实用 | HDFS维护状态

2017-07-07 Cloudera中国

引言

更新操作系统以及应用安全补丁或修补程序等系统维护操作是任何数据中心的常规操作。需要开展这种维护操作的DataNodes可能在任何地方脱机工作几分钟到几个小时。按照设计，Apache Hadoop HDFS可以处理DataNodes停止。但是，同时在多个DataNode上进行任何非协同维护操作可能会导致临时性的数据可用性问题。
目前为了执行计划好的维护任务，HDFS支持以下功能：
1. 滚动升级。
2. 退役。
3. 使用维护状态（自CDH 5.11版本开始提供这一功能）

滚动升级过程有助于升级集群软件，而无需使集群脱机。当来自相同或不同机架的多个DataNode进行停机升级时，最好选择使得块和文件的副本可用性不受影响的DataNode。但是，并没有什么简单的办法可以实现这一点。

HDFS支持“退役”功能，以减少滚动升级时遭遇到的数据可用性问题。当DataNodes请求退役时，NameNode将其转换为“退役正在进行中（decommission in-progress）”的状态，其中所有块都将被复制到其他实时DataNodes上，以满足使用dfs.replication属性或文件特定复制因子的全局块复制需求。当所有这些副本都被充分复制到其他DataNodes时，NameNode将处于“退役正在进行中”状态下的DataNodes转换为最终“已经退役”状态。

当在多个DataNode上同时运行退役操作时，因为必须为所有DataNodes上的所有块完成足够的复制，退役操作可能会非常耗费时间，比如在整个机架上执行维护操作。退役操作还会增加集群中的网络使用率，并且可能影响性能SLA。

维护状态

上游Jira：HDFS-7877

设计文档：Doc

HDFS维护状态这一新功能旨在克服滚动升级和退役功能存在的缺点，并使计划性维护活动更加无缝地进行。维护状态功能仅适用于HDFS DataNode角色。

维护状态功能通过允许块复制因子在短暂时间段内小于配置的级别以避免不必要的块复制。也就是说，它不会马上计划复制要求开展维护活动的DataNodes上的块。即使这些DataNodes停机进行维护，集群也将继续为副本使用最小的复制限值（该限值可予以配置，稍后详细讨论）以维持运行，而不是满足全局块复制因子或文件特定复制因子设置所需级别。

短期性维护活动（例如批量性滚动升级或修理机架交换机）是此功能的最佳适用用例。

在更深入地了解这个功能之前，让我们先看看这个配置。

配置

dfs.namenode.maintenance.replication.min:

此NameNode配置定义了所有正在维护的DataNodes块所需的最小数量实时副本。对于处于维护状态的DataNodes，在其他正常运行的DataNodes中存在的块满足该参数时，则不会触发复制。否则，将触发复制以满足此标准。此配置的允许范围为[0至dfs.replication]。

当维护状态最小复制因子为1时，表示属于DataNodes的块在维护持续时间内，集群可仅使用1个活动副本。如果在即将进入维护状态的DataNode上存在块的唯一副本时，则该功能首先确保副本被充分复制到另一个DataNode，以保证更好的数据可用性。

当维护状态最小复制因子等于全局复制因子时，即使DataNodes处于维护状态，集群也至少需要dfs.replication副本块，在触发块复制时这会导致维护状态功能的表现与退役功能类似。

当维护状态最小复制因子为0时，其表示：对于正在进行维护的DataNodes上的块，NameNode不会对任何最小复制系数进行验证。系统管理员可能希望将此值设置为维护最小复制因子，以通过避免副本扫描来加快维护操作，但该设置会增加数据无法访问的风险。

dfs.namenode.hosts.provider.classname:

该NameNode配置指定了提供主机文件访问权限的类。在维护状态功能诞生之前，节点重新加入需要使用dfs.hosts配置属性指定的文件，并且退役节点需要使用dfs.hosts.exclude配置属性指定的排除文件。这些文件具有使用逗号分隔的节点名称或IP地址作为值的简单格式。默认使用org.apache.hadoop.hdfs.server.blockmanagement.HostFileManager加载dfs.hosts和dfs.hosts.exclude指定的文件。

由于向后兼容性原因，这些较旧的文件格式不可扩展。新的维护状态功能使用JSON格式的新型组合主机文件，可用于包含和排除（退役或维护状态）DataNodes。仅使用此较新的组合主机文件方可支持维护状态功能。提供的程序类别名称必须为org.apache.hadoop.hdfs.server.blockmanagement.CombinedHostFileManager以便加载在dfs.hosts中定义的JSON文件。

以下是主机组合文件的格式：

“adminState”表示节点的“hostName”希望转换到的状态。其允许的状态包括：

```
NORMAL
DECOMMISSIONED
IN_MAINTENANCE
```

“maintenanceExpireTimeInMS”表示的是维护到期时间，当达到该到期时间后，相应的DataNodes内部状态将在NameNode中恢复为NORMAL，因为维护最小值将不再有效，这可能会导致复制。该维护到期时间配置是DataNode特定的，并且可以从一个DataNode转换到另一个DataNode。与需要快速重新启动修复程序的DataNodes相比，利用完整操作系统更新的DataNodes可能需要更长的到期时间。维护到期时间以从纪元至今所经过的毫秒为单位进行表示。

操作

根据使用案例、可靠性和性能要求，确定维护状态最小复制的最优值。如下例所示，需要配置dfs.namenode.maintenance.replication.min：

仅新型的组合主机文件格式支持维护状态功能。如下图所示，对主机文件提供程序进行配置：

使用dfs.hosts属性指定组合主机文件的位置。

所有上述值的任何更改都需要NameNode重新启动才能生效。Cloudera公司建议您在打开维护窗口之前确定上述参数。对DataNodes更新组合主机文件/etc/hadoop/conf/maintenance需要进行如下所述维护操作。

对于datanode-100、datanode-101和datanode-102的一个小时维护窗口而言，其维护到期时间的计算如下所述：

```
namenode-host# echo $((`date +%s` * 1000 + 60 * 60 * 1000))
1492543534000
namenode-host# cat /etc/hadoop/conf/maintenance
{
   "hostName": "datanode-100",
   "port": 50010,
   "adminState": "IN_MAINTENANCE",
   "maintenanceExpireTimeInMS": 1492543534000
}
{
   "hostName": "datanode-101",
   "port": 50010,
   "adminState": "IN_MAINTENANCE",
   "maintenanceExpireTimeInMS": 1492543534000
}
{
   "hostName": "datanode-102",
   "port": 50010,
   "adminState": "IN_MAINTENANCE",
   "maintenanceExpireTimeInMS": 1492543534000
}
```

在更新完组合主机文件后，运行以下命令触发NameNode刷新节点列表，并针对指定的DataNodes启动维护状态转换。

namenode-host$ hdfs dfsadmin -refreshNodes

在摘要页面的NameNode WebUI中可以监控DataNodes状态转换到维护状态的过程。一旦DataNodes转换到IN_MAINTENANCE状态，就可以将其从集群中安全地删除，以进行维护操作。

Configured Capacity:	1.36 TB
DFS Used:	147.65 MB (0.01%)
Non DFS Used:	473.33 GB
DFS Remaining:	920.9 GB (65.97%)
Block Pool Used:	147.65 MB (0.01%)
DataNodes usages% (Min/Median/Max/stdDev):	0.01% / 0.01% / 0.01% / 0.00%
Live Nodes	4 (Decommissioned: 0, In Maintenance: 1)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	Fri Dec 16 18:35:47 -0800 2016
Last Checkpoint Time	Fri Dec 16 18:35:46 -0800 2016

要使完成维护操作的DataNodes重新加入到该集群中，将使用相同的组合主机文件重复refreshNodes管理操作，但是将DataNodes的adminState将被更新为NORMAL，如下所示：

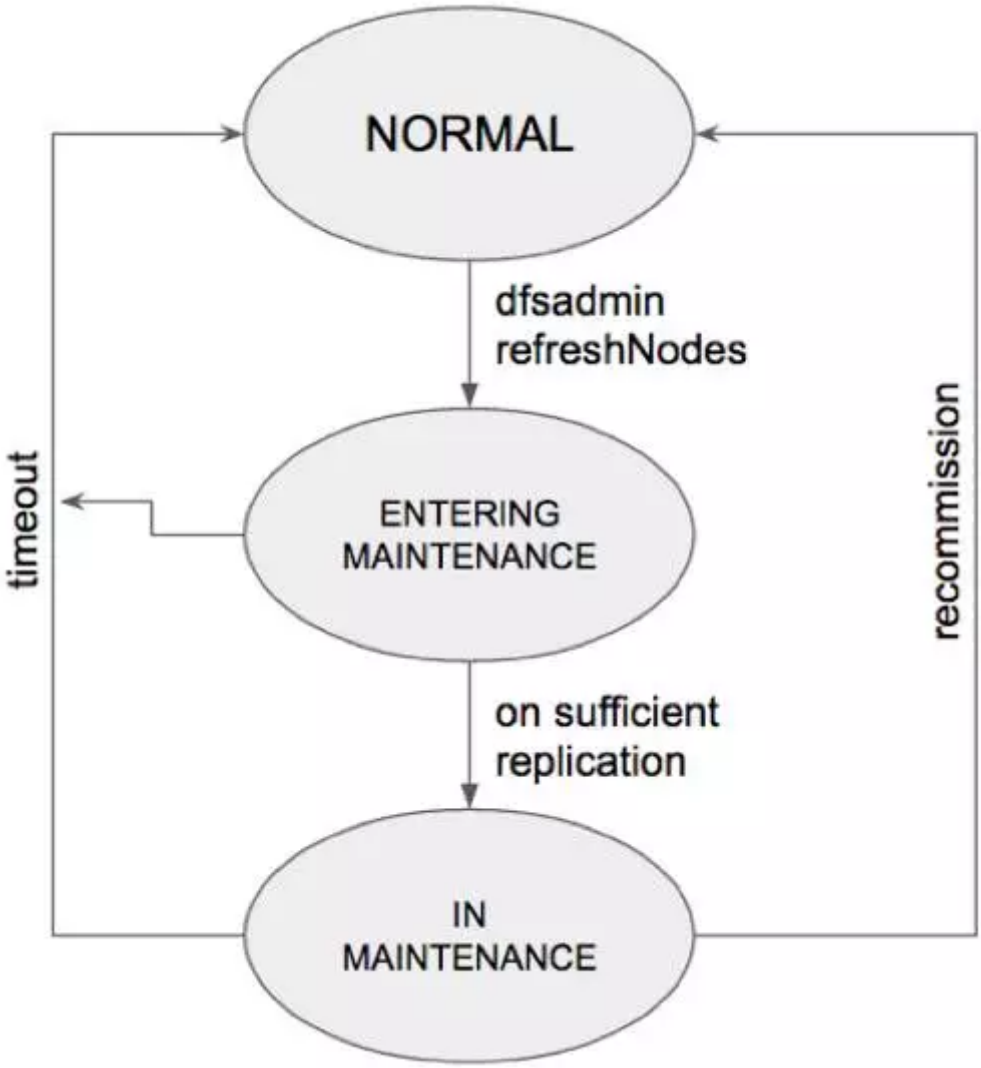
```
{
   "hostName": "datanode-100",
   "port": 50010,
   "adminState": "NORMAL"
}
...
```

内部构件

NameNode确保位于正在请求维护操作的DataNodes上的块按照维护状态最小复制配置dfs.namenode.maintenance.replication.min进行充分复制。这个短暂的等待期被称为ENTERING_MAINTENANCE状态。当块被充分复制时，DataNodes将转移到最终安全状态，在该状态下，已做好从集群中删除的准备。此最终安全状态称为IN_MAINTENANCE状态。

如果维护状态最小复制因子被设置为0，那么DataNode将直接转换为IN_MAINTENANCE。从状态转换的角度来看，ENTERING_MAINTENANCE => IN_MAINTENANCE与DECOMMISSION_INPROGRESS => DECOMMISSIONED的操作类似。对于维护状态而言，为了验证是否有足够的复制，将使用dfs.namenode.replication.maintenance.min属性；而对于退役而言，则将使用文件复制因子。

下图所示为非常简易的状态转换示意图。更多关于状态转换和事件的详图，请参阅设计文档。



DataNodes四个状态中的任何一个状态ENTERING_MAINTENANCE、IN_MAINTENANCE、DECOMMISSION_INPROGRESS或DECOMMISSIONED都被称为暂停服务节点。

对于写入请求而言，暂停服务的节点将完全被屏蔽。默认块放置策略将所有暂停服务的DataNodes作为副本候选。对于读取请求而言，暂停服务的节点将部分受到防护。对于某一给定的块读取请求，BlockManager返回一个已排除IN_MAINTENANCE节点的LocatedBlock，但仍可包含活动的ENTERING_MAINTENANCE节点。ENTERING_MAINTENANCE DataNodes将继续为持续的读取和写入请求提供服务。

NameNode BlockMap继续保留来自ENTERING_MAINTENANCE和IN_MAINTENANCE DataNodes的所有块，因此从块复制的角度来看，它们被认为是有效的副本。NameNode在维护状态到期之前不会使这些DataNodes中的任何块无效，无论这些节点是否死亡或活着。

磁盘均衡器和移动器（Disk Balancer and Mover）可以避免出现任何块移动造成的暂停DataNodes服务。

缺点

当维护状态最小复制因子为1时，如果托管唯一副本的DataNode暂时停止，则可能会导致数据可用性问题。另外，为副本所有读取请求提供服务的单个DataNode可能会导致出现性能问题。因此，当同时对许多个DataNodes执行维护操作时，Cloudera公司建议您将维护最小复制因子设置为2，以防止出现上述问题。

总结

维护状态功能HDFS-7877在对DataNodes进行任何快速维护操作时，克服了滚动升级和退役功能的复杂性和性能损失等缺陷。CDH5.11中采用的HDFS版本完全支持此维护状态功能。

致谢

HDFS-7877是由来自Twitter公司的Ming Ma和来自Cloudera公司的Lei（Eddy）Xu、Manoj Govindasamy共同合作开发完成。

阅读原文