목차

CEPH pg,object repair 단계별 정리 – ceph status 메시지 설명 3
MANUALLY REPAIR OBJECT 및 PG Repare 3
상태 3
문제 PG 찾기 3
문제 찾기 3
문제 해당 object 찾기 4
문제 해결 4
ceph -s에서 출력되는 pg 에러 예제 4
문제가 있는 PG 번호를 출력한다. 5
해당 PG들에 대해 repair를 수행한다. 5
PLACEMENT GROUP STATES 5
참조링크 6

CEPH pg,object repair 단계별 정리 – ceph status 메시지 설명

MANUALLY REPAIR OBJECT 및 PG Repare

상태

```
health HEALTH_ERR 1 pgs inconsistent; 2 scrub errors
```

의 상태라고 가정하고 다음을 진행하여 보자.

문제 PG 찾기

```
$ sudo ceph health detail

HEALTH_ERR 1 pgs inconsistent; 2 scrub errors
pg 17.1c1 is active+clean+inconsistent, acting [21,25,30] 2 scrub errors
```

문제시 되는 PG는 17.1c1 이고 이는 OSD 21,25,30 위에 있다.

문제 찾기

문제시 되는 원인을 찾기 위해, OSD logs 파일을 봐야 한다.

```
$ grep -Hn 'ERR' /vat/logs/ceph/ceph-osd.21.log
```

만약 로그가 rotate한다면 grep 대신 zgrep을 사용할 수 있다.

```
log [ERR] : 17.1c1 shard 21: soid
58bcc1c1/rb.0.90213.238e1f29.00000001232d/head//17 digest 0 != known digest
3062795895
log [ERR] : 17.1c1 shard 25: soid
58bcc1c1/rb.0.90213.238e1f29.00000001232d/head//17 digest 0 != known digest
3062795895
```

이 내용은 digest가 3062795895이여야 하는데 실제 digest는 0이다.

문제 해당 object 찾기

위의 1,2,3으로 우리가 알 수 있는 문제점은 아래와 같다.

1. PG: `17.1c1` (문제시되는 PG)
2. `OSD ID` (문제시되는 PG가 위치한 OSD)
3. `rb.0.90213.238e1f29.00000001232d` (문제시 되는 Object name)

위의 내용으로 우리는 Object를 찾을 수 있다.

```
$ sudo find /var/lib/ceph/osd/ceph-21/current/17.1c1_head/ -name
'rb.0.90213.238e1f29.00000001232d*' -ls

671193536 4096 -rw-r—r— 1 root root 4194304 Feb 14 01:05
/var/lib/ceph/osd/ceph-21/current/17.1c1_head/DIR_1/DIR_C/DIR_1/DIR_C/rb.0.9
0213.238e1f29.00000001232d__head_58BCC1C1__1
```

문제시되는 `object size` 와
`MD5(rb.0.90213.238e1f29.00000001232d__head_58BCC1C1__1)`를 볼 수 있다.

문제 해결

1. 문제시 되는 PG를 담당하고 있는 OSD stop
2. 저널 플러시 (`ceph-osd -i —flush-journal`)
3. 다른 위치로 `bad object` 를 이동시켜라.
4. OSD를 start

```
$ ceph pg repair 17.1c1
```

ceph -s에서 출력되는 pg 에러 예제

```
[root@storage004 ~] ceph -s

health HEALTH_ERR
5 pgs inconsistent
9 scrub errors
```

- `active`: Request를 처리할 수 있는 상태
- `clean`: 데이터가 Replica 카운트대로 Replication 되어있는 상태
- `inconsistent`: Replica 중 일부가 잘못된 데이터를 가지고있는 상태. 이전에 **디스크 오류**가 있었을 수 있고 `scrub errors` 카운트가 존재할 수 있다.

문제가 있는 PG 번호를 출력한다.

```
[root@storage004 ~] ceph health detail
HEALTH_ERR 5 pgs inconsistent; 9 scrub errors
pg 9.180 is active+clean+inconsistent, acting [43,36,18] # pg 9.180
pg 9.e1 is active+clean+inconsistent, acting [20,41,23] # pg 9.e1
pg 2.12 is active+clean+inconsistent, acting [24,37,15] # pg 2.12
pg 9.3e3 is active+clean+inconsistent, acting [32,26,7] # pg 9.3e3
pg 9.265 is active+clean+inconsistent, acting [24,10,3] # pg 9.265
9 scrub errors
```

해당 PG들에 대해 repair를 수행한다.

```
[root@storage004 ~] ceph pg repair 9.180
instructing pg 9.180 on osd.43 to repair

[root@storage004 ~] ceph pg repair 9.e1
instructing pg 9.e1 on osd.20 to repair

[root@storage004 ~] ceph pg repair 2.12
instructing pg 2.12 on osd.24 to repair

[root@storage004 ~] ceph pg repair 9.3e3
instructing pg 9.3e3 on osd.32 to repair

[root@storage004 ~] ceph pg repair 9.265
instructing pg 9.265 on osd.24 to repair

# repair 완료
[root@storage004 ~] ceph health detail
HEALTH_OK
```

PLACEMENT GROUP STATES

When checking a cluster's status (e.g., running `ceph -w` or `ceph -s`), Ceph will report on the status of the placement groups. A placement group has one or more states. The optimum state for placement groups in the placement group map is `active + clean`.

- `Creating` – Ceph가 게재 위치그룹을 만드는 중.
- `Active` – Ceph가 게재 위치그룹에 대한 요청 처리중.
- `Clean` – Ceph가 게재 위치그룹의 모든 개체를 올바른 횟수만큼 복제했다.
- `Down` – Ceph가 필요한 데이터가있는 복제본이 다운되어 있으므로 게재 위치 그룹이 오프라인 상태입니다.
- `Replay` – Ceph가 게재 위치 그룹은 OSD가 손상된 후 클라이언트가 작업을 재생할 때까지 기다리고 있습니다.
- `Splitting` – Ceph가 게재 위치 그룹을 여러 게재 위치 그룹으로 분할 중입니다.

- `Scrubbing` – Ceph가 게재 위치 그룹의 불일치를 확인 중입니다.
- `Degraded` – Ceph는 게재 위치 그룹의 일부 개체를 올바른 횟수만큼 복제하지 않았습니다.
- `Inconsistent` – Ceph는 게재 위치 그룹에있는 개체의 하나 이상의 복제본에서 불일치를 감지합니다. (예 : 개체의 크기가 잘못되었거나 복구 완료 후 개체가 하나의 복제본에서 누락 됨 등).
- `Peering` – 게재 위치 그룹이 피어링 프로세스를 진행 중입니다.
- `Repair` – Ceph가 게재 위치 그룹을 확인하고 발견 한 불일치를 수정 중입니다. (if possible).
- `Recovering` – Ceph는 개체와 해당 복제본을 마이그레이션 / 동기화합니다.
- `Backfill` – Ceph는 최근 작업의 로그에서 어떤 내용을 동기화해야 하는지를 추론하는 대신 배치 그룹의 전체 내용을 검색하고 동기화합니다. Backfill은 복구의 특별한 경우입니다.
- `Wait-backfill` – 게재 위치 그룹은`Backfill`를 시작하기 위해 줄을 서서 기다리고 있습니다.
- `Backfill-toofull` -대상 OSD가 전체 비율을 초과하기 때문에 백필 작업이 대기 중입니다.
- `Incomplete` – ceph는 쓰기 또는 정상적인 복사본이 없는 배치 그룹 정보가 발생될 때 정보가 누락되었음을 감지한다. 이 상태가 표기되면 필요한 정보가 포함 된 실패된 OSD를 시작하거나, 일시적으로 `minb_size`를 조정하여 복구 할 수 있다.
- `Stale` – 게재 위치 그룹이 알 수 없는 상태입니다. 게재 위치 그룹 매핑이 변경된 이후에 모니터가 업데이트를받지 못했습니다.
- `Remapped` – 배치 그룹은CRUSH가 지정한 것과 다른 OSD 세트에 임시로 맵핑됩니다.
- `Undersized` – 배치 그룹은 구성된 풀 복제 수준보다 적은 수의 사본입니다.
- `Peered` – 배치 그룹이 피어링되었지만 풀의 구성된`min_size` 매개 변수에 도달하기에 충분한 사본이 없기 때문에 클라이언트 IO를 제공 할 수 없습니다. 이 상태에서 복구가 발생할 수 있으므로 결국 pg가 최대 `min_size`까지 치료할 수 있습니다.

참조링크

- https://idchowto.com/?p=31090
- https://access.redhat.com/solutions/1589113

From:
https://atl.kr/dokuwiki/ - **AllThatLinux!**

Permanent link:
https://atl.kr/dokuwiki/doku.php/ceph_pg_object_repair_%EB%8B%A8%EA%B3%84%EB%B3%84_%EC%A0%95%EB%A6%AC_ceph_status_%EB%A9%94%EC%8B%9C%EC%A7%80_%EC%84%A4%EB%AA%85

Last update: **2021/07/19 12:12**