Ceph分部署存储

什么是分布式文件存储:

分布式文件存储是指文件系统管理的物理存储资源不一定直接链接在本地节点上而是通过计算机网络与节点相连.分布式文件系统的设计基于客户机和服务器模式

常用分布式文件系统:

Lustre Hadoop FastDFS ceph GlusterFS

什么是ceph

具有高扩展,高可用,高性能的特点 ,可以提供对象存储,块存储,文件系统存储

可以提供PB级别的存储空间(PB>TB>GB)

DAS存储 :直连存储

NAS存储 :

SAN存储

Ceph组件:

OSDs : 存储设备

Monitors:集群监控组件

RadosGateway (RGW) : 对象存储网关

MDSs : 存放文件系统的元数据(对象和块存储不需要)

Client:ceph客户端

确保时间同步,yum源同步,四台主机实现无密码链接(自己本身也要做),在物理机上为每一个机器添加3个磁盘(客户端除外)

Node1作为主机,能同事控制其他机器:

yum -y install ceph-deploy

mkdir ceph

cd ceph

部署ceph集群

]# ceph-deploy new node1 node2 node3 //给所有节点都为mon

]# for i in node1 node2 node3

> do

> ssh $i "yum -y install ceph-mon ceph-osd ceph-mds ceph-redosgw"

> done

///给所有节点安装ceph软件包

ceph-deploy mon create-initial //初始化所有节点的mon服务

准备日志磁盘分区

把/vdb分成两个区 作为vdc和vdd的缓存盘

/vdb1-----/vdc

/vdb2-----/vdd

ssd盘

]# for i in node1 node2 node3

do

ssh $i "parted /dev/vdb mklabel gpt"

ssh $i "parted /dev/vdb mkpart primary 1 50%"

ssh $i "parted /dev/vdb mkpart primary 50% 100%"

done

2）磁盘分区后的默认权限无法让ceph软件对其进行读写操作，需要修改权限。

node1、node2、node3都需要操作，这里以node1为例。

]# chown ceph.ceph /dev/vdb1

]# chown ceph.ceph /dev/vdb2 //由于只是暂时修改用户和组所以要写入配置文件

创建OSD存储空间

db1和vdb2这两个分区用来做存储服务器的journal缓存盘。

ceph-cluster]# for i in node1 node2 node3

do

ssh $i "parted /dev/vdb mklabel gpt"

ssh $i "parted /dev/vdb mkpart primary 1 50%"

ssh $i "parted /dev/vdb mkpart primary 50% 100%"

done

2）磁盘分区后的默认权限无法让ceph软件对其进行读写操作，需要修改权限。

node1、node2、node3都需要操作，这里以node1为例。

ceph-cluster]# chown ceph.ceph /dev/vdb1

ceph-cluster]# chown ceph.ceph /dev/vdb2

ceph-cluster]# vim /etc/udev/rules.d/70-vdb.rules

ENV{DEVNAME}=="/dev/vdb1",OWNER="ceph",GROUP="ceph"

ENV{DEVNAME}=="/dev/vdb2",OWNER="ceph",GROUP="ceph"

3）初始化清空磁盘数据（仅node1操作即可）。

]# ceph-deploy disk zap node1:vdc node1:vdd

]# ceph-deploy disk zap node2:vdc node2:vdd

]# ceph-deploy disk zap node3:vdc node3:vdd

4）创建OSD存储空间（仅node1操作即可）

重要：很多同学在这里会出错！将主机名、设备名称输入错误！！！

]# ceph-deploy osd create node1:vdc:/dev/vdb1 node1:vdd:/dev/vdb2

//创建osd存储设备，vdc为集群提供存储空间，vdb1提供JOURNAL缓存，

//一个存储设备对应一个缓存设备，缓存需要SSD，不需要很大

]# ceph-deploy osd create \

node2:vdc:/dev/vdb1 node2:vdd:/dev/vdb2

]# ceph-deploy osd create \

node3:vdc:/dev/vdb1 node3:vdd:/dev/vdb2

常见错误及解决方法（非必须操作）。

使用osd create创建OSD存储空间时，如提示下面的错误提示：

[ceph\_deploy][ERROR ] RuntimeError: bootstrap-osd keyring not found; run 'gatherkeys'

可以使用如下命令修复文件，重新配置ceph的密钥文件：

[root@node1 ceph-cluster]# ceph-deploy gatherkeys node1 node2 node3

步骤四：验证测试

1) 查看集群状态。

[root@node1 ~]# ceph -s

2）常见错误（非必须操作）。

如果查看状态包含如下信息：

health: HEALTH\_WARN

clock skew detected on node2, node3…

clock skew表示时间不同步，解决办法：请先将所有主机的时间都使用NTP时间同步！！！

Ceph要求所有主机时差不能超过0.05s，否则就会提示WARN，如果使用NTP还不能精确同步时间，可以手动修改所有主机的ceph.conf，在[MON]下面添加如下一行：

mon clock drift allowed = 1

如果状态还是失败，可以尝试执行如下命令，重启ceph服务：

[root@node1 ~]# systemctl restart ceph\\*.service ceph\\*.target

创建磁盘镜像:

]# rbd create damo-image(镜像名) --image-feature layering --size 50G(大小)

]# rbd list //查看镜像

]# rbd info damo-image //查看指定的镜像详细信息

]# rbd resize --size 55G damo-image --allow-shrink //扩容已有的镜像大小

在客户端:

客户端需要安装:ceph-common软件包

~]# yum -y install ceph-common

远端的拷贝到本地:

~]# scp 192.168.4.11:/etc/ceph/ceph.conf /etc/ceph/

~]# scp 192.168.4.11:/etc/ceph/ceph.client.admin.keyring /etc/ceph/

映像到本地磁盘

~]# rbd map damo-image

~]# lsblk //查看镜像在本地是否存在

~]# rbd showmapped

id pool image snap device

0 rbd damo-image - /dev/rbd0

客户端格式化,挂载分区:

~]# mkfs.xfs /dev/rbd0

~]# mount /dev/rbd0 /mnt

在echo输入内容,创建快照

~]# rbd snap ls damo-image //查看有没有镜像快照

~]# rbd snap create damo-image(快照名) --snap image-snap1 //创建镜像快照

~]# rbd snap ls damo-image //查看镜像快照

SNAPID NAME SIZE

4 image-snap1 56320 MB

使用快照恢复数据:

先删除客户端写入的文件

在卸载

~]# rbd snap rollback damo-image --snap image-snap1 //还原快照

在挂载,查看就有了之前的数据

~]# rbd snap purge rbd/damo-image //删除所有快照

分布式文件存储

新建一台虚拟机:

IP:192.168.4.14

配置yum源

安装ceph-mds 软件

]# ceph-deploy mds create node4 //部署新建主机的节点

]# ceph-deploy admin node4 //同步配置文件可key

创建存储池:

~]# ceph osd pool create cephfs\_data 120

创建两个存储池:

一个用于存储数据

一个用于存储元数据

~]# ceph osd pool create cephfs\_data 120

~]# ceph mds stat //查看mds状态

~]# ceph fs new myfs1 cephfs\_data (名称)cephfs\_data集群池名称 //查看集群池

~]# ceph fs ls //查看集群池名称

name: myfs1, metadata pool: cephfs\_data, data pools: [cephfs\_data ]

~]# ceph mds stat //查看状态

e5: 1/1/1 up {0=node4=up:active}

客户端挂载:

Linux内核需要支持ceph文件系统(不需要软件)

~]# mount -t ceph 192.168.4.11:6789:/ /mnt/cephfs -o name=admin,secret=AQDCUFFdda1eORAANHRVVkoqUb08KZ2IoCdvNw==

//在客户端挂载

文件系统类型为ceph

//192.168.4.11为mon节点的IP

//admin是用户名,secret是秘钥

秘钥可以再//etc/ceph/ceph.client.admin.keyring 中找到