Ceph OSD

Contents

**** 版本修订记录 ****

版本号	*修订时间*	*修订内容*
v1.0	*2018-08-09*	*初版修订*

一．简介

Ceph OSD 主要用来存储数据，通常来说，一个 OSD 进程对应了单个的存储设备，比如传统的 HDD，或则 SSD，也有列外情况，比如在生产环境下，为了增加集群性能，一个 OSD 中使用 HDD 作为数据的存储，SSD 作为元数据的存储，这样的设备搭配使用也是可以的。OSD 的后端存储可以使用 bluestore,filestore 以及 kvstore 等；

为了保证 OSD 中各个副本对象的内容的一致性，OSD 提供了数据清洗功能，轻度清洗主要对比副本的大小和属性是否一致，深度清洗会读取数据并使用校验和来验证数据的完整性；

二．原理

2.1 架构

三．操作

3.1 新增 OSD

```
# Filestore
sudo mkdir /var/lib/ceph/osd/${clusterid}-${id} # 创建 osd 的目录
sudo mkfs -t {fstype} /dev/{drive} # 格式化磁盘
sudo mount -o user_xattr /dev/{drive} /var/lib/ceph/osd/${cluster}-${id} # 挂在到 osd 目录
ceph-osd -i {osd-num} --mkfs –mkkey # 初始化 osd 目录
ceph auth add osd.{osd-num} osd “allow *” mon “allow rwx” -i /var/lib/ceph/osd/${cluster}-${id}/keyring # 创建 cephx 认证
ceph osd crush add {id-or-name} {weight} [{bucket-type}={bucket-name}] # 添加 osd 到 crush map 中
```

3.2 更换 OSD

```
# 将 Filestore 更换为 Bluestore
ceph osd destroy {id} --yes-i-really-really-mean-it
ceph-volume lvm zap /dev/drive

ceph-volume lvm prepare --osd-id {id} --data /dev/{drive}
ceph-volume lvm activate {id} {fsid}
或
ceph-volume lvm create --osd-id {id} --data /dev/{drive}

sudo systemctl start ceph-osd@{osd_name}
```

3.3 删除 OSD

```
ceph osd out {osd-num}

# 如何在标记 OSD 为 osd 的过程中 pg 的状态一直卡在了 active+remapped 那么可用下面的方法;
ceph osd in {osd-num}
ceph osd crush reweight osd.{osd-num} 0

# 停止 OSD 进程
sudo systemcel stop ceph-osd@{osd-num}

# 移除集群中的 OSD 信息
ceph osd purge {id} --yes-i-really-really-mean-it
edit /etc/ceph.conf # 删除配置文件中的 osd 信息
```

3.4 OSD 参数

	一般
osd uuid	OSD 进程的唯一标识，是针对单个进程的，fsid 是针对真个集群的;
osd data	OSD 存储数据的目录，不推荐修改，默认为 /var/lib/ceph/osd/$cluster-$id;
osd max write size	OSD 最大写的大小，单位为 MB；默认为 90
osd max object size	集群中对象的最大值，单位为 MB，默认为 128
osd client message size cap	内存中的客户端数据消息最大值，默认 500MB
osd class dir	RADOS 插件的类路径，默认为$libdir/rados-classes
	文件系统(针对 Filestore)
osd mkfs options {fs-type}	当创建的 OSD 的文件系统为{fs-type}时使用的参数
osd mount options {fs-type}	当挂在文件类型为{fs-type}的 OSD 时的参数
	日志设置(JOURNAL)
osd journal	OSD 日志的位置，可以为一个文件夹或是一个设备，默认目录为/var/lib/ceph/osd/$cluster-$id/journal
osd journal size	OSD 日志的大小，当使用目录存储日志的时候，该值表示日志的最大值，默认为 5120MB，当使用块设备时，该值无效，会使用块设备的所有空间，该值至少是: osd journal size = {2 * (expected throughput * filestore max sync interval)}
osd max scrubs	
osd scrub begin hour	
osd scrub end hour	
osd scrub during recovery	
osd scrub thread timeout	
osd scrub finalize thread timeout	

osd scrub load threshold	
osd scrub min interval	
osd scrub max interval	
osd scrub chunk min	
osd scrub chunk max	
osd scrub sleep	
osd deep scrub interval	
osd scrub interval randomize ratio	
osd deep scrub stride	
OPRATIONS	
osd op queue	默认 prio, 可选值为 prio, wpq, mclock_opclass, mclock_client；
osd op queue cut off	
osd client op priority	
osd recovery op priority	
osd scrub priority	
osd snap trim priority	
osd op thread timeout	
osd op complaint time	
osd disk threads	
osd disk thread ioprio class	
osd disk thread ioprio priority	
osd op history size	
osd op history duration	
osd op log threshold	
QOS BASED ON MCLOCK	
osd push per object cost	
osd recovery max chunk	
osd op queue mclock client op res	
osd op queue mclock client op wgt	
osd op queue mclock osd subop res	
osd op queue mclock osd subop wgt	
osd op queue mclock osd subop lim	
osd op queue mclock snap res	
osd op queue mclock snap wgt	
osd op queue mclock snap lim	
osd op queue mclock recov res	
osd op queue mclock recov wgt	
osd op queue mclock recov lim	
osd op queue mclock scrub res	
osd op queue mclock scrub wgt	
osd op queue mclock scrub lim	
BACKFILLING	

osd max backfills	
osd backfill scan min	
osd backfill scan max	
osd backfill retry interval	
	OSD MAP
osd map dedup	
osd map cache size	
osd map cache bl size	
osd map cache bl inc size	
osd map message max	
	RECOVERY
osd recovery delay start	
osd recovery max active	
osd recovery max chunk	
osd recovery max single start	
osd recovery thread timeout	
osd recover clone overlap	
osd recovery sleep	
osd recovery sleep hdd	
osd recovery sleep ssd	
osd recovery sleep hybrid	
	TIERING
osd agent max ops	
osd agent max low ops	
	MISCELLANEOUS
osd snap trim thread timeout	
osd backlog thread timeout	
osd default notify timeout	
osd check for log corruption	
osd remove thread timeout	
osd command thread timeout	
osd command max records	
osd auto upgrade tmap	
osd tmapput sets users tmap	
osd fast fail on connection refused	

3.4 其他命令

```
ceph osd dump # 输出所有的 osd 信息
ceph osd tree # 输出 osd 的树形结构信息
ceph osd ls pool ls detail # 获取 pool 的详细信息
```

四．参考资料

【1】Adding/Removing OSDs