**Hadoop2的高可用并取代SecondaryNamenode**

在hadoop2.0之前，namenode只有一个，存在单点问题（虽然hadoop1.0有secondarynamenode，checkpointnode，buckcupnode这些，但是单点问题依然存在），在hadoop2.0引入了HA机制。hadoop2.0的HA机制官方介绍了有2种方式，一种是NFS（Network File System）方式，另外一种是QJM（Quorum Journal Manager）方式。

2 基本原理

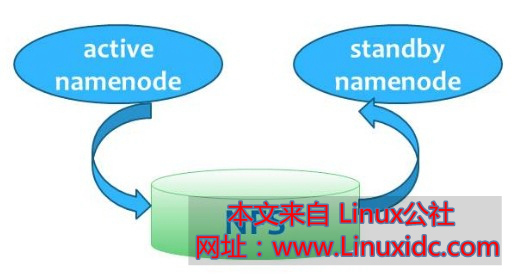
hadoop2.0的HA 机制有两个namenode，一个是active namenode，状态是active；另外一个是standby namenode，状态是standby。两者的状态是可以切换的，但不能同时两个都是active状态，最多只有1个是active状态。只有active namenode提供对外的服务，standby namenode是不对外服务的。active namenode和standby namenode之间通过NFS或者JN（journalnode，QJM方式）来同步数据。

active namenode会把最近的操作记录写到本地的一个edits文件中（edits file），并传输到NFS或者JN中。standby namenode定期的检查，从NFS或者JN把最近的edit文件读过来，然后把edits文件和fsimage文件合并成一个新的fsimage，合并完成之后会通知active namenode获取这个新fsimage。active namenode获得这个新的fsimage文件之后，替换原来旧的fsimage文件。

这样，保持了active namenode和standby namenode的数据的实时同步，standby namenode可以随时切换成active namenode（譬如active namenode挂了）。而且还有一个原来hadoop1.0的secondarynamenode，checkpointnode，buckcupnode的功能：合并edits文件和fsimage文件，使fsimage文件一直保持更新。所以启动了hadoop2.0的HA机制之后，secondarynamenode，checkpointnode，buckcupnode这些都不需要了。

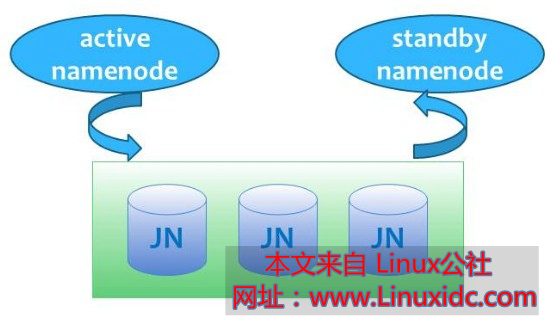
3 NFS方式

NFS作为active namenode和standby namenode之间数据共享的存储。active namenode会把最近的edits文件写到NFS，而standby namenode从NFS中把数据读过来。这个方式的缺点是，如果active namenode或者standby namenode有一个和NFS之间网络有问题，则会造成他们之前数据的同步出问题。



4 QJM（Quorum Journal Manager ）方式

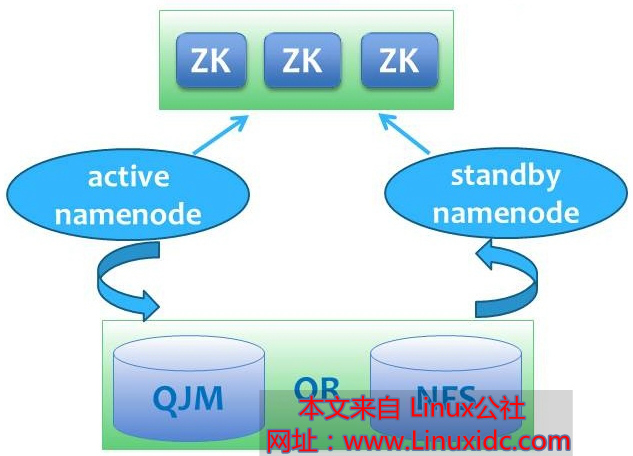
QJM的方式可以解决上述NFS容错机制不足的问题。active namenode和standby namenode之间是通过一组journalnode（数量是奇数，可以是3,5,7...,2n+1）来共享数据。active namenode把最近的edits文件写到2n+1个journalnode上，只要有n+1个写入成功就认为这次写入操作成功了，然后standby namenode就可以从journalnode上读取了。可以看到，QJM方式有容错的机制，可以容忍n个journalnode的失败。



5 主备节点的切换

active namenode和standby namenode可以随时切换。当active namenode挂掉后，也可以把standby namenode切换成active状态，成为active namenode。可以人工切换和自动切换。人工切换是通过执行HA管理的命令来改变namenode的状态，从standby到active，或者从active到standby。自动切换则在active namenode挂掉的时候，standby namenode自动切换成active状态，取代原来的active namenode成为新的active namenode，HDFS继续正常工作。

主备节点的自动切换需要配置zookeeper。active namenode和standby namenode把他们的状态实时记录到zookeeper中，zookeeper监视他们的状态变化。当zookeeper发现active namenode挂掉后，会自动把standby namenode切换成active namenode。



6 实战tips

* QJM方式有明显的优点，一是本身就有fencing的功能，二是通过多个journal节点增强了系统的健壮性，所以建议在生成环境中采用QJM的方式。
* journalnode消耗的资源很少，不需要额外的机器专门来启动journalnode，可以从hadoop集群中选几台机器同时作为journalnode。

介绍完hadoop2.0的HA的基本原理之后，后面的文章会分别详细介绍这两种方式的配置和原理。