ceph集群搭建

一、集群规划

节点名称	节点IP	节点角色
node1	192.168.1.112	MON,OSD
node2	192.168.1.111	MON,OSD
node3	192.168.1.110	MON,OSD
node4	192.168.1.109	OSD
node5	192.168.1.108	OSD
node6	192.168.1.107	OSD
node7	192.168.1.106	OSD
node8	192.168.1.105	OSD
node9	192.168.1.104	OSD

二、准备工作

1. 时间同步

```
# node1节点
apt -y install ntp
# 清空配置文件
echo '' > /etc/ntp.conf
vim /etc/ntp.conf
# 文件中写入以下信息，同步阿里云时间服务器的时间。
server ntp.aliyun.com
# 设置时区
timedatectl set-timezone Asia/Shanghai
# 重启NTP服务
systemctl restart ntp
systemctl enable ntp

# 其他节点
# 如果其他节点可以连接外网，那么server ntp.aliyun.com，否则server 192.168.1.112
```

2. 安装cephadm

```
# 1.通过二进制安装
# 下载二进制命令程序包，速度20kb/s。嫌慢的同志使用下一种方法
curl --silent --remote-name --location
https://github.com/ceph/ceph/raw/quincy/src/cephadm/cephadm
```

```
# 命令添加可执行权限
chmod +x cephadm

# 验证cephadm命令是否可用
./cephadm --help

# 其实到这一步，cephadm就已经能够部署集群了，但是没有安装全部功能，也没有把命令安装成操作系统命令

# 添加ceph指定版本的系统包镜像源，这里我们安装quincy版。本地apt或yum库中会多出一些镜像地址。
./cephadm add-repo --release quincy

# 开始安装cephadm命令
./cephadm install

# 2.直接安装
apt install -y cephadm

# 验证cephadm已经成为系统命令
which cephadm
```

三、创建集群

如果以前创建过集群，可以通过 `cephadm rm-cluster --force  --fsid <fsid>` 卸载集群

1. 引导创建初始集群

```
cephadm bootstrap --mon-ip 192.168.1.112
```

```
Ceph Dashboard is now available at:

             URL: https://node1:8443/
            User: admin
        Password: 18ra8rfm2q

You can access the Ceph CLI with:

    sudo /usr/sbin/cephadm shell --fsid c586de66-5aa2-11ed-86ff-461c91943ba5 -c /etc/ceph/ceph.conf -k /etc/ceph/ceph.client.admin.keyring

Please consider enabling telemetry to help improve Ceph:

    ceph telemetry on

For more information see:

    https://docs.ceph.com/docs/master/mgr/telemetry/

Bootstrap complete
```

2. 安装ceph命令工具包

```
# 执行以下命令安装ceph-common包
cephadm add-repo --release quincy
cephadm install ceph-common

# 检验ceph命令在主机上安装成功
ceph -v

# 检验主机上的ceph命令能成功连接集群，获取集群状态
ceph -s
```

3. 安装OSD

前提要求

主机上的每块非系统硬盘都可作为一个OSD。但是能安装OSD的硬盘必须满足以下条件：

硬盘设备不能有分区
硬盘设备不能被其他LVM占用或声明
硬盘设备不能已挂载
硬盘不能包含文件系统
硬盘设备不能是包含Ceph Bluestore存储引擎的OSD
硬盘设备不能小于5GB

```
# 查看集群目前所有的OSD设备
ceph orch device ls
--------------------------------------------------------------------------------
----
Hostname  Path      Type  Serial                          Size   Health   Ident
Fault  Available
Hostname  Path            Type  Serial           Size   Health   Ident  Fault
Available
node1     /dev/mtdblock0  ssd                    4194k  Unknown  N/A    N/A    No

node1     /dev/sda        hdd   0123456789ABCDE  1000G  Unknown  N/A    N/A    No

node1     /dev/zram0      ssd                     519M  Unknown  N/A    N/A    No

node1     /dev/zram1      ssd                     519M  Unknown  N/A    N/A    No

node1     /dev/zram2      ssd                     519M  Unknown  N/A    N/A    No

node1     /dev/zram3      ssd                     519M  Unknown  N/A    N/A
No
# 注意Available是Yes的，表示这个设备满足条件，可以安装成为OSD，如果是No表示设备已经是OSD了，
或者设备无法安装OSD。


# 指定查看某一台机器的OSD
# 格式是：ceph orch device ls [--hostname=...] [--wide] [--refresh]
ceph orch device ls --hostname=10.0.1.15 --wide --refresh
```

1. 将主机添加到集群

```
ceph cephadm get-pub-key > ~/ceph.pub
```

```
# 重复下面二个操作，对每个机器
ssh-copy-id -f -i ~/ceph.pub root@node2
ceph orch host add node2 IP地址

ssh-copy-id -f -i ~/ceph.pub root@node3
ceph orch host add node3 IP地址
.....

# 再次查看集群目前所有设备
ceph orch host ls
--------------------------------------------------------------------------------
----
HOST   ADDR   LABELS  STATUS
node1  node1
node2  node2
node3  node3
node4  node4
node5  node5
node6  node6
node7  node7
node8  node8
node9  node9
```

2. 增加OSD

```
# 单独增加
ceph orch daemon add osd node1:/dev/sda
# 如果出现 Created no osd(s) on host node-01; already created? 错误，但是ceph -s 没有OSD，那么需要将下线得磁盘，初始化重新加入磁盘
# 需要zap该磁盘，使其可重新被使用。
ceph orch device zap node1 /dev/sda --force
# 出现异常
Traceback (most recent call last):
  File "<stdin>", line 6251, in <module>
  File "<stdin>", line 1358, in _infer_fsid
  File "<stdin>", line 1441, in _infer_image
  File "<stdin>", line 3713, in command_ceph_volume
  File "<stdin>", line 1120, in call_throws
# 需要wipefs -af /dev/sda
# 如果出现zap出现 probing initialization failed: Device or resource错误，需要手动进行
dd if=/dev/zero of=/dev/sda bs=512K count=1
# 然后重启 reboot

## 对所有可用设备添加到OSDS
#部署osd,磁盘不能小于5G
ceph orch apply osd --all-available-devices
```

如果上次安装ceph时，使用过/dev/sda，所有它会是NO状态，所有需要重新进行擦除化。

3. 添加MDS(可用不添加，因为不用cephFS)

```
ceph fs volume create sdnsdn ndoe1,node3,node4
```

4. **创建RGW(关键)**

几个重要得概念

> 1、zone:可用区，有一个或多个对象网关实例组成。zone不可以跨集群，配置zone不同于其他典型配置，因为不需要在ceph.conf中配置。
> 2、zonegroup:以前叫做"region"，有多个zone组成,一个zonegroup里面有一个master zone，在同一个zonegroup中的多个zone 可以同步元数据和数据，提供灾难恢复能力。
> 3、realm:代表一个唯一的命名空间，有一个或多个zonegroup组成。在同一个realm中的不同zonegroup只能同步元数据。在realm中有period的概念，表示zonegroup的配置状态，修改zonegroup，必须更新period。
> 要明确一点，一个zone属于一个zonegroup，两者都属于一个realm（域名）
> 创建RGW的命令中realm和zone是不可缺省的
> 原文链接：https://blog.csdn.net/HzauTriste/article/details/122480450

`orch apply rgw <realm_name> <zone_name> [<subcluster>] [<port:int>] [--ssl] [<placement>] [--dry-run] [plain|json|json-pretty|yaml]`

realm_name 和 zone_name 是必须得

```
#创建RGW
 ceph orch apply rgw rgw01 cn-shenzhen --placement="3 node1 node2 node3"  --
port=8000

------------------------------------------------------------------
 services:
    mon: 2 daemons, quorum node1,node3 (age 29m)
    mgr: node1.ovctpk(active, since 14h), standbys: node3.ubilql
    mds: sdnsdn:1 {0=node1.node3.aszewa=up:active}
    osd: 3 osds: 3 up (since 27m), 3 in (since 27m)
    rgw: 2 daemons active (foo.cn-shenzhen.node3.pnlzft, foo.cn-
shenzhen.node4.pgmikh)

 root@node1:~# ceph orch ls
NAME                       RUNNING  REFRESHED  AGE  PLACEMENT    IMAGE NAME
                         IMAGE ID
alertmanager                   1/1  14m ago    14h  count:1
quay.io/prometheus/alertmanager:v0.20.0   5eb21d2cb030
crash                          3/3  14m ago    14h  *
quay.io/ceph/ceph:v15                     mix
grafana                        1/1  14m ago    14h  count:1
quay.io/ceph/ceph-grafana:6.7.4           7e059a4bf8c7
mds.node1                      1/1  4m ago     24m  node3
quay.io/ceph/ceph:v15                     d7e2f380c45b
mgr                            2/2  14m ago    14h  count:2
quay.io/ceph/ceph:v15                     mix
mon                            2/2  14m ago    29m  node1;node3
quay.io/ceph/ceph:v15                     mix
node-exporter                  1/3  14m ago    14h  *
quay.io/prometheus/node-exporter:v0.18.1  8e76a0ec7d90
osd.None                       2/0  4m ago     -    <unmanaged>
quay.io/ceph/ceph:v15                     d7e2f380c45b
osd.all-available-devices      1/1  14m ago    14h  *
quay.io/ceph/ceph:v15                     260ffa0f98ff
```

```
prometheus                      1/1  14m ago    14h  count:1
quay.io/prometheus/prometheus:v2.18.1     a0ca3ee7950c
rgw.foo.cn-shenzhen             2/2  4m ago     5m   count:2
quay.io/ceph/ceph:v15                     d7e2f380c45b

# 集成到dashboard,也可以不用
radosgw-admin user create --uid=rgw --display-name=foo --system
```

5. s3接口配置

S3客户端配置

```
# 下载
pip install s3cmd
# 创建S3用户
root@node1:~# radosgw-admin user create --uid=s3user01 --display-name=s3user01
{
    "user_id": "s3user01",
    "display_name": "s3user01",
    "email": "",
    "suspended": 0,
    "max_buckets": 1000,
    "subusers": [],
    "keys": [
        {
            "user": "s3user01",
            "access_key": "R5YOAIXNAWW9MBH5PQKS",
            "secret_key": "ZSy8K9YBOLs24TqCWBAD8O4bQ4QO1WULZnv16j4O"
        }
    ],
    "swift_keys": [],
    "caps": [],
    "op_mask": "read, write, delete",
    "default_placement": "",
    "default_storage_class": "",
    "placement_tags": [],
    "bucket_quota": {
        "enabled": false,
        "check_on_raw": false,
        "max_size": -1,
        "max_size_kb": 0,
        "max_objects": -1
    },
    "user_quota": {
        "enabled": false,
        "check_on_raw": false,
        "max_size": -1,
        "max_size_kb": 0,
        "max_objects": -1
    },
    "temp_url_keys": [],
    "type": "rgw",
    "mfa_ids": []
}
```

```
# 配置s3客户端
s3cmd --configure
# 几个关键点
Access Key  与上面得access_key一致
Secret Key  与上面得secret_key一致
S3 Endpoint 主机ip+端口号 192.168.1.112:7480
DNS-style bucket+hostname:port template for accessing a bucket  %
(bucket)s.192.168.1.112
Use HTTPS protocol no
```