> 先达到极限，然后再突破它

HA高可用

HA概述

1. 所谓HA（High Available），即高可用（7*24小时不中断服务）。
2. 实现高可用最关键的策略是**消除单点故障。**HA严格来说应该分成各个组件的HA机制：HDFS的HA和YARN的HA。
3. Hadoop2.0之前，在HDFS集群中NameNode存在单点故障（SPOF）。
4. NameNode主要在以下两个方面影响HDFS集群

   NameNode机器发生意外，如宕机，集群将无法使用，直到管理员重启

   NameNode机器需要升级，包括软件、硬件升级，此时集群也将无法使用

   > HDFS HA功能通过**配置Active/Standby两个NameNodes**实现在集群中对NameNode的热备来解决上述问题。如果出现故障，如机器崩溃或机器需要升级维护，这时可通过此种方式将NameNode很快的切换到另外一台机器。

HDFS-HA工作要点

1.元数据管理方式需要改变

> 内存中各自保存一份元数据；Edits日志只有Active状态的NameNode节点可以做写操作；两个NameNode都可以读取Edits；共享的Edits放在一个共享存储中管理（qjournal和NFS两个主流实现）；

2.需要一个状态管理功能模块

> **实现了一个zkfailover，常驻在每一个namenode所在的节点，每一个zkfailover负责监控自己所在NameNode节点**，利用zk进行状态标识，当需要进行状态切换时，由zkfailover来负责切换，切换时需要防止brain split现象的发生。

3.必须保证两个NameNode之间能够ssh无密码登录

4.隔离（Fence），即同一时刻仅仅有一个NameNode对外提供服务

HDFS-HA自动故障转移工作机制

手动转移命令

`hdfs haadmin -failover`

但是手动转移不够方便，我们需要配置自动故障转移，自动故障转移为HDFS部署增加了两个新组件ZooKeeper和ZKFailoverController（ZKFC）进程

HA的自动故障转移依赖于ZooKeeper的以下功能：

- 故障检测：集群中的每个NameNode在ZooKeeper中**维护了一个持久会话**，如果机器崩溃，ZooKeeper中的会话将终止，ZooKeeper通知另一个NameNode需要触发故障转移。

- 现役NameNode选择：ZooKeeper提供了一个简单的机制用于**唯一的选择一个节点为active状态。**如果目前现役NameNode崩溃，另一个节点可能从ZooKeeper获得特殊的排外锁以表明它应该成为现役NameNode。

  ZKFC是自动故障转移中的另一个新组件，是ZooKeeper的客户端，也**监视和管理NameNode的状态。**每个运行NameNode的主机也运行了一个ZKFC进程，ZKFC负责；

- 健康监测：ZKFC使用一个**健康检查命令定期地ping与之在相同主机的NameNode**，只要该NameNode及时地回复健康状态，ZKFC认为该节点是健康的。如果该节点崩溃，冻结或进入不健康状态，健康监测器标识该节点为非健康的。

- ZooKeeper会话管理：**当本地NameNode是健康的，ZKFC保持一个在ZooKeeper中打开的会话。**如果本地NameNode处于active状态，ZKFC也保持一个特殊的znode锁，该锁使用了ZooKeeper对短暂节点的支持，如果会话终止，锁节点将自动删除。

- 基于ZooKeeper的选择：如果本地NameNode是健康的，且ZKFC发现没有其它的节点当前持有znode锁，它将为自己获取该锁。如果成功，则它已经赢得了选择，**并负责运行故障转移进程以使它的本地NameNode为Active。**故障转移进程与前面描述的手动故障转移相似，首先如果必要保护之前的现役NameNode，然后本地NameNode转换为Active状态。