Hadoop2.2.0中HDFS的高可用性实现原理

在Hadoop2.0.0之前，NameNode(NN)在HDFS集群中存在单点故障（single point of failure），每一个集群中存在一个NameNode，如果NN所在的机器出现了故障，那么将导致整个集群无法利用，直到NN重启或者在另一台主机上启动NN守护线程。

主要在两方面影响了HDFS的可用性：

（1）、在不可预测的情况下，如果NN所在的机器崩溃了，整个集群将无法利用，直到NN被重新启动；

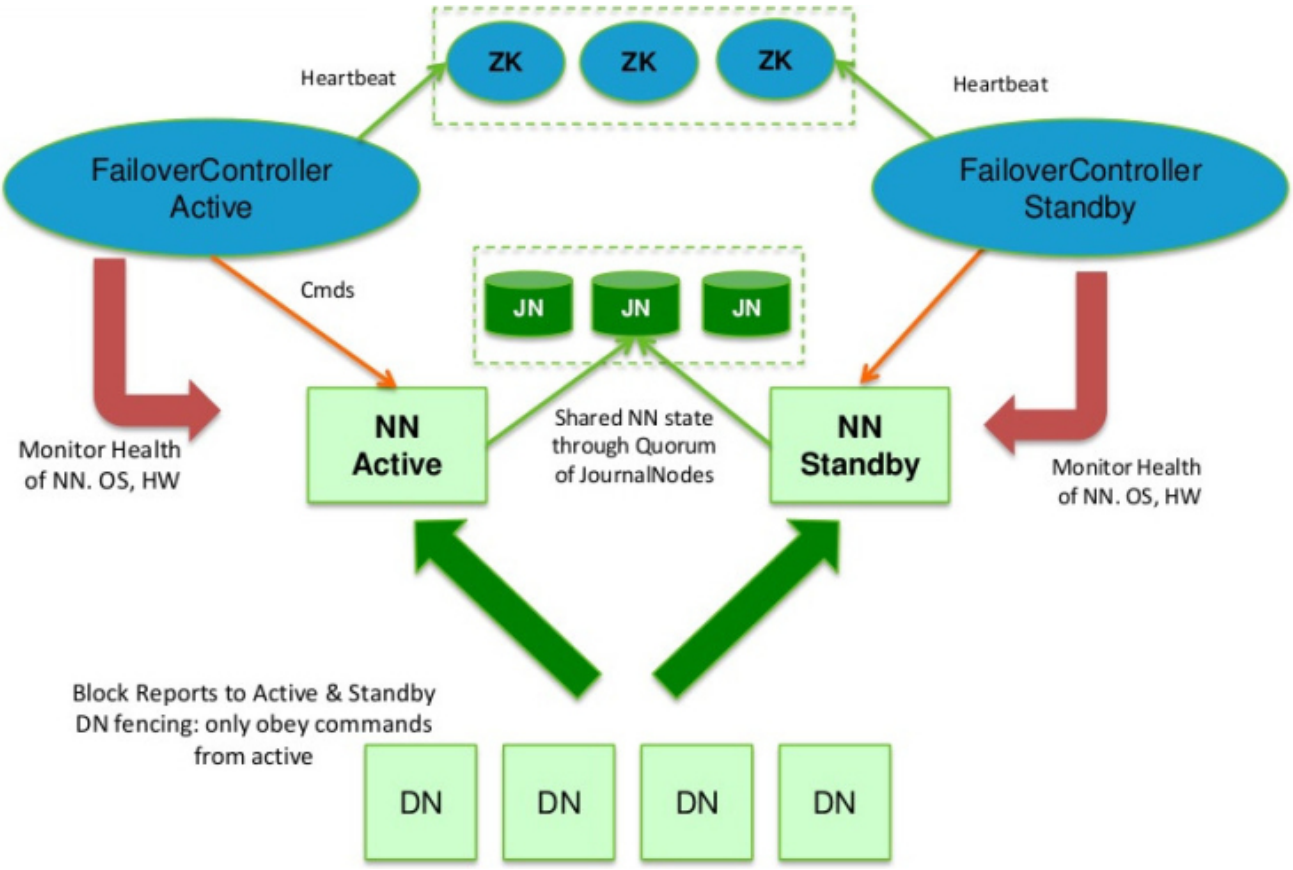
（2）、在可预知的情况下，比如NN所在的机器硬件或者软件需要升级，将导致集群宕机。

HDFS的高可用性将通过在同一个集群中运行两个NN（active NN & standby NN）来解决上面两个问题，这种方案允许在机器破溃或者机器维护快速地启用一个新的NN来恢复故障。

在典型的HA集群中，通常有两台不同的机器充当NN。在任何时间，只有一台机器处于Active状态；另一台机器是处于Standby状态。Active NN负责集群中所有客户端的操作；而Standby NN主要用于备用，它主要维持足够的状态，如果必要，可以提供快速的故障恢复。

为了让Standby NN的状态和Active NN保持同步，即元数据保持一致，它们都将会和JournalNodes守护进程通信。当Active NN执行任何有关命名空间的修改，它需要持久化到一半以上的JournalNodes上(通过edits log持久化存储)，而Standby NN负责观察edits log的变化，它能够读取从JNs中读取edits信息，并更新其内部的命名空间。一旦Active NN出现故障，Standby NN将会保证从JNs中读出了全部的Edits，然后切换成Active状态。Standby NN读取全部的edits可确保发生故障转移之前，是和Active NN拥有完全同步的命名空间状态。

为了提供快速的故障恢复，Standby NN也需要保存集群中各个文件块的存储位置。为了实现这个，集群中所有的Database将配置好Active NN和Standby NN的位置，并向它们发送块文件所在的位置及心跳，如下图所示：



Hadoop2.2.0中HDFS的高可用性实现原理

在任何时候，集群中只有一个NN处于Active 状态是极其重要的。否则，在两个Active NN的状态下NameSpace状态将会出现分歧，这将会导致数据的丢失及其它不正确的结果。为了保证这种情况不会发生，在任何时间，JNs只允许一个NN充当writer。在故障恢复期间，将要变成Active状态的NN将取得writer的角色，并阻止另外一个NN继续处于Active状态。

为了部署HA集群，你需要准备以下事项：

（1）、NameNode machines：运行Active NN和Standby NN的机器需要相同的硬件配置；

（2）、JournalNode machines：也就是运行JN的机器。JN守护进程相对来说比较轻量，所以这些守护进程可以可其他守护线程（比如NN，YARN ResourceManager）运行在同一台机器上。在一个集群中，最少要运行3个JN守护进程，这将使得系统有一定的容错能力。当然，你也可以运行3个以上的JN，但是为了增加系统的容错能力，你应该运行奇数个JN（3、5、7等），当运行N个JN，系统将最多容忍(N-1)/2个JN崩溃。

在HA集群中，Standby NN也执行namespace状态的checkpoints，所以不必要运行Secondary NN、CheckpointNode和BackupNode；事实上，运行这些守护进程是错误的。