20-namenode高可用2（ha）（热备）-原理

----------成都尚学堂-mr-zeng----------

hadoop2对namenode单点问题的解决-方案

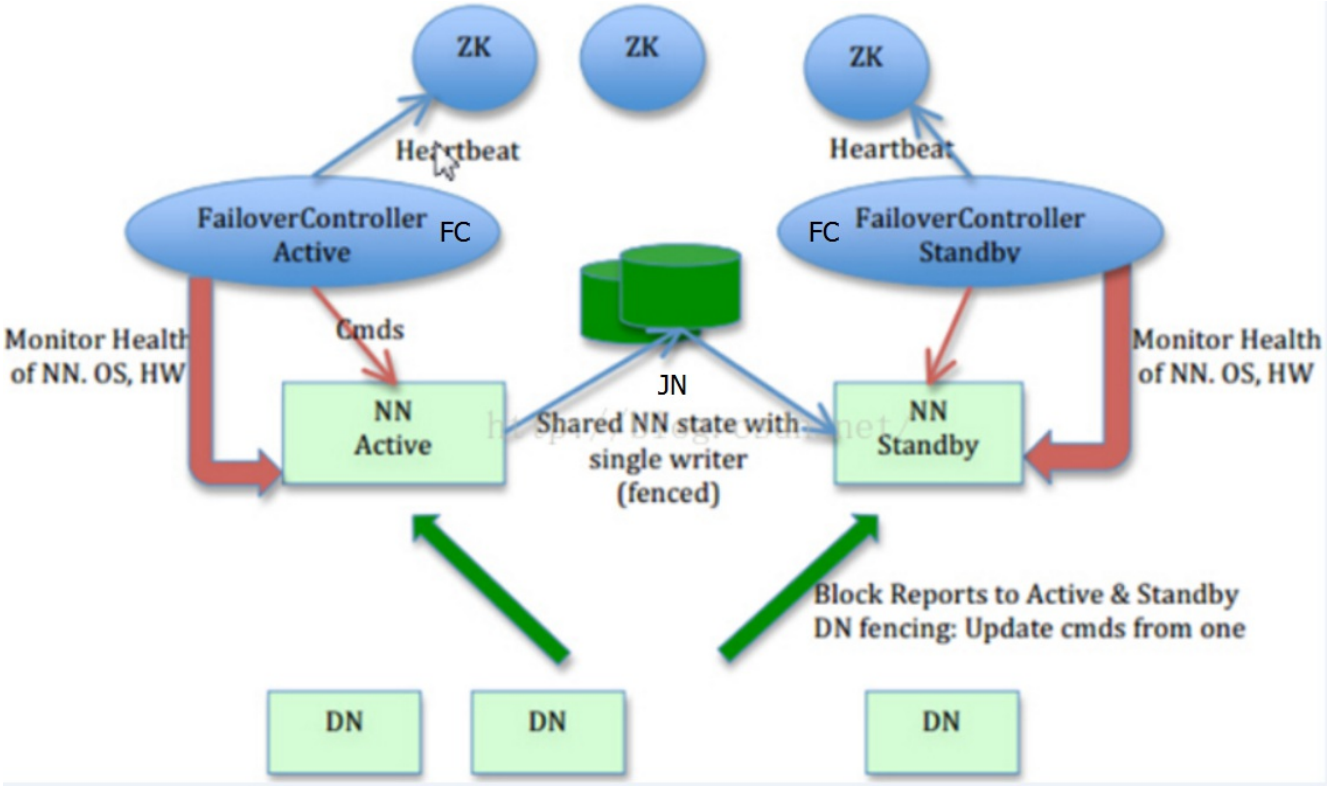
为了解决hadoop1中的单点问题，在hadoop2中新的NameNode不再是只有一个，可以有多个（目前只支持2个）。每一个都有相同的职能。一个是active(激活)状态的，一个是standby（备用）状态的。当集群运行时，只有active状态的NameNode是正常工作的，standby状态的NameNode是处于待命状态的，**时刻同步**active状态NameNode的数据。一旦active状态的NameNode不能工作，**通过（手工切换-命令）或者（自动切换-zookeeper事件-监听）**，standby状态的NameNode就可以转变为active状态的，就可以继续工作了。这就是高可靠。

Namenode元数据的存储

使用JournalNode实现NameNode(Active和Standby)数据的共享

Hadoop2.0中，2个NameNode的数据其实是实时共享的。新HDFS采用了一种共享机制，Quorum Journal Node（JournalNode）集群或者Nnetwork File System（NFS）进行共享。NFS是操作系统层面的，JournalNode是hadoop层面的，我们这里使用JournalNode集群进行数据共享（这也是主流的做法）。

hadoop2里Namenode的HA架构



ZK：zookeeper服务器，提供对在线状态namenode信息进行存储-和在有namenode挂机时提醒FailoverController进行故障转移。

FC：FailoverController，检测namenode健康，并在故障做转移的程序-与对应的namenode在同一个节点。

NN：namenode，访问hdfs的入口，提供对datanode的管理。--》注意元数据不在NN存储。

JN：JournalNode，存储Namenode要使用的元数据的服务器。

DN：hdfs的存储数据块的服务器。

NameNode之间的故障切换

对于HA集群而言，确保同一时刻只有一个NameNode处于active状态是至关重要的。否则，**两个NameNode的数据状态就会产生分歧(数据冲突)**，可能丢失数据，或者产生错误的结果。为了保证这点，这就需要利用使用ZooKeeper了。首先HDFS集群中的两个NameNode都在ZooKeeper中注册（在node节点存储数据），当active状态的NameNode出故障时，ZooKeeper能检测到这种情况，它就会自动（通过Watcher监听node节点数据变化-事件！）把standby状态的NameNode切换为active状态。