ceph集群扩容和权重调整要点

集群扩容

该文档中只涉及集群扩容中的心得体会，不涉及任何具体操作过程，具体操作过程请查看该文档:实验环境手工配置crush map

物理拓扑

	rack-01	rack-02	rack-03
pool1 ssd	主机01 (osd11~13)	主机02 (osd14~16)	主机03 (osd17~19)
pool2 sata	主机04 (osd21~23)	主机05 (osd24~26)	主机06 (osd27~29)
	主机07 (osd31~33)	主机08 (osd34~36)	主机09 (osd37~39)

扩容作用

集群扩容通过向集群中增加osd来达到增大ceph集群存储容量的目的

扩容实现方式:

a. 通过向集群中添加osd来实现扩容

该种方式需要向ceph集群位于同一pool中的三个主机中添加相同数量的osd，如下图所示

	rack-01	rack-02	rack-03
pool1 ssd	主机01 (osd11~13) 新增osd41~43	主机02 (osd14~16) 新增osd44~46	主机03 (osd17~19) 新增osd47~49
pool2 sata	主机04 (osd21~23)	主机05 (osd24~26)	主机06 (osd27~29)
	主机07 (osd31~33) 新增osd51~53	主机08 (osd34~36) 新增osd54~56	主机09 (osd37~39) 新增osd57~59

b. 通过向集群中添加主机实现扩容

该种方式又有两种实现方法：添加单台主机或添加多台主机

添加单台主机的扩容方式在实际生产环境中不建议使用，实现方式如下图所示

	rack-01	rack-02	rack-03
pool1 ssd	主机01 (osd11~13)	主机02 (osd14~16)	主机03 (osd17~19)
	新增主机10 osd41~43	新增主机10 osd44~46	新增主机10 osd47~49
pool2 sata	主机04 (osd21~23)	主机05 (osd24~26)	主机06 (osd27~29)
	主机07 (osd31~33)	主机08 (osd34~36)	主机09 (osd37~39)

添加多台主机实现方式如下如所示

	rack-01	rack-02	rack-03
pool1 ssd	主机01 (osd11~13)	主机02 (osd14~16)	主机03 (osd17~19)
	新增主机11 osd41~43	新增主机12 osd44~46	新增主机13 osd47~49
pool2 sata	主机04 (osd21~23)	主机05 (osd24~26)	主机06 (osd27~29)
	主机07 (osd31~33)	主机08 (osd34~36)	主机09 (osd37~39)

c. 通过向集群中增加机架实现扩容

	rack-01	rack-02	rack-03
pool1 ssd	主机01 (osd11~13)	主机02 (osd14~16)	主机03 (osd17~19)
pool2 sata	主机04 (osd21~23)	主机05 (osd24~26)	主机06 (osd27~29)
	主机07 (osd31~33)	主机08 (osd34~36)	主机09 (osd37~39)
	rack-11	rack-12	rack-13
	新增主机11 osd111~113	新增主机12 osd114~116	新增主机13 osd117~119
	新增主机14 osd121~123	新增主机15 osd124~126	新增主机16 osd127~129
	新增主机17 osd131~133	新增主机18 osd134~136	新增主机19 osd137~139

扩容操作过程

构建物理拓扑

物理拓扑就是ceph集群中的osd节点在主机，机架的对应放置位置

物理拓扑中使用host domain来标识主机位置，rack domain来标识机架位置，root domain来标识机房名称

物理拓扑的构建顺序为：

a. 创建rack机架类型的bucket(指定机架实例存在则不用创建)

b. 将创建的rack移动至default的root bucket中

c. 创建host主机类型的bucket(指定主机实例存在则不用创建)

d. 将指定ods实例添加至指定host主机实例中，并指定该osd的weight权重值

osd的weight权重值是通过该osd对应磁盘容量(单位GB)/1024得到

构建逻辑拓扑

逻辑拓扑是ceph集群中数据实际的流转方向

逻辑拓扑中使用failure domain来标识失效域，replica domain来标识复制域，osd domain来表示osd域

逻辑拓扑的构建顺序为：

a. 创建replica-domain类型的bucket(指定replica domain复制域实例存在则不用创建)

b. 创建osd-domain类型的bucket(指定osd domainosd域实例存在则不用创建)

c. 将指定osd实例添加至指定osd domain实例中，并指定该osd的权重，权重的计算方式与物理拓扑中的权重计算方式相同

d. 将指定osd domain实例移动至指定replica domain实例中

e. 创建failure-domain类型的bucket(指定failure domain失效域实例存在则不用添加)

逻辑拓扑构建过程中的要点

a. 在已有数据的replica domain中添加osd domain之后会产生数据迁移，可能会对集群性能或集群健康状态产生影响

因此在进行集群扩容时一般会创建新的replica domain，再向其中加入osd domain，用于避免发生数据迁移

b. 构建逻辑拓扑时的顺序是从低到高：先创建osd domain，将osd进程实例添加至osd domain；再创建replica domain，将osd domain实例添加至replica domain；最后再将replica domain添加至新建或已有的failure domain中，避免过早发生数据迁移

编辑并应用rule规则

编辑并应用rule规则的过程为：

a. 导出当前已有的rule规则，并保存为文本文件格式

b. 对文本文件格式的rule规则进行编辑，添加指定类型的rule规则实例

c. 将文本文件格式的rule规则保存为二进制格式文件导入即可

扩容过程中的相关命名规范

当前生产环境中会同时使用ssd和sata两种类型的磁盘，因为同一主机中会存在两种类型的磁盘，
在物理拓扑中不能对两种类型磁盘对应的bucket进行分别命名，在逻辑拓扑中可以对两种类型磁盘的bucket(EX：osd-domain，replica-domain，failure-domain)以及在对应failure-domain上创建的pool命名时要便于区分：

ssd类型磁盘：osd-domain命名为osd-group-a-xx，replica-domain命名为replica-a-xx，failure-domain命名为apple，
在ssd上创建的pool命名为openstack-00

sata类型磁盘: osd-domain命名为osd-group-b-xx, replica-domain命名为replica-b-xx, failure-domain命名为sata01,
在sata上创建的pool命名为sata-00

---

osd权重调整

权重作用

weight权重是ceph集群用于衡量每个osd的存储能力的度量值, osd的权重值越大, 表明该osd的存储能力越强,
ceph集群就会将更多的数据分配至该osd上进行存储

权重实际作用过程

osd本身的weight权重变化 ==> PG向该osd映射的变化 ==> osd之间发生对应的数据迁移

权重调整规则

a. 调整权重时, 需要保证一个osd-domain内的权重重量保持不变, EX: 降低了osd.1实例权重值0.2, 就要升高该osd域内中的osd.3实例权重值0.2

b. 在一个osd-domain内调整权重时, 需要选择权重值一高一低的两个osd实例进行调整, 而不能选择两个高的或两个低的osd实例进行调整

c. 对一高一低两个实例进行调整时, 需要先将高权重的osd实例调低, 再将权重低的osd实例调高

d. 调整权重值的单位为0.01: 将高权重的osd实例权重下降0.01, 再将低权重的osd实例升高0.01

e. 在对2个osd实例调整过程中, 需要实时查看ceph -s和df -h命令的输出结果来查看调整权重过程中的数据迁移状态,
但集群状态稳定在ok状态时才能进行下一次权重调整

f. 当某个osd的磁盘使用率达到94%的极端情况下, 可以立刻对该osd的权重下降0.01, 再找出同一osd-domain下的权重较小的osd进程升高0.01