如何在所有的mon的损坏情况下将数据恢复如初

jc

前置条件：
拥有一个正常的集群，使用了rbd存储池，并且rbd池中存储了一些镜像文件，然后集群的mon出现无法启动的情况

现在需要知道的信息：
原始集群的rbd存储池的pg数(这个在集群创建的时候自己肯定是知道的记录下来，这个最好pg的名称也记录下来)
这个可以用这个脚本:

```
[root@node1 ~]# ceph pg dump pgs|awk '/^[0-9a-f]+\.[0-9a-f]+/  {print $1}' > /root/mypg
```

原始集群的fsid（这个在配置文件里面有/etc/ceph/ceph.conf）

我的测试集群

```
[root@node1 ~]# ceph -s
    cluster 88c14091-07c1-4457-9790-4efbe6417196
     health HEALTH_OK
     monmap e1: 1 mons at {node1=192.168.0.22:6789/0}
            election epoch 2, quorum 0 node1
     osdmap e113: 2 osds: 2 up, 2 in
      pgmap v370: 80 pgs, 1 pools, 580 MB data, 155 objects
            3259 MB used, 37676 MB / 40936 MB avail
                  80 active+clean
```

镜像状态

```
[root@node1 ~]# rbd info test
rbd image 'test':
    size 4000 MB in 1000 objects
    order 22 (4096 kB objects)
```

```
    block_name_prefix: rb.0.19fc.6b8b4567
    format: 1
```

这个镜像对应的数据为：

```
[root@node1 ~]# ll /mnt/
total 155656
-rw-r--r--. 1 root root 159383552 Dec 12 01:07 a
-rw-r--r--. 1 root root       393 Dec 12 01:05 ceph.conf
drwxr-xr-x. 2 root root      4096 Dec 12 01:18 dir
```

现在先停止掉node1整个集群，这个操作是模拟的mon无法启动，整个集群无法启动了

现在准备倒出集群1的全部的数据，这个在实际环境当中是去倒出所有的主本或者副本的数据，这个是每个pg主本或者副本倒出一个即可,我的环境因为只有两个osd，所以只用倒出osd0上的数据即可，这个在大环境的时候需要去各个机器上倒出需要的数据

倒出pg的命令是这个命令：

```
[root@node1 ~]# ceph-objectstore-tool --op export --pgid 0.0 --data-path /var/lib/ceph/osd/ceph-0/ --journal-path /var/lib/ceph/osd/ceph-0/journal --file /mytest/0.0
```

这个地方因为是很多的pg，那么我就用脚本一次倒出来
脚本内容如下：

```
#! /bin/sh
for a in `cat /root/mypg`
do
ceph-objectstore-tool --op export --pgid $a --data-path /var/lib/ceph/osd/ceph-0/ --journal-path /var/lib/ceph/osd/ceph-0/journal --file /mytest/$a
done
```

将这些数据发送到一个新的空的机器上，我的机器为node2，我准备在node2上创建一个新的集群进行数据的恢复

```
[root@node1 ~]# scp -r /mytest node2:/root/
```

在node2上创建新的集群
创建初始化配置文件

```
[root@node2 ~]# mkdir /root/myrecovery
[root@node2 ~]# cd /root/myrecovery
[root@node2 myrecovery]# ceph-deploy new node2
```

修改配置文件的fsid跟原始集群一样的
配置好集群后

修改新的集群的pg数跟原始的集群的pg数一样的

```
[root@node2 myrecovery]# ceph osd pool set rbd pg_num 80
[root@node2 myrecovery]# ceph osd pool set rbd pgp_num 80
```

现在有个问题需要处理下，在导入数据的时候，以为原始的pg数里面记录了osd map epoch的id号，这个导入的时候要求当前的集群的epoch的编号要大于倒出数据的那个时候的id号
如果不处理会显示如下

```
Importing pgid 0.1d
ERROR: Export map_epoch 122 > osd epoch 13
```

这个时候有个办法就是频繁的重启osd这个会增加epoch的编号（这个地方暂时还没有研究出来其他的办法修改那个epoch的编号，目前重启的方式增加）
osdmap e50: 2 osds: 2 up, 2 in
这个e50的编号
重启脚本如下：

```
[root@node2 myrecovery]# seq 10|xargs -i /etc/init.d/ceph restart osd
```

当编号超过原始集群的编号的时候，上面的导入的122这个编号的时候，就可以停止集群的osd

```
[root@node2 myrecovery]# /etc/init.d/ceph stop osd
```

删除集群的pg数据和目录,这个是初始的集群的必须删除掉才能导入，主副本的pg都删除掉

```
[root@node2 myrecovery]# rm -rf /var/lib/ceph/osd/ceph-0/current/0.*
[root@node2 myrecovery]# rm -rf /var/lib/ceph/osd/ceph-1/current/0.*
```

然后导入数据
命令是下面的命令

```
ceph-objectstore-tool --op import  --data-path /var/lib/ceph/osd/ceph-0/ --journal-path /var/lib/ceph/osd/ceph-0/journal --file /root/mytest/0.0
```

这里同样使用脚本的方式，mypg就是之前保留下来的pg的编号,导入的时候同样是主副本的机器的osd都需要导入

```
#! /bin/sh
for a in `cat mypg`
do
ceph-objectstore-tool --op import  --data-path /var/lib/ceph/osd/ceph-0/ --journal-path /var/lib/ceph/osd/ceph-0/journal --file /root/mytest/$a
ceph-objectstore-tool --op import  --data-path /var/lib/ceph/osd/ceph-1/ --journal-path /var/lib/ceph/osd/ceph-1/journal --file /root/mytest/$a
done
```

然后重启新的整个集群，然后去检查下跟之前的集群的数据是不是一致的

是不是很神奇的数据都回来了，以前只以为gluster有如此健壮性，现在发现ceph同样很强大，好了希望谁都不需要去用到这个技术，掌握了就好