_filter_blocks_with_high_priority: 1
  pin_l0_filter_and_index_blocks_in_cache: 1
  index_type: 0
  hash_index_allow_collision: 1
  checksum: 1
  no_block_cache: 0
  block_cache: 0x5572fff1f500
  block_cache_name: LRUCache
  block_cache_options:
    capacity : 134217728
    num_shard_bits : 4
    strict_capacity_limit : 0
    high_pri_pool_ratio: 0.000
  block_cache_compressed: (nil)
  persistent_cache: (nil)
  block_size: 4096
  block_size_deviation: 10
  block_restart_interval: 16
  index_block_restart_interval: 1
  filter_policy: rocksdb.BuiltinBloomFilter
  whole_key_filtering: 1
  format_version: 2

2018-02-28 08:10:49.661765 7f233425a400  4 rocksdb:        Options.write_buffer_size: 67108864
2018-02-28 08:10:49.661767 7f233425a400  4 rocksdb:  Options.max_write_buffer_number: 2
2018-02-28 08:10:49.661769 7f233425a400  4 rocksdb:          Options.compression: NoCompression
2018-02-28 08:10:49.661771 7f233425a400  4 rocksdb:                 Options.bottommost_compression: Disabled
2018-02-28 08:10:49.661773 7f233425a400  4 rocksdb:       Options.prefix_extractor: nullptr
2018-02-28 08:10:49.661774 7f233425a400  4 rocksdb:   Options.memtable_insert_with_hint_prefix_extractor: null
ptr
2018-02-28 08:10:49.661776 7f233425a400  4 rocksdb:             Options.num_levels: 7
2018-02-28 08:10:49.661777 7f233425a400  4 rocksdb:        Options.min_write_buffer_number_to_merge: 1
2018-02-28 08:10:49.661779 7f233425a400  4 rocksdb:     Options.max_write_buffer_number_to_maintain: 0
2018-02-28 08:10:49.661780 7f233425a400  4 rocksdb:            Options.compression_opts.window_bits: -14
2018-02-28 08:10:49.661782 7f233425a400  4 rocksdb:                  Options.compression_opts.level: -1
2018-02-28 08:10:49.661783 7f233425a400  4 rocksdb:               Options.compression_opts.strategy: 0
2018-02-28 08:10:49.661784 7f233425a400  4 rocksdb:         Options.compression_opts.max_dict_bytes: 0
2018-02-28 08:10:49.661785 7f233425a400  4 rocksdb:      Options.level0_file_num_compaction_trigger: 4
2018-02-28 08:10:49.661787 7f233425a400  4 rocksdb:          Options.level0_slowdown_writes_trigger: 20
2018-02-28 08:10:49.661788 7f233425a400  4 rocksdb:              Options.level0_stop_writes_trigger: 36
2018-02-28 08:10:49.661789 7f233425a400  4 rocksdb:                   Options.target_file_size_base: 67108864
2018-02-28 08:10:49.661791 7f233425a400  4 rocksdb:             Options.target_file_size_multiplier: 1
2018-02-28 08:10:49.661792 7f233425a400  4 rocksdb:                Options.max_bytes_for_level_base: 268435456
2018-02-28 08:10:49.661793 7f233425a400  4 rocksdb: Options.level_compaction_dynamic_level_bytes: 0
2018-02-28 08:10:49.661795 7f233425a400  4 rocksdb:          Options.max_bytes_for_level_multiplier: 10.000000
2018-02-28 08:10:49.661799 7f233425a400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[0]: 1
2018-02-28 08:10:49.661801 7f233425a400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[1]: 1
2018-02-28 08:10:49.661802 7f233425a400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[2]: 1
2018-02-28 08:10:49.661803 7f233425a400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[3]: 1
2018-02-28 08:10:49.661804 7f233425a400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[4]: 1
2018-02-28 08:10:49.661806 7f233425a400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[5]: 1
2018-02-28 08:10:49.661827 7f233425a400  4 rocksdb: Options.max_bytes_for_level_multiplier_addtl[6]: 1
2018-02-28 08:10:49.661829 7f233425a400  4 rocksdb:       Options.max_sequential_skip_in_iterations: 8
```

```
2018-02-28 08:10:49.661831 7f233425a400  4 rocksdb:                    Options.max_compaction_bytes: 167772160
0
2018-02-28 08:10:49.661832 7f233425a400  4 rocksdb:                        Options.arena_block_size: 8388608
2018-02-28 08:10:49.661833 7f233425a400  4 rocksdb:   Options.soft_pending_compaction_bytes_limit: 68719476736
2018-02-28 08:10:49.661834 7f233425a400  4 rocksdb:   Options.hard_pending_compaction_bytes_limit: 27487790694
4
2018-02-28 08:10:49.661836 7f233425a400  4 rocksdb:       Options.rate_limit_delay_max_milliseconds: 100
2018-02-28 08:10:49.661837 7f233425a400  4 rocksdb:                Options.disable_auto_compactions: 0
2018-02-28 08:10:49.661840 7f233425a400  4 rocksdb:                         Options.compaction_style: kCompact
ionStyleLevel
2018-02-28 08:10:49.661841 7f233425a400  4 rocksdb:                           Options.compaction_pri: kByCompe
nsatedSize
2018-02-28 08:10:49.661843 7f233425a400  4 rocksdb:  Options.compaction_options_universal.size_ratio: 1
2018-02-28 08:10:49.661844 7f233425a400  4 rocksdb: Options.compaction_options_universal.min_merge_width: 2
2018-02-28 08:10:49.661845 7f233425a400  4 rocksdb: Options.compaction_options_universal.max_merge_width: 4294
967295
2018-02-28 08:10:49.661847 7f233425a400  4 rocksdb: Options.compaction_options_universal.max_size_amplificatio
n_percent: 200
2018-02-28 08:10:49.661848 7f233425a400  4 rocksdb: Options.compaction_options_universal.compression_size_perc
ent: -1
2018-02-28 08:10:49.661850 7f233425a400  4 rocksdb: Options.compaction_options_fifo.max_table_files_size: 1073
741824
2018-02-28 08:10:49.661851 7f233425a400  4 rocksdb:                  Options.table_properties_collectors:
2018-02-28 08:10:49.661852 7f233425a400  4 rocksdb:                  Options.inplace_update_support: 0
2018-02-28 08:10:49.661854 7f233425a400  4 rocksdb:                Options.inplace_update_num_locks: 10000
2018-02-28 08:10:49.661855 7f233425a400  4 rocksdb:              Options.memtable_prefix_bloom_size_ratio: 0.
000000
2018-02-28 08:10:49.661857 7f233425a400  4 rocksdb:   Options.memtable_huge_page_size: 0
2018-02-28 08:10:49.661859 7f233425a400  4 rocksdb:                           Options.bloom_locality: 0
2018-02-28 08:10:49.661860 7f233425a400  4 rocksdb:                    Options.max_successive_merges: 0
2018-02-28 08:10:49.661861 7f233425a400  4 rocksdb:                Options.optimize_filters_for_hits: 0
2018-02-28 08:10:49.661863 7f233425a400  4 rocksdb:                Options.paranoid_file_checks: 0
2018-02-28 08:10:49.661864 7f233425a400  4 rocksdb:                Options.force_consistency_checks: 0
2018-02-28 08:10:49.661865 7f233425a400  4 rocksdb:                Options.report_bg_io_stats: 0
2018-02-28 08:10:49.667931 7f233425a400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/version_set.cc:2859] Re
covered from manifest file:/var/lib/ceph/osd/cephprod-78/current/omap/MANIFEST-000371 succeeded,manifest_file_
number is 371, next_file_number is 373, last_sequence is 19073448, log_number is 0,prev_log_number is 0,max_co
lumn_family is 0

2018-02-28 08:10:49.667959 7f233425a400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/version_set.cc:2867] Co
lumn family [default] (ID 0), log number is 370

2018-02-28 08:10:49.668162 7f233425a400  4 rocksdb: EVENT_LOG_v1 {"time_micros": 1519801849668152, "job": 1, "
event": "recovery_started", "log_files": [372]}
2018-02-28 08:10:49.668175 7f233425a400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/db_impl_open.cc:482] Re
covering log #372 mode 2
2018-02-28 08:10:49.668323 7f233425a400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/version_set.cc:2395] Cr
eating manifest 374

2018-02-28 08:10:49.712054 7f233425a400  4 rocksdb: EVENT_LOG_v1 {"time_micros": 1519801849712047, "job": 1, "
event": "recovery_finished"}
2018-02-28 08:10:49.773877 7f233425a400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/db_impl_open.cc:1063] D
B pointer 0x5572fffa0000
2018-02-28 08:10:50.907709 7f233425a400  0 filestore(/var/lib/ceph/osd/cephprod-78) mount(1757): enabling WRIT
EAHEAD journal mode: checkpoint is not enabled
2018-02-28 08:10:50.910646 7f233425a400  1 journal _open /var/lib/ceph/osd/cephprod-78/journal fd 21: 10737418
240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:10:50.911138 7f233425a400 -1 journal do_read_entry(626839552): bad header magic
2018-02-28 08:10:50.911157 7f233425a400 -1 journal do_read_entry(626839552): bad header magic
2018-02-28 08:10:50.911347 7f233425a400  1 journal _open /var/lib/ceph/osd/cephprod-78/journal fd 21: 10737418
240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:10:50.912616 7f233425a400  1 filestore(/var/lib/ceph/osd/cephprod-78) upgrade(1364)
Cluster fsid=f9dfd27f-c704-4d53-9aa0-4a23d655c7c4
Supported features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object loca
tor,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharde
d objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in mis
sing set}
On-disk features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object locato
r,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharded
objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in missi
ng set}
1 pgs to scan
Scanning 11.5f_head, 0/1 completed
Error getting attr on : 11.5f_head,#-13:fa000000:::scrub_11.5f:head#, (61) No data available
2018-02-28 08:10:50.914622 7f233425a400  1 journal close /var/lib/ceph/osd/cephprod-78/journal
```

```
2018-02-28 08:10:50.917589 7f233425a400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/db_impl.cc:217] Shutdow
n: canceling all background work
2018-02-28 08:10:50.919123 7f233425a400  4 rocksdb: [/build/ceph-12.2.2/src/rocksdb/db/db_impl.cc:343] Shutdow
n complete

================================================================================================================
==============
```

iccluster003

```
================================================================================================================
==============
ceph-objectstore-tool --cluster cephprod --data-path /var/lib/ceph/osd/cephprod-154 --pgid 11.5f --journal /va
r/lib/ceph/osd/cephprod-154/journal --debug .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8EDF
__b get-omaphdr --op list

2018-02-28 08:09:20.546358 7f4e27184400  0 filestore(/var/lib/ceph/osd/cephprod-154) backend xfs (magic 0x5846
5342)
2018-02-28 08:09:20.546869 7f4e27184400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-154) detect_feat
ures: FIEMAP ioctl is disabled via 'filestore fiemap' config option
2018-02-28 08:09:20.546886 7f4e27184400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-154) detect_feat
ures: SEEK_DATA/SEEK_HOLE is disabled via 'filestore seek data hole' config option
2018-02-28 08:09:20.546888 7f4e27184400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-154) detect_feat
ures: splice() is disabled via 'filestore splice' config option
2018-02-28 08:09:20.558458 7f4e27184400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-154) detect_feat
ures: syncfs(2) syscall fully supported (by glibc and kernel)
2018-02-28 08:09:20.558556 7f4e27184400  0 xfsfilestorebackend(/var/lib/ceph/osd/cephprod-154) detect_feature:
 extsize is disabled by conf
2018-02-28 08:09:20.559415 7f4e27184400  0 filestore(/var/lib/ceph/osd/cephprod-154) start omap initiation
2018-02-28 08:09:22.246165 7f4e27184400  0 filestore(/var/lib/ceph/osd/cephprod-154) mount(1757): enabling WRI
TEAHEAD journal mode: checkpoint is not enabled
2018-02-28 08:09:22.248878 7f4e27184400  1 journal _open /var/lib/ceph/osd/cephprod-154/journal fd 11: 1073741
8240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:09:22.249433 7f4e27184400 -1 journal do_read_entry(8727834624): bad header magic
2018-02-28 08:09:22.249458 7f4e27184400 -1 journal do_read_entry(8727834624): bad header magic
2018-02-28 08:09:22.249644 7f4e27184400  1 journal _open /var/lib/ceph/osd/cephprod-154/journal fd 11: 1073741
8240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:09:22.250832 7f4e27184400  1 filestore(/var/lib/ceph/osd/cephprod-154) upgrade(1364)
Cluster fsid=f9dfd27f-c704-4d53-9aa0-4a23d655c7c4
Supported features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object loca
tor,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharde
d objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in mis
sing set}
On-disk features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object locato
r,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharded
objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in missi
ng set}
1 pgs to scan
Scanning 11.5f_head, 0/1 completed
2018-02-28 08:09:22.252699 7f4e27184400  1 journal close /var/lib/ceph/osd/cephprod-154/journal

================================================================================================================
==============
ceph-objectstore-tool --cluster cephprod --data-path /var/lib/ceph/osd/cephprod-154 --pgid 11.5f --journal /va
r/lib/ceph/osd/cephprod-154/journal --debug .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8EDF
__b list-omap --op list

2018-02-28 08:09:22.403113 7ff8390cf400  0 filestore(/var/lib/ceph/osd/cephprod-154) backend xfs (magic 0x5846
5342)
2018-02-28 08:09:22.403566 7ff8390cf400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-154) detect_feat
ures: FIEMAP ioctl is disabled via 'filestore fiemap' config option
2018-02-28 08:09:22.403581 7ff8390cf400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-154) detect_feat
ures: SEEK_DATA/SEEK_HOLE is disabled via 'filestore seek data hole' config option
2018-02-28 08:09:22.403582 7ff8390cf400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-154) detect_feat
ures: splice() is disabled via 'filestore splice' config option
2018-02-28 08:09:22.423052 7ff8390cf400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-154) detect_feat
ures: syncfs(2) syscall fully supported (by glibc and kernel)
2018-02-28 08:09:22.423151 7ff8390cf400  0 xfsfilestorebackend(/var/lib/ceph/osd/cephprod-154) detect_feature:
 extsize is disabled by conf
2018-02-28 08:09:22.423696 7ff8390cf400  0 filestore(/var/lib/ceph/osd/cephprod-154) start omap initiation
2018-02-28 08:09:24.205131 7ff8390cf400  0 filestore(/var/lib/ceph/osd/cephprod-154) mount(1757): enabling WRI
TEAHEAD journal mode: checkpoint is not enabled
```

```
2018-02-28 08:09:24.207720 7ff8390cf400  1 journal _open /var/lib/ceph/osd/cephprod-154/journal fd 11: 1073741
8240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:09:24.208188 7ff8390cf400 -1 journal do_read_entry(8727834624): bad header magic
2018-02-28 08:09:24.208197 7ff8390cf400 -1 journal do_read_entry(8727834624): bad header magic
2018-02-28 08:09:24.208373 7ff8390cf400  1 journal _open /var/lib/ceph/osd/cephprod-154/journal fd 11: 1073741
8240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:09:24.208928 7ff8390cf400  1 filestore(/var/lib/ceph/osd/cephprod-154) upgrade(1364)
Cluster fsid=f9dfd27f-c704-4d53-9aa0-4a23d655c7c4
Supported features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object loca
tor,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharde
d objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in mis
sing set}
On-disk features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object locato
r,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharded
objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in missi
ng set}
1 pgs to scan
Scanning 11.5f_head, 0/1 completed
2018-02-28 08:09:24.209959 7ff8390cf400  1 journal close /var/lib/ceph/osd/cephprod-154/journal

================================================================================================================
==============
```

iccluster021

```
================================================================================================================
==============
ceph-objectstore-tool --cluster cephprod --data-path /var/lib/ceph/osd/cephprod-170 --pgid 11.5f --journal /va
r/lib/ceph/osd/cephprod-170/journal --debug .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8EDF
__b get-omaphdr --op list

2018-02-28 08:09:53.154905 7f8ef1852400  0 filestore(/var/lib/ceph/osd/cephprod-170) backend xfs (magic 0x5846
5342)
2018-02-28 08:09:53.155402 7f8ef1852400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-170) detect_feat
ures: FIEMAP ioctl is disabled via 'filestore fiemap' config option
2018-02-28 08:09:53.155420 7f8ef1852400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-170) detect_feat
ures: SEEK_DATA/SEEK_HOLE is disabled via 'filestore seek data hole' config option
2018-02-28 08:09:53.155422 7f8ef1852400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-170) detect_feat
ures: splice() is disabled via 'filestore splice' config option
2018-02-28 08:09:53.157589 7f8ef1852400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-170) detect_feat
ures: syncfs(2) syscall fully supported (by glibc and kernel)
2018-02-28 08:09:53.157682 7f8ef1852400  0 xfsfilestorebackend(/var/lib/ceph/osd/cephprod-170) detect_feature:
 extsize is disabled by conf
2018-02-28 08:09:53.158407 7f8ef1852400  0 filestore(/var/lib/ceph/osd/cephprod-170) start omap initiation
2018-02-28 08:09:54.759126 7f8ef1852400  0 filestore(/var/lib/ceph/osd/cephprod-170) mount(1757): enabling WRI
TEAHEAD journal mode: checkpoint is not enabled
2018-02-28 08:09:54.761726 7f8ef1852400  1 journal _open /var/lib/ceph/osd/cephprod-170/journal fd 11: 1073741
8240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:09:54.766991 7f8ef1852400 -1 journal do_read_entry(7316729856): bad header magic
2018-02-28 08:09:54.767185 7f8ef1852400  1 journal _open /var/lib/ceph/osd/cephprod-170/journal fd 11: 1073741
8240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:09:54.767756 7f8ef1852400  1 filestore(/var/lib/ceph/osd/cephprod-170) upgrade(1364)
Cluster fsid=f9dfd27f-c704-4d53-9aa0-4a23d655c7c4
Supported features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object loca
tor,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharde
d objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in mis
sing set}
On-disk features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object locato
r,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharded
objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in missi
ng set}
1 pgs to scan
Scanning 11.5f_head, 0/1 completed
2018-02-28 08:09:54.815818 7f8ef1852400  1 journal close /var/lib/ceph/osd/cephprod-170/journal

================================================================================================================
==============
ceph-objectstore-tool --cluster cephprod --data-path /var/lib/ceph/osd/cephprod-170 --pgid 11.5f --journal /va
r/lib/ceph/osd/cephprod-170/journal --debug .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8EDF
__b list-omap --op list

2018-02-28 08:09:54.977291 7fd5b7f2b400  0 filestore(/var/lib/ceph/osd/cephprod-170) backend xfs (magic 0x5846
```

```
5342)
2018-02-28 08:09:54.977831 7fd5b7f2b400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-170) detect_feat
ures: FIEMAP ioctl is disabled via 'filestore fiemap' config option
2018-02-28 08:09:54.977850 7fd5b7f2b400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-170) detect_feat
ures: SEEK_DATA/SEEK_HOLE is disabled via 'filestore seek data hole' config option
2018-02-28 08:09:54.977851 7fd5b7f2b400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-170) detect_feat
ures: splice() is disabled via 'filestore splice' config option
2018-02-28 08:09:54.989337 7fd5b7f2b400  0 genericfilestorebackend(/var/lib/ceph/osd/cephprod-170) detect_feat
ures: syncfs(2) syscall fully supported (by glibc and kernel)
2018-02-28 08:09:54.989416 7fd5b7f2b400  0 xfsfilestorebackend(/var/lib/ceph/osd/cephprod-170) detect_feature:
 extsize is disabled by conf
2018-02-28 08:09:54.990103 7fd5b7f2b400  0 filestore(/var/lib/ceph/osd/cephprod-170) start omap initiation
2018-02-28 08:09:56.617268 7fd5b7f2b400  0 filestore(/var/lib/ceph/osd/cephprod-170) mount(1757): enabling WRI
TEAHEAD journal mode: checkpoint is not enabled
2018-02-28 08:09:56.619856 7fd5b7f2b400  1 journal _open /var/lib/ceph/osd/cephprod-170/journal fd 11: 1073741
8240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:09:56.620331 7fd5b7f2b400 -1 journal do_read_entry(7316729856): bad header magic
2018-02-28 08:09:56.620340 7fd5b7f2b400 -1 journal do_read_entry(7316729856): bad header magic
2018-02-28 08:09:56.620512 7fd5b7f2b400  1 journal _open /var/lib/ceph/osd/cephprod-170/journal fd 11: 1073741
8240 bytes, block size 4096 bytes, directio = 1, aio = 1
2018-02-28 08:09:56.621089 7fd5b7f2b400  1 filestore(/var/lib/ceph/osd/cephprod-170) upgrade(1364)
Cluster fsid=f9dfd27f-c704-4d53-9aa0-4a23d655c7c4
Supported features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object loca
tor,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharde
d objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in mis
sing set}
On-disk features: compat={},rocompat={},incompat={1=initial feature set(~v.18),2=pginfo object,3=object locato
r,4=last_epoch_clean,5=categories,6=hobjectpool,7=biginfo,8=leveldbinfo,9=leveldblog,10=snapmapper,11=sharded
 objects,12=transaction hints,13=pg meta object,14=explicit missing set,15=fastinfo pg attr,16=deletes in missi
ng set}
1 pgs to scan
Scanning 11.5f_head, 0/1 completed
2018-02-28 08:09:56.622137 7fd5b7f2b400  1 journal close /var/lib/ceph/osd/cephprod-170/journal

================================================================================================================
==============
```

#17 - 02/28/2018 10:33 AM - Yoann Moulin

is that normal all files in 11.5f_head have size=0 on each replicate of the PG ?

```
root@iccluster020:/var/lib/ceph/osd/cephprod-78/current/11.5f_head# ll
total 16
drwxr-xr-x   2 ceph ceph   110 Feb  8 03:10 ./
drwxr-xr-x 386 ceph ceph 12288 Feb  7 16:20 ../
-rw-r--r--   1 ceph ceph     0 Feb  8 03:10 \.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8ED
F__b
-rw-r--r--   1 ceph ceph     0 Feb  7 16:20 __head_0000005F__b

root@iccluster003:/var/lib/ceph/osd/cephprod-154/current/11.5f_head# ll
total 20
drwxr-xr-x   2 ceph ceph   110 Feb  3 01:57 ./
drwxr-xr-x 372 ceph ceph 16384 Feb  7 16:20 ../
-rw-r--r--   1 ceph ceph     0 Feb  3 01:57 \.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8ED
F__b
-rw-r--r--   1 ceph ceph     0 Jan 31 17:26 __head_0000005F__b

root@iccluster021:/var/lib/ceph/osd/cephprod-170/current/11.5f_head# ll
total 20
drwxr-xr-x   2 ceph ceph   110 Feb  3 01:57 ./
drwxr-xr-x 362 ceph ceph 16384 Feb  8 00:47 ../
-rw-r--r--   1 ceph ceph     0 Feb  3 01:57 \.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8ED
F__b
-rw-r--r--   1 ceph ceph     0 Jan 31 17:26 __head_0000005F__b
```

Yoann

#18 - 03/01/2018 01:11 AM - Brad Hubbard

Can you dump the object with something like the following.

```
ceph-objectstore-tool --cluster cephprod --data-path /var/lib/ceph/osd/cephprod-170 --pgid 11.5f --journal /va
r/lib/ceph/osd/cephprod-170/journal --debug .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8EDF
__b list-omap dump
```

Can you also "getfattr -d -e hex" to list the extended attributes for the files on disk?

#19 - 03/01/2018 01:55 AM - David Zafman

Yoann Moulin wrote:

> is that normal all files in 11.5f_head have size=0 on each replicate of the PG ?
>
> [...]
>
> Yoann

In this case the object is used for storing omap keys in rocksdb.

To fix this type of omap_digest issue I suggest the following steps with all OSDs up and in.

1. rados -p default.rgw.buckets.index setomapval .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19 temporary-key anything
2. ceph pg deep-scrub 11.5f ; ceph -w
   ...
   2018-02-28 12:57:17.409463 osd.1 [INF] 1.0 deep-scrub starts
   2018-02-28 12:57:17.418817 osd.1 [INF] 1.0 deep-scrub ok
   ^C
3. rados -p default.rgw.buckets.index rmomapkey .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19 temporary-key

#20 - 03/01/2018 09:06 AM - Yoann Moulin

- File osd-170.log added

- File osd-78.log added

- File osd-154.log added

in attachment the result of the dump for each OSD

and the extended attributes for the files on disk :

```
root@iccluster020:/var/lib/ceph/osd/cephprod-78/current/11.5f_head# getfattr -d -e hex  *
# file: \134.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8EDF__b
user.ceph._=0x11082c0100000403540000000000000330000002e6469722e63393732346166662d356661302d346464392d62343934
2d353762646234386661623465 2e3331343532382e3139feffffffffffffffdf8e7f0800000000000b0000000000000006031c0000000b
00000000000000ffffffff000000000000000ffffffffffffffff00000000d46531010000000009d10100d36531010000000009d10100
02021500000008d4647601000000 00ac0900000000000000000000000000000000000000b18e975aade472200202150000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000d4
user.ceph._@1=0x65310100000000000000000000000001c000000b18e975a619e7a20ffffffffffffffff0000000000000000000000000
000000000000000000
user.ceph.snapset=0x03021d0000000000000000000000010000000000000000000000000000000000000000000000
user.cephos.seq=0x010110000000acb1130000000000000000000200000000
user.cephos.spill_out=0x3000

# file: __head_0000005F__b
user.cephos.spill_out=0x3000
```

```
root@iccluster003:/var/lib/ceph/osd/cephprod-154/current/11.5f_head# getfattr -d -e hex  *
# file: \134.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8EDF__b
user.ceph._=0x11082c0100000403540000000000000330000002e6469722e63393732346166662d356661302d346464392d62343934
2d353762646234386616234652e3331343532382e3139feffffffffffffffdf8e7f0800000000000b0000000000000006031c0000000b
00000000000000ffffffff000000000000000000ffffffffffffffff00000000d46531010000000009d10100d36531010000000009d10100
02021500000008d464760100000000ac0900000000000000000000000000000000000000b18e975aade47220020215000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000d4
user.ceph._@1=0x653101000000000000000000000000001c000000b18e975a619e7a20ffffffffffffffff000000000000000000000000
000000000000000000
user.ceph.snapset=0x03021d0000000000000000000000010000000000000000000000000000000000000000000000
user.cephos.seq=0x01011000000072d154060000000000000000200000000
user.cephos.spill_out=0x3000

# file: __head_0000005F__b
user.cephos.spill_out=0x3000



root@iccluster021:/var/lib/ceph/osd/cephprod-170/current/11.5f_head# getfattr -d -e hex  *
# file: \134.dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19__head_087F8EDF__b
user.ceph._=0x11082c0100000403540000000000000330000002e6469722e63393732346166662d356661302d346464392d62343934
2d353762646234386616234652e3331343532382e3139feffffffffffffffdf8e7f0800000000000b0000000000000006031c0000000b
00000000000000ffffffff000000000000000000ffffffffffffffff00000000d46531010000000009d10100d36531010000000009d10100
02021500000008d464760100000000ac0900000000000000000000000000000000000000b18e975aade47220020215000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000d4
user.ceph._@1=0x653101000000000000000000000000001c000000b18e975a619e7a20ffffffffffffffff000000000000000000000000
000000000000000000
user.ceph.snapset=0x03021d0000000000000000000000010000000000000000000000000000000000000000000000
user.cephos.seq=0x01011000000022613e000000000000000000200000000
user.cephos.spill_out=0x3000

# file: __head_0000005F__b
user.cephos.spill_out=0x3000
```

Thanks

#21 - 03/01/2018 09:08 AM - Yoann Moulin

David Zafman wrote:

> Yoann Moulin wrote:
>
> > is that normal all files in 11.5f_head have size=0 on each replicate of the PG ?
> >
> > [...]
> >
> > Yoann
>
> In this case the object is used for storing omap keys in rocksdb.
>
> To fix this type of omap_digest issue I suggest the following steps with all OSDs up and in.
>
> 1. rados -p default.rgw.buckets.index setomapval .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19 temporary-key anything
> 2. ceph pg deep-scrub 11.5f ; ceph -w
>   ...
>   2018-02-28 12:57:17.409463 osd.1 [INF] 1.0 deep-scrub starts
>   2018-02-28 12:57:17.418817 osd.1 [INF] 1.0 deep-scrub ok
>   ^C
> 3. rados -p default.rgw.buckets.index rmomapkey .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19 temporary-key

What do those commands do exactly ?

Yoann

#22 - 03/01/2018 09:26 AM - Yoann Moulin

- File osd-154_dump.log added

- File osd-78_dump.log added

- File osd-170_dump.log added

in attachment the result of the dump for each OSD with the good args

#23 - 03/01/2018 09:55 AM - Yoann Moulin

Yoann Moulin wrote:

> David Zafman wrote:
>
> > Yoann Moulin wrote:
> >
> > > is that normal all files in 11.5f_head have size=0 on each replicate of the PG ?
> > >
> > > [...]
> > >
> > > Yoann
> >
> > In this case the object is used for storing omap keys in rocksdb.
> >
> > To fix this type of omap_digest issue I suggest the following steps with all OSDs up and in.
> >
> > 1. rados -p default.rgw.buckets.index setomapval .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19 temporary-key anything
> > 2. ceph pg deep-scrub 11.5f ; ceph -w
> >   ...
> >   2018-02-28 12:57:17.409463 osd.1 [INF] 1.0 deep-scrub starts
> >   2018-02-28 12:57:17.418817 osd.1 [INF] 1.0 deep-scrub ok
> >   ^C
> > 3. rados -p default.rgw.buckets.index rmomapkey .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19 temporary-key

Thanks David, my cluster is HEALTH_OK now

I ran the command following command as suggested :

```
$ rados -p default.rgw.buckets.index setomapval .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19 temporary-key anything
$ ceph pg deep-scrub 11.5f
```

I got this in ceph -w :

```
2018-03-01 10:46:21.350430 mon.iccluster002.iccluster.epfl.ch [ERR] overall HEALTH_ERR 3 scrub errors; Possible data damage: 1 pg inconsistent
2018-03-01 10:46:25.310039 mon.iccluster002.iccluster.epfl.ch [INF] Health check cleared: OSD_SCRUB_ERRORS (was: 3 scrub errors)
2018-03-01 10:46:25.310097 mon.iccluster002.iccluster.epfl.ch [INF] Health check cleared: PG_DAMAGED (was: Possible data damage: 1 pg inconsistent)
2018-03-01 10:46:25.310133 mon.iccluster002.iccluster.epfl.ch [INF] Cluster is now healthy
2018-03-01 10:47:21.350629 mon.iccluster002.iccluster.epfl.ch [INF] overall HEALTH_OK
2018-03-01 10:47:49.367624 mon.iccluster002.iccluster.epfl.ch [INF] mon.1 10.90.37.11:6789/0
2018-03-01 10:47:49.367711 mon.iccluster002.iccluster.epfl.ch [INF] mon.2 10.90.37.19:6789/0
```

then I removed the temporary-key

```
rados -p default.rgw.buckets.index rmomapkey .dir.c9724aff-5fa0-4dd9-b494-57bdb48fab4e.314528.19 temporary-key
```

re run ceph pg deep-scrub 11.5f

and got this in the osd-78 log file :

```
2018-03-01 10:49:18.916969 7fa609f15700  0 log_channel(cluster) log [DBG] : 11.5f deep-scrub starts
2018-03-01 10:49:19.129079 7fa609f15700  0 log_channel(cluster) log [DBG] : 11.5f deep-scrub ok
```

Thanks a lot again for your help !

Yoann

#24 - 03/21/2018 05:57 PM - Ryan Anstey

I'm trying to fix my problems but I'm kind of a noob, having trouble getting things to work. My cluster seems to be different than Yoann's, so I'm hoping I'm modifying my commands correctly.

```
2018-03-21 10:12:23.434810 osd.13 osd.13 192.168.1.103:6808/5264 723 : cluster [ERR] 0.66 shard 8: soid 0:662d
1777:::rb.0.854e.238e1f29.000000140b6d:head data_digest 0xf019dc86 != data_digest 0xc55f058d from auth oi 0:66
2d1777:::rb.0.854e.238e1f29.000000140b6d:head(18685'550335 client.6158162.0:2246141 dirty|data_digest|omap_dig
est s 4194304 uv 550335 dd c55f058d od ffffffff alloc_hint [0 0 0])
2018-03-21 10:12:23.434815 osd.13 osd.13 192.168.1.103:6808/5264 724 : cluster [ERR] 0.66 shard 13: soid 0:662
d1777:::rb.0.854e.238e1f29.000000140b6d:head data_digest 0xf019dc86 != data_digest 0xc55f058d from auth oi 0:6
62d1777:::rb.0.854e.238e1f29.000000140b6d:head(18685'550335 client.6158162.0:2246141 dirty|data_digest|omap_di
gest s 4194304 uv 550335 dd c55f058d od ffffffff alloc_hint [0 0 0])
2018-03-21 10:12:23.434818 osd.13 osd.13 192.168.1.103:6808/5264 725 : cluster [ERR] 0.66 shard 17: soid 0:662
d1777:::rb.0.854e.238e1f29.000000140b6d:head data_digest 0xf019dc86 != data_digest 0xc55f058d from auth oi 0:6
62d1777:::rb.0.854e.238e1f29.000000140b6d:head(18685'550335 client.6158162.0:2246141 dirty|data_digest|omap_di
gest s 4194304 uv 550335 dd c55f058d od ffffffff alloc_hint [0 0 0])
2018-03-21 10:12:23.434821 osd.13 osd.13 192.168.1.103:6808/5264 726 : cluster [ERR] 0.66 soid 0:662d1777:::rb
.0.854e.238e1f29.000000140b6d:head: failed to pick suitable auth object
2018-03-21 10:15:51.382696 mon.ce1 mon.0 192.168.1.101:6789/0 2739 : cluster [ERR] Health check update: Possib
le data damage: 4 pgs inconsistent (PG_DAMAGED)
2018-03-21 10:15:47.386268 osd.13 osd.13 192.168.1.103:6808/5264 727 : cluster [ERR] 0.66 repair 3 errors, 0 f
ixed
```

Referencing the code above, I ran

```
rados -p rbd setomapval rb.0.854e.238e1f29.000000140b6d temporary-key anything
ceph pg deep-scrub 0.66
```

Still broken, but waiting for deep scrub to finish.

#25 - 03/21/2018 06:15 PM - David Zafman

You have a data_digest issue not an omap_digest one. You can remove the temporary omap entry. Since shards 8, 13 and 17 all have data_digest of 0xf019dc86 then presumably the object info data digest of 0xc55f058d is somehow wrong. With client activity stopped, read the data from this object and write it again using rados get then rados put.

#26 - 03/21/2018 08:38 PM - Ryan Anstey

David Zafman wrote:

> With client activity stopped, read the data from this object and write it again using rados get then rados put.

I can't seem to put the object back on the cluster. Not sure if it's supposed to take a really long time or not?

```
# fix_osd=8
# bad_object=rb.0.854e.238e1f29.000000140b6d
# service ceph-osd@${fix_osd} stop
# ceph-objectstore-tool --data-path /var/lib/ceph/osd/ceph-${fix_osd} ${bad_object} get-bytes /tmp/badobj.obj
Error getting attr on : 0.17d_head,#-2:be800000:::scrub_0.17d:head#, (61) No data available
# rados -p rbd put ${bad_object} /tmp/badobj.obj
```

(...waiting for about 30-60 mins so far)

#27 - 03/21/2018 09:00 PM - David Zafman

The time to write should be on the same order as reading.

You forgot to restart your osd before running rados.

#28 - 03/21/2018 09:31 PM - Ryan Anstey

I'm trying the same thing on a different broken pg, while it's stuck the pg detail is:

```
# ceph health detail | grep 1a8
    pg 0.1a8 is stuck stale for 137.823084, current state stale+undersized+degraded+inconsistent+peered, last
acting [2]
    pg 0.1a8 is stale+undersized+degraded+inconsistent+peered, acting [2]
    pg 0.1a8 is stuck undersized for 147.980149, current state stale+undersized+degraded+inconsistent+peered,
last acting [2]
```

However, when I start up all the OSDs during `rados put` (I had them down for the object dump), the command finishes. Not sure if it worked or not though? I'm not sure if all OSDs are supposed to be brought up before `rados put`.

#29 - 03/21/2018 10:08 PM - Ryan Anstey

So for some reason my rados get/put commands are working now, not sure why. After I complete all my steps, the repair commands don't seem to initiate a repair. I've run a deep scrub to see if that does anything. I can only wait.

My notes (that may or may not work) for my recovery are here, in case I'm doing anything wrong or if anyone needs it:

```
Obvious reminder: Find and replace all pgs in this document with the new ID before continuing.

# Find the object ID

Option 1:

- `cat /var/log/ceph/ceph.log | grep 0.1d7 | grep fail`

Option 2:

- Initiate repair `ceph pg repair 0.1d7`
- Watch the logs and get the object ID: `tail -f /var/log/ceph/ceph.log`later with
- I've got `rb.0.854e.238e1f29.000000149eed`

# Prepare for get/put

- Obvious reminder: Find and replace all objects IDs in this document with the new ID before continuing.
- (Make sure all my vm's are turned off.)
- Disable recovery `ceph osd set noout`
- Get OSDs the pg is on `ceph health detail | grep 0.1d7`
- Find the server for each:

```
ceph osd find 4
ceph osd find 15
ceph osd find 12
```

- Update below, then run on respective servers:

```
# CE1
fix_osd=12
bad_object=rb.0.854e.238e1f29.000000140b6d
service ceph-osd@${fix_osd} stop

# CE3
fix_osd=4
bad_object=rb.0.854e.238e1f29.000000140b6d
service ceph-osd@${fix_osd} stop

# CE4
fix_osd=15
bad_object=rb.0.854e.238e1f29.000000140b6d
service ceph-osd@${fix_osd} stop
```

Update on pg status

Health detail reports it as `pg 0.1d7 is stale+undersized+degraded+inconsistent+peered, acting [6]`

rados

```
rados -p rbd stat ${bad_object}
rados -p rbd get ${bad_object} /tmp/${bad_object}
ls -l /tmp/${bad_object}
rados -p rbd put ${bad_object} /tmp/${bad_object}
```

Clean up

- Start up osds ON EACH SERVER: `service ceph-osd@${fix_osd} start`
- Enable recovery: `ceph osd unset noout`
- Wait for recovery (from the downed osds).
- Repair the bad OSD `ceph pg repair 0.1d7` (for some reason the repair doesn't start?) Try deep scrub `ceph p
g deep-scrub 0.1d7` instead?
- Check if it's okay `cat /var/log/ceph/ceph.log | grep 0.1d7 | grep errors`
```

```
(Not sure if problem solved yet...)
```

#30 - 03/21/2018 10:39 PM - Ryan Anstey

So the stuff I did above did not work, the result of the repair after get/put:

```
9
2018-03-21 15:17:01.877225 osd.4 osd.4 192.168.1.103:6804/8355 4 : cluster [ERR] 0.1d7 shard 4: soid 0:ebdd235
c:::rb.0.854e.238e1f29.000000149eed:head data_digest 0x21792920 != data_digest 0x17ba3aab from auth oi 0:ebdd2
35c:::rb.0.854e.238e1f29.000000149eed:head(16876'632423 osd.3.0:63207 dirty|data_digest|omap_digest s 4194304
uv 57087 dd 17ba3aab od ffffffff alloc_hint [0 0 0])
2018-03-21 15:17:01.877230 osd.4 osd.4 192.168.1.103:6804/8355 5 : cluster [ERR] 0.1d7 shard 12: soid 0:ebdd23
5c:::rb.0.854e.238e1f29.000000149eed:head data_digest 0x21792920 != data_digest 0x17ba3aab from auth oi 0:ebdd
235c:::rb.0.854e.238e1f29.000000149eed:head(16876'632423 osd.3.0:63207 dirty|data_digest|omap_digest s 4194304
 uv 57087 dd 17ba3aab od ffffffff alloc_hint [0 0 0])
2018-03-21 15:17:01.877233 osd.4 osd.4 192.168.1.103:6804/8355 6 : cluster [ERR] 0.1d7 shard 15: soid 0:ebdd23
5c:::rb.0.854e.238e1f29.000000149eed:head data_digest 0x21792920 != data_digest 0x17ba3aab from auth oi 0:ebdd
235c:::rb.0.854e.238e1f29.000000149eed:head(16876'632423 osd.3.0:63207 dirty|data_digest|omap_digest s 4194304
 uv 57087 dd 17ba3aab od ffffffff alloc_hint [0 0 0])
2018-03-21 15:17:01.877235 osd.4 osd.4 192.168.1.103:6804/8355 7 : cluster [ERR] 0.1d7 soid 0:ebdd235c:::rb.0.
854e.238e1f29.000000149eed:head: failed to pick suitable auth object
```

Any idea what I'm doing wrong?

#31 - 03/21/2018 11:00 PM - David Zafman

I don't know how rados get/put could work while the PGs OSDs are all stopped. Also, 'rados get' will give EIO error if data_digest of data doesn't match the object info unless the object size exceeds the default rados read size. (4194304)

Also, you could use 'rados list-inconsistent-obj <pgid>' to see what scrub errors are found instead of searching osd logs.

#32 - 03/21/2018 11:12 PM - Ryan Anstey

Ah, I got confused by stuff. Should I just not stop the OSDs, or just stop during the get, start for the put?

I just tried it with all OSDs up for get and put, deep-scrubbed and still the same issue (failed to pick suitable auth object).

#33 - 03/21/2018 11:38 PM - David Zafman

I'm confused because you are dealing with 2 different objects.

Does rb.0.854e.238e1f29.000000140b6d still have a data_digest problem?
Does rb.0.854e.238e1f29.000000149eed still have a data_digest problem?

1.

ceph osd set noout
ceph osd set pause
stop OSD
ceph-objectstore-tool .... get-bytes
start OSD
ceph osd unset noout
ceph osd unset pause

1. client operations better not be happening or you'll undo them here
   rados ... put...

2.

1. Using a size smaller than the object size if object size <= 4194304
   rados ... -b size get...
   rados ... put ...

#34 - 03/22/2018 02:02 AM - Ryan Anstey

Got it! I couldn't use the pause feature because none of the get/get-bytes/stat stuff would work, it all got stuck.

However, I seem to have pasted some dumb commands in and ended up wiping out one of the objects. (I "getted" a file that ended up being 0 bytes and then put it in the cluster, ugh.) Do I just mark this as unfound/lost?

Not sure what to do when I break an object.

```
# rados -p rbd stat ${ceph_object}
rbd/rb.0.854e.238e1f29.0000001efdb6 mtime 2018-03-21 18:20:17.000000, size 0
```

Ran a pg repair with no luck.

#35 - 02/07/2019 07:23 PM - David Zafman

- Project changed from Ceph to RADOS

- Category deleted (OSD)

#36 - 04/11/2019 04:48 PM - David Zafman

- Duplicates Feature #25085: Allow repair of an object with a bad data_digest in object_info on all replicas added

#37 - 04/11/2019 04:50 PM - David Zafman

- Status changed from 12 to Duplicate

This was merged to master Jul 31, 2018 in https://github.com/ceph/ceph/pull/23217 for a different tracker.

Files

osd-170.log	23.7 KB	03/01/2018	Yoann Moulin
osd-78.log	52.2 KB	03/01/2018	Yoann Moulin
osd-154.log	14.9 KB	03/01/2018	Yoann Moulin
osd-154_dump.log	15 KB	03/01/2018	Yoann Moulin
osd-78_dump.log	32.2 KB	03/01/2018	Yoann Moulin
osd-170_dump.log	15.2 KB	03/01/2018	Yoann Moulin