ceph pg inconsistent问题修复方法

# unexpected clone

## 现象

2020-09-03 13:45:24.851042 7f69209d6700 -1 log\_channel(cluster) log [ERR] : deep-scrub 2.7f 2:fe380759:::rbd\_data.c99f4850a276

.000000000000105b:ad45 is an unexpected clone

2020-09-03 13:45:24.851565 7f69209d6700 -1 log\_channel(cluster) log [ERR] : deep-scrub 2.7f 2:fe380759:::rbd\_data.c99f4850a276

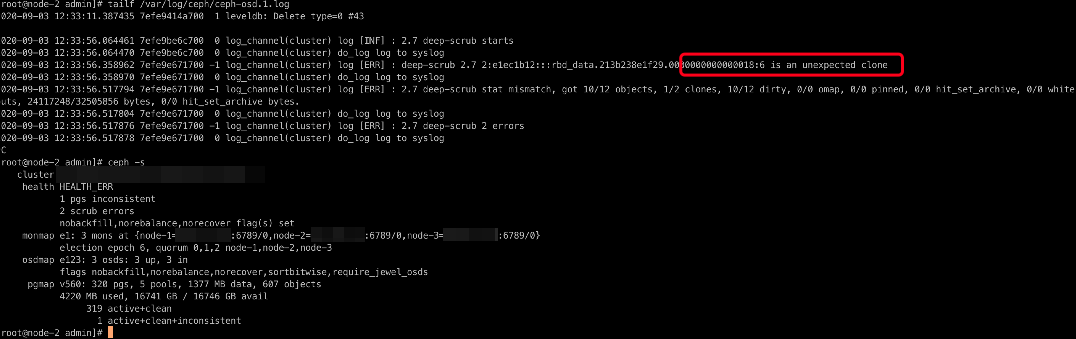
.000000000000105b:ab6d is an unexpected clone

## 原因

快照数据丢失，kv db中的数据还在

模拟复现方法：

1. 创建一个卷
2. 从卷创建一个快照
3. 找到快照数据目录直接用rm命令删除（3副本同时删除，否则repair会自动修复）



## 解决方法

osd报错unexpected clone，快照删除步骤

1. 查看deep-scrub执行后的osd log找到有问题的object id

例如：ceph pg deep-scrub 2.7f

2020-09-03 13:45:24.851042 7f69209d6700 -1 log\_channel(cluster) log [ERR] : deep-scrub 2.7f 2:fe380759:::rbd\_data.c99f4850a276

.000000000000105b:ad45 is an unexpected clone

2020-09-03 13:45:24.851565 7f69209d6700 -1 log\_channel(cluster) log [ERR] : deep-scrub 2.7f 2:fe380759:::rbd\_data.c99f4850a276

.000000000000105b:ab6d is an unexpected clone

这里有2个快照对象

rbd\_data.c99f4850a276.000000000000105b:ad45

rbd\_data.c99f4850a276.000000000000105b:ab6d

2. 设置集群进入维护模式

cat /var/admin/setflag.sh

#!/bin/bash

ceph osd set norebalance

ceph osd set norecover

ceph osd set nobackfill

3. 找到pg存放的3个osd，依次操作，每次只能停1个osd

例如：ceph pg map 2.7f

osdmap e72817 pg 2.7f (2.7f) -> up [10,55,42] acting [10,55,42]

或者看集群状态信息

ceph health detail

HEALTH\_ERR 5 pgs inconsistent; 2657 scrub errors

pg 2.2e2 is active+clean+inconsistent, acting [0,22,55]

pg 2.249 is active+clean+inconsistent, acting [55,2,28]

pg 2.367 is active+clean+inconsistent, acting [31,17,65]

pg 2.3c6 is active+clean+inconsistent, acting [66,57,20]

pg 2.7f is active+clean+inconsistent, acting [10,55,42]

4. 登录到节点并停止osd进程

例如：ceph osd find 10 说明osd.10在node-7上

{

"osd": 10,

"ip": "x.x.x.x:6800\/3978063",

"crush\_location": {

"host": "node-7",

"root": "default"

}

}

ssh node-7

systemctl stop ceph-osd@10

5. 找到有问题的快照对象,43885和44357转成16进制后分别为0xab6d和0xad45,也就是上面日志中找到的快照对象

ls -l /var/lib/ceph/osd/ceph-10/current/2.7f\_head | grep 000000000000105b

ceph-objectstore-tool --pgid 2.7f --data-path /var/lib/ceph/osd/ceph-10/ --op list | grep 000000000000105b

ceph-objectstore-tool --pgid 2.7f --data-path /var/lib/ceph/osd/ceph-10/ '["2.7f",{"oid":"rbd\_data.c99f4850a276.0000000

00000105b","key":"","snapid":43885,"hash":2598378623,"max":0,"pool":2,"namespace":"","max":0}]' remove

ceph-objectstore-tool --pgid 2.7f --data-path /var/lib/ceph/osd/ceph-10/ '["2.7f",{"oid":"rbd\_data.c99f4850a276.0000000

00000105b","key":"","snapid":44357,"hash":2598378623,"max":0,"pool":2,"namespace":"","max":0}]' remove

6. 重启osd，退出维护模式

systemctl start ceph-osd@10

cat /var/admin/unsetflag.sh

#!/bin/bash

ceph osd unset norebalance

ceph osd unset norecover

ceph osd unset nobackfill

7. ceph -s观察集群状态，等待恢复，确保所有object都是active+clean后操作另外2个osd

8. 再次执行deep-scrub，看日志object报错信息消失，同时inconsistent object减少，说明有问题的快照已删除。

# missing clone

## 现象

osd日志报错信息如下：

/var/log/ceph/ceph-osd.0.log

2020-09-03 17:48:46.733341 7f2088f04700 -1 log\_channel(cluster) log [ERR] : deep-scrub 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:4

74e542d:::rbd\_data.fc90b3cae5697.0000000000000890:ad1a

2020-09-03 17:48:46.733343 7f2088f04700 0 log\_channel(cluster) do\_log log to syslog

2020-09-03 17:48:46.733393 7f2088f04700 -1 log\_channel(cluster) log [ERR] : deep-scrub 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:4

74e542d:::rbd\_data.fc90b3cae5697.0000000000000890:ac2f

2020-09-03 17:48:46.733395 7f2088f04700 0 log\_channel(cluster) do\_log log to syslog

2020-09-03 17:48:46.733416 7f2088f04700 -1 log\_channel(cluster) log [ERR] : deep-scrub 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:4

74e542d:::rbd\_data.fc90b3cae5697.0000000000000890:ab47

2020-09-03 17:48:46.733418 7f2088f04700 0 log\_channel(cluster) do\_log log to syslog

2020-09-03 17:48:46.733448 7f2088f04700 -1 log\_channel(cluster) log [ERR] : deep-scrub 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:4

74e542d:::rbd\_data.fc90b3cae5697.0000000000000890:aa50

2020-09-03 17:48:46.733450 7f2088f04700 0 log\_channel(cluster) do\_log log to syslog

2020-09-03 17:48:46.733468 7f2088f04700 0 log\_channel(cluster) log [INF] : deep-scrub 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head 5 missing clone(s)

2020-09-03 17:48:46.733469 7f2088f04700 0 log\_channel(cluster) do\_log log to syslog

2020-09-03 17:49:24.979128 7f2088f04700 -1 log\_channel(cluster) log [ERR] : 2.2e2 deep-scrub stat mismatch, got 645/645 objects, 190/190 clones, 645/645 dirty, 0/0 omap,

0/0 pinned, 0/0 hit\_set\_archive, 0/0 whiteouts, 2100887040/2101206528 bytes, 0/0 hit\_set\_archive bytes.

2020-09-03 17:49:24.979144 7f2088f04700 0 log\_channel(cluster) do\_log log to syslog

2020-09-03 17:49:24.979246 7f2088f04700 -1 log\_channel(cluster) log [ERR] : 2.2e2 deep-scrub 6 errors

2020-09-03 17:49:24.979249 7f2088f04700 0 log\_channel(cluster) do\_log log to syslog

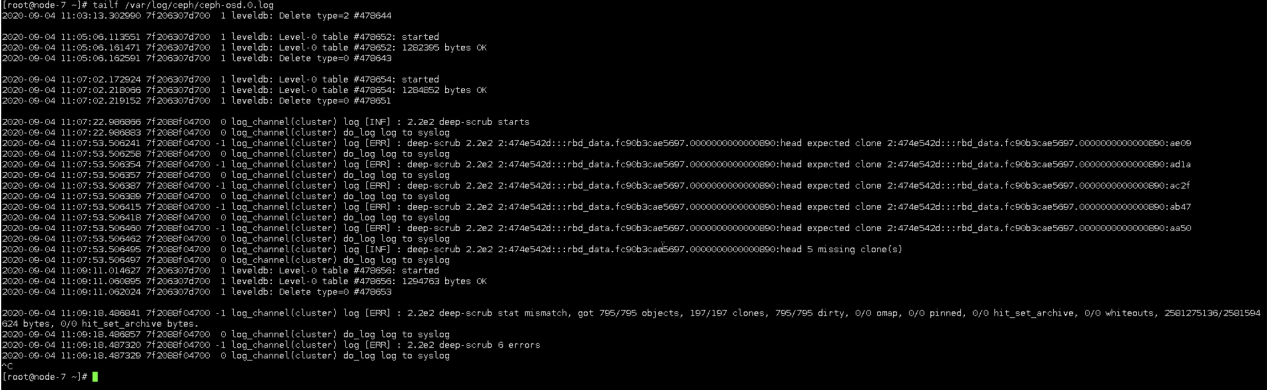
## 原因

中间快照数据丢失

模拟复现方法：

1. 创建一个卷
2. 从卷创建一个快照
3. 从快照克隆一个卷

删除中间快照的某个object，触发deep-scrub即可复现



## 解决方法

和问题1方法解决方法相同，不同的是可能会有多个对象需要删除，丢失的快照可能涉及多个对象，删除过程中注意观察日志

# expected clone

## 现象

2020-09-09 10:36:52.803501 7f1574f93700 0 log\_channel(cluster) log [INF] : 2.2e2 repair starts

2020-09-09 10:36:52.803516 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941316 7f1574f93700 -1 log\_channel(cluster) log [ERR] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:bd95

2020-09-09 10:37:31.941336 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941525 7f1574f93700 -1 log\_channel(cluster) log [ERR] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:bcb8

2020-09-09 10:37:31.941530 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941584 7f1574f93700 -1 log\_channel(cluster) log [ERR] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:bbe4

2020-09-09 10:37:31.941586 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941612 7f1574f93700 -1 log\_channel(cluster) log [ERR] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:bb07

2020-09-09 10:37:31.941614 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941642 7f1574f93700 -1 log\_channel(cluster) log [ERR] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:ba33

2020-09-09 10:37:31.941644 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941670 7f1574f93700 -1 log\_channel(cluster) log [ERR] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:ae09

2020-09-09 10:37:31.941673 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941697 7f1574f93700 -1 log\_channel(cluster) log [ERR] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:ad1a

2020-09-09 10:37:31.941700 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941725 7f1574f93700 -1 log\_channel(cluster) log [ERR] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:ac2f

2020-09-09 10:37:31.941727 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941751 7f1574f93700 -1 log\_channel(cluster) log [ERR] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:ab47

2020-09-09 10:37:31.941754 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941843 7f1574f93700 -1 log\_channel(cluster) log [ERR] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head expected clone 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:aa50

2020-09-09 10:37:31.941851 7f1574f93700 0 log\_channel(cluster) do\_log log to syslog

2020-09-09 10:37:31.941902 7f1574f93700 0 log\_channel(cluster) log [INF] : repair 2.2e2 2:474e542d:::rbd\_data.fc90b3cae5697.0000000000000890:head 10 missing clone(s)

## 原因

对象快照数据丢失，kv db中残留快照信息

## 解决方法

参考问题1的方法删除各osd上的clone-metadata后再repair

ceph-objectstore-tool --pgid <pgid> --data-path /var/lib/ceph/osd/ceph-<osd-id>/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata <object\_snap\_id>

Removal of clone bcb8 complete

Use pg repair after OSD restarted to correct stat information

示例：

ceph-objectstore-tool --pgid 2.2e2 --data-path /var/lib/ceph/osd/ceph-0/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata 48533

ceph-objectstore-tool --pgid 2.2e2 --data-path /var/lib/ceph/osd/ceph-0/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata 48312

ceph-objectstore-tool --pgid 2.2e2 --data-path /var/lib/ceph/osd/ceph-0/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata 48100

ceph-objectstore-tool --pgid 2.2e2 --data-path /var/lib/ceph/osd/ceph-0/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata 47879

ceph-objectstore-tool --pgid 2.2e2 --data-path /var/lib/ceph/osd/ceph-0/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata 47667

ceph-objectstore-tool --pgid 2.2e2 --data-path /var/lib/ceph/osd/ceph-0/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata 44553

ceph-objectstore-tool --pgid 2.2e2 --data-path /var/lib/ceph/osd/ceph-0/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata 44314

ceph-objectstore-tool --pgid 2.2e2 --data-path /var/lib/ceph/osd/ceph-0/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata 44079

ceph-objectstore-tool --pgid 2.2e2 --data-path /var/lib/ceph/osd/ceph-0/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata 43847

ceph-objectstore-tool --pgid 2.2e2 --data-path /var/lib/ceph/osd/ceph-0/ rbd\_data.fc90b3cae5697.0000000000000890 remove-clone-metadata 43600

注意，object\_snap\_id需要将16进制转换成10进制后再remove-clone-metada

例如0xbd95 -> 48533

# dirty|data\_digest|omap\_digest

## 现象

2020-09-02 16:48:58.942996 7f69209d6700 -1 log\_channel(cluster) log [ERR] : 2.7f shard 42: soid 2:fe04b2ba:::rbd\_data.5fcd567875eeb.00000000000065d3:aaf9 data\_digest 0x43d61c5d != data\_digest 0x6b8aa19c from shard 10, data\_digest 0x43d61c5d != data\_digest 0x6b8aa19c from auth oi 2:fe04b2ba:::rbd\_data.5fcd567875eeb.00000000000065d3:aaf9(59131'4398942 osd.10.0:148516892 [aaf9] dirty|data\_digest|omap\_digest s 4194304 uv 4353559 dd 6b8aa19c od ffffffff)

2020-09-02 16:48:58.943131 7f69209d6700 -1 log\_channel(cluster) log [ERR] : 2.7f shard 55: soid 2:fe04b2ba:::rbd\_data.5fcd567875eeb.00000000000065d3:aaf9 data\_digest 0x43d61c5d != data\_digest 0x6b8aa19c from shard 10, data\_digest 0x43d61c5d != data\_digest 0x6b8aa19c from auth oi 2:fe04b2ba:::rbd\_data.5fcd567875eeb.00000000000065d3:aaf9(59131'4398942 osd.10.0:148516892 [aaf9] dirty|data\_digest|omap\_digest s 4194304 uv 4353559 dd 6b8aa19c od ffffffff)

使用命令也可以看到相同的报错信息

rados list-inconsistent-obj <pgid> --format json-pretty

{

"object": {

"name": "rbd\_data.783f254cba81.0000000000004cda",

"nspace": "",

"locator": "",

"snap": 22905,

"version": 2220729

},

"errors": [

"data\_digest\_mismatch"

],

"union\_shard\_errors": [

"data\_digest\_mismatch\_oi"

],

"selected\_object\_info": "2:fe0bd7c5:::rbd\_data.783f254cba81.0000000000004cda:5979(30644'2304580 osd.10.0:56968876 [5979,586b] dirty|data\_digest|omap\_digest s 4194304 uv 2220729 dd 8a530ecf od ffffffff)",

"shards": [

{

"osd": 10,

"errors": [],

"size": 4194304,

"omap\_digest": "0xffffffff",

"data\_digest": "0x8a530ecf"

},

{

"osd": 42,

"errors": [

"data\_digest\_mismatch\_oi"

],

"size": 4194304,

"omap\_digest": "0xffffffff",

"data\_digest": "0x43d61c5d"

},

{

"osd": 55,

"errors": [

"data\_digest\_mismatch\_oi"

],

"size": 4194304,

"omap\_digest": "0xffffffff",

"data\_digest": "0x43d61c5d"

}

]

}

## 原因

3副本中对象校验和crc32值不一致，有些object是数据还有一些数据快照

## 解决方法

1. 查看pg的详细错误信息

rados list-inconsistent-obj <pgid> --format json-pretty

1. 通过rados get/put的方式修复，先把对象取出来

rados get -p volumes rbd\_data.783f254cba81.0000000000004cda rbd\_data.783f254cba81.0000000000004cda

1. 再put回去

rados put -p volumes rbd\_data.783f254cba81.0000000000004cda rbd\_data.783f254cba81.0000000000004cda

# Trim\_object Snap not in clones

## 现象

2020-09-03 23:56:12.341071 7f042ece1700 -1 log\_channel(cluster) log [ERR] : trim\_object Snap aa48 not in clones

## 原因

待分析

## 解决方法

待分析