超人学院技术篇：hadoop hdfs 滚动升级

2016-05-31 BJ-CRXY

HDFS 滚动升级允许对单独的HDFS进程升级。例如，datanodes可以被单独的升级而不依赖Namenodes。一个namenode可以被升级而不依赖其他的namenodes。Namenodes可以被升级而不依赖datanodes和journal 节点

在 hadoop v2，HDFS支持高可用的namenode服务和写兼用。这些功能可以让HDFS再升级的时候不需要停机、为了使HDFS集群无停机时间，集群必须设置为HA

如果在任何新版本中启用了任何新特性，升级后可能无法在旧版本上使用。在这种情况下(t.dbdao.com)，升级应遵循下列步骤：

1.禁用新功能

2.升级集群

3.启用新功能

注意：滚动升级只能从hadoop-2.4.0之后

2.1 无停机时间升级

在一个HA集群，有2个或者更多的NameNodes(NNs),一些journalNodes（JNS）和一些ZookeeperNodes（ZKNS）。在大部分情况下，当升级HDFS时，Jns是相对稳定的，不需要进行升级。在这里描述的滚动升级，只针对NNs和DNs，不考虑JNs和ZKNs。升级Jns和ZKNs可能导致集群停止。

升级非联合集群：

假设这里有2个namenodes NN1和NN2，NN1是actve，NN2是standby状态。下面是升级一个HA集群的步骤(t.dbdao.com)：

1.准备滚动升级

（1）.运行hdfs dfsadmin -rollingUpgrade prepare 来为滚动创建一个fsimage

（2）.运行hdfs dfsadmin -rollingUpgrade query 来检查回滚image的状态。等待和重新运这个命令，直到出现“Proceed with rolling upgrade”

2.升级active和standby NNs

(1).关闭和升级N2

(2).使用-rollingUpgrade started选项将NN2作为standby启动

(3).将NN1故障转移到NN2，这样NN2变成了active，NN1成为standby

(4).关闭和升级NN1

(5).使用-rollingUpgrade started选项将NN1作为standby启动

3.升级DNs

1.选择一小部分DataNodes子集(例如在一个机架中的datanodes)(t.dbdao.com)
2.运行hdfs dfsadmin -shutdownDatanode <DATANODE_HOST:IPC_PORT> upgrade命令来关闭这些DataNode
3.运行hdfs dfsadmin -getDatanodeInfo <DATANODE_HOST:IPC_PORT>来检查，并等到DataNode停止
4.升级和重启dataNode
5.在平行的子集上的dataNode执行上面所有的步骤
6.重复上面步骤直到集群中的所有DataNode都被升级。
7.完成滚动升级
运行hdfs dfsadmin -rollingUpgrade finalize 来完成滚动升级
升级联合集群：
在一个联合的机器，这里有多个Namespace，每个Namespace都有一对active和standby NNS。联合集群的升级和非联合集群的升级很相似，除了步骤1和步骤4在每个namespace上执行之外，在每对active和standby NNs上执行步骤2，例如(t.dbdao.com)：
1.为每个Namespace准备滚动升级
2.为每个Namespace升级每对active和standby NN
3.升级DNs
4.为每个Namespace 完成滚动升级
2.2 有停机时间的升级
对于非HA 集群，不可能不停机来进行升级，因为需要重启namenodes。但是datanodes仍然可以以滚动的方式升级(t.dbdao.com)。
升级非HA集群：
在一个非HA集群，这里有一个NN，SNN和许多DNs。升级非HA集群和升级HA集群很相似，除了第二步“升级active和stadby NNs”变为：
升级 NN和SNN
1.停止SNN
2.停止和升级NN
3.使用-rollingUpgrade started选项启动NN
4.升级和重启SNN
3.降级和回滚
当升级的版本不合适，或者由于一些不太可能的情况，(t.dbdao.com)升级失败（由于新版本的bug），管理员可以考虑降级HDFS到升级之前的版本，或者回滚HDFS到升级之前的版本和升级之前的状态。
注意降级可以以滚动的方式进行，但是回滚不行。回滚需要集群停机。
注意 降级和回滚只能在滚动升级被启动和升级被终止之前。一个升级可以被 完成，降级，回滚来终止。因此，在执行完成或降级之后执行回滚，或者在完成之后执行降级，可能不可能。
降级将软件重建到升级之前的版本，并保持用户数据。假设时间T是滚动升级开始的时间，并且升级被降级终止。那么，在T之前或之后创建的文件仍然在HDFS中可用。在T之前或之后的文件仍然会被HDFS删除。

一个新版本可以被降级到之前的版本，只有在namenode布局版本和DataNode布局均在2个版本之间无变化

4.1 无停机时间的降级

在一个HA 集群，当从一个就版本滚动升级到进软件版本的时候，可能需要降级，在滚动的方式中，被升级的机器被回退到就软件版本。和之前一样，假设NN1和NN2分别为active和standby状态。滚动降级的步骤如下：

1.降级DNs

选择一小部分DataNodes子集(例如在一个机架中的datanodes)

1）.运行hdfs dfsadmin -shutdownDatanode <DATANODE_HOST:IPC_PORT> upgrade来停止其中选择的DataNode

2）.运行hdfs dfsadmin -getDatanodeInfo <DATANODE_HOST:IPC_PORT>来检查，并等待datanode关闭

3）.降级和重启DataNode

4）.重复上面的操作，直到平行子集中的多(t.dbdao.com)有DataNode全都执行了上述步骤

2.重复上述步骤直到所有的DataNode都被降级

降级active和standby NNs

1.停止并且降级NN2

2.以普通standby方式启动NN2（注意，这里使用-rollingUpgrade downgrade不正确。）

3.从NN1故障转移到NN2这样NN2成为了active 并且NN1成为了standby。

4.关闭和升级NN1

5.以普通standby方式启动NN1（注意，这里使用-rollingUpgrade downgrade不正确。）

3.完成回滚降级

运行hdfs dfsadmin -rollingUpgrade finalize来完成回滚降级(t.dbdao.com)

请注意，datanode 必须在namenode之前进行降级，因为协议的变更可能是向后兼容，当时不会是向前兼容，例如，就datanode可以和新Namenode通信，但是反过来就不行

4.2 停机降级

管理可以选择首先关闭集群和降级。一下的步骤：

1.关闭所有NNs和DNs

2.在所有机器上重建之前的版本

3.使用-rollingUpgrade downgrade选项启动NNs

4.以普通方式启动DNs(t.dbdao.com)

回滚不仅重建软件到升级之前的版本，还恢复用户数据到升级之前的状态。假设滚动升级开始的时间是T，并且升级被回滚终止。在T之前创建的文件在HDFS仍然可用，但是在T之后创建的文件不可用。在T之前删除的文件，在HDFS仍然删除，但是在T之后删除的文件会被重建(t.dbdao.com)。

从一个新版本回滚到之前的版本总是被支持的。但是这个不能以滚动的方式进行。其需要集群停机。假设NN1和NN2分别为active和standby状态。回滚的步骤如下：

回滚HDFS
1.停止所有的NNs和DNs
2.在所有机器上重建升级前的版本
3.以acive 使用 -rollingUpgrade rollback选项启动NN1
4.在NN2上运行-bootstrapStandby’ ，以普通方式以standby启动
5.使用-rollback选项启动DNs
6. 滚动升级的命令和启动选项
6.1 DFSAdmin 命令
Dfsadmin –rollingUpgrafe
hdfsdfsadmin -rollingUpgrade <query|prepare|finalize>
选项：执行下列升级动作:

query	查询当前滚动升级的状态
prepare	准备一个新的滚动升级
finalize	完成当前的滚动升级

dfsadmin –getDatanodeInfo
hdfsdfsadmin -getDatanodeInfo <DATANODE_HOST:IPC_PORT>
从给定的datanode获得选项。这个命令用来检测datanode是否存活，可unix命令ping相似(t.dbdao.com)。
dfsadmin –shutdownDatanode
hdfsdfsadmin -shutdownDatanode <DATANODE_HOST:IPC_PORT> [upgrade]
注意这个命令不会等待datanode关闭完成。dfsadmin –getDatanodeInfo命令可以用来检测datanode是否关闭完成对给定的datanode提交关闭请求。如果选项参数指定了upgrade，客户端访问datanode将被建议等待期重启，并且快速启动默认将被启用。当在一定时间没有重启时，客户端会由于超时而忽略这个datanode。在这种情况下，快速启动模式也将被禁用。
6.2 NameNode 启动选项
namenode -rollingUpgrade
hdfsnamenode -rollingUpgrade <downgrade|rollback|started>
选项：当在进行一个滚动升级时，-rollingUpgrade namenode启动选项可以指定多种滚动升级选项

downgrade	namenode到升级之前的版本并且保持用户数据
rollback	重建namenode到升级之前的版本但是也恢复用户数据到升级状态之前
started	指定一个滚动升级已经开始，这样namenode应该允许image目录在启动的时候有不同布局版本。

阅读原文