**HDFS HA配置详解**

HDFS 集群中NameNode 存在单点故障（SPOF ）。对于只有一个NameNode 的集群，如果NameNode 机器出现意外downtime，那么整个集群将无法使用，直到NameNode 重新启动。HDFS 的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Active NN的downtime，就会切换到Standby使得NN服务不间断。HDFS HA依赖zookeeper，下面是测试的过程。

环境如下  
主机：master、slave[1-2]，CentOS 6.5  
Hadoop 2.6.5  
ZooKeeper 3.4.5

|  |  |  |
| --- | --- | --- |
| **Host** | **HDFS** | **ZooKeeper** |
| master | NN,ZKFC,JournalNode | Server |
| slave1 | NN,ZKFC,JournalNode,DN | Server |
| slave2 | DN | Server |

**1. 启动ZooKeeper**

三台机器启动zookeeper

[hadoop@master ~]$ zkServer.sh start

JMX enabled by default

Using config: /opt/bigdata/zookeeper/bin/../conf/zoo.cfg

Starting zookeeper ... STARTED

[hadoop@master ~]$ zkServer.sh status

JMX enabled by default

Using config: /opt/bigdata/zookeeper/bin/../conf/zoo.cfg

Mode: follower

**2. 修改Hadoop配置**

core-site中需要使用ha.zookeeper.quorum设置ZooKeeper服务器节点。另外fs.defaultFS需要设置成HDFS的逻辑服务名（需与hdfs-site.xml中的dfs.nameservices一致）。

$ core-site.xml

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://bigdata</value>

</property>

<property>

<name>hadoop.tmp.dir</name>

<value>/data/hdfs\_tmp/</value>

</property>

<property>

<name>hadoop.logfile.size</name>

<value>104857600</value>

</property>

<property>

<name>hadoop.logfile.count</name>

<value>10</value>

</property>

<property>

<name>io.file.buffer.size</name>

<value>131072</value>

</property>

<property>

<name>ha.zookeeper.quorum</name>

<value>master,slave1,slave2</value>

</property>

</configuration>

hdfs-site.xml中需要添加的设置较多：  
dfs.nameservices —– HDFS NN的逻辑名称，使用上面设置的bigdata  
dfs.ha.namenodes.bigdata —– 给定服务逻辑名称bigdata的节点列表  
dfs.namenode.rpc-address.bigdata.nn1 —– bigdata中nn1节点对外服务的RPC地址  
dfs.namenode.http-address.bigdata.nn1 —– bigdata中nn1节点对外服务的http地址  
dfs.namenode.shared.edits.dir —– 设置一组 journalNode 的 URI 地址，active NN 将 edit log 写入这些JournalNode，而 standby NameNode 读取这些 edit log，并作用在内存中的目录树中。如果journalNode有多个节点则使用分号分割。该属性值应符合以下格式qjournal://host1:port1;host2:port2;host3:port3/journalId  
dfs.journalnode.edits.dir —– JournalNode 所在节点上的一个目录，用于存放 editlog 和其他状态信息。  
dfs.ha.automatic-failover.enabled —– 启动自动failover。自动failover依赖于zookeeper集群和ZKFailoverController（ZKFC），后者是一个zookeeper客户端，用来监控NN的状态信息。每个运行NN的节点必须要运行一个zkfc。zkfs提供了下面的功能：  
**Health monitoring** zkfc定期对本地的NN发起health-check的命令，如果NN正确返回，那么这个NN被认为是OK的。否则被认为是失效节点。  
**ZooKeeper session management** 当本地NN是健康的时候，zkfc将会在zk中持有一个session。如果本地NN又正好是active的，那么zkfc还有持有一个”ephemeral”的节点作为锁，一旦本 地NN失效了，那么这个节点将会被自动删除。  
**ZooKeeper-based election** 如果本地NN是健康的，并且zkfc发现没有其他的NN持有那个独占锁。那么他将试图去获取该锁，一旦成功，那么它就需要执行Failover，然后成为active的NN节点。Failover的过程是：第一步，对之前的NN执行fence，如果需要的话。第二步，将本地NN转换到active状态。  
启动zkfc的方法如下：hadoop-daemon.sh start zkfc。通过start-dfs.sh会自动启动该进程，一般无需手动起停。  
dfs.client.failover.proxy.provider.myhadoop —– 客户端与 active NameNode 进行交互的 Java 实现类，DFS 客户端通过该类寻找当前的active NN。  
dfs.ha.fencing.methods —– 解决HA集群脑裂问题（即出现两个 master 同时对外提供服务，导致系统处于不一致状态）。在 HDFS HA中，JournalNode 只允许一个 NameNode 写数据，不会出现两个 active NameNode 的问题，  
但是，当主备切换时，之前的 active NameNode 可能仍在处理客户端的 RPC 请求，为此，需要增加隔离机制（fencing）将之前的 active NameNode 杀死。常用的fence方法是sshfence，要指定ssh通讯使用的密钥dfs.ha.fencing.ssh.private-key-files和连接超时时间。

$ hdfs-site.xml

<configuration>

<property>

<name>dfs.nameservices</name>

<value>bigdata</value>

</property>

<property>

<name>dfs.ha.namenodes.bigdata</name>

<value>nn1,nn2</value>

</property>

<property>

<name>dfs.namenode.rpc-address.bigdata.nn1</name>

<value>master:8020</value>

</property>

<property>

<name>dfs.namenode.rpc-address.bigdata.nn2</name>

<value>slave1:8020</value>

</property>

<property>

<name>dfs.namenode.http-address.bigdata.nn1</name>

<value>master:50070</value>

</property>

<property>

<name>dfs.namenode.http-address.bigdata.nn2</name>

<value>slave1:50070</value>

</property>

<property>

<name>dfs.namenode.shared.edits.dir</name>

<value>qjournal://master:8485;slave1:8485;slave2:8485/bigdata </value>

</property>

<property>

<name>dfs.ha.automatic-failover.enabled</name>

<value>true</value>

</property>

<property>

<name>dfs.journalnode.edits.dir</name>

<value>/data/journal</value>

</property>

<property>

<name>dfs.client.failover.proxy.provider.bigdata</name>

<value>org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider</value>

</property>

<property>

<name>dfs.ha.fencing.methods</name>

<value>sshfence</value>

<description>how to communicate in the switch process</description>

</property>

<property>

<name>dfs.ha.fencing.ssh.private-key-files</name>

<value>/home/hadoop/.ssh/id\_rsa</value>

<description>the location stored ssh key</description>

</property>

<property>

<name>dfs.ha.fencing.ssh.connect-timeout</name>

<value>5000</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>/data/hdfs/data</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>/data/hdfs/name</value>

</property>

<property>

<name>dfs.namenode.handler.count</name>

<value>8</value>

</property>

<property>

<name>dfs.replication</name>

<value>2</value>

</property>

</configuration>

分发到其他节点：

[hadoop@master hadoop]$ scp core-site.xml hdfs-site.xml hadoop@slave1:/opt/bigdata/hadoop/etc/hadoop

[hadoop@master hadoop]$ scp core-site.xml hdfs-site.xml hadoop@slave2:/opt/bigdata/hadoop/etc/hadoop

创建journal目录（master、slave1、slave2）

[hadoop@master data]$ cd /data/

[hadoop@master data]$ sudo mkdir journal

[hadoop@master data]$ sudo chown hadoop journal/

[hadoop@master data]$ sudo chgrp hadoop journal/

**3. 启动NameNode HA**

第一次启动格式化HDFS。 格式化HDFS的过程中，HA会journalnode通讯，所以需要先把三个节点的journalnode启动。（master、slave1、slave2）

[hadoop@master data]$ /opt/bigdata/hadoop/sbin/hadoop-daemon.sh start journalnode

starting journalnode, logging to /opt/bigdata/hadoop/logs/hadoop-hadoop-journalnode-master.out

其中一台namenode初始化zkfc（Master）

[hadoop@master hadoop]$ hadoop zkfc -formatZK

17/12/11 20:43:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

17/12/11 20:43:48 INFO tools.DFSZKFailoverController: Failover controller configured for NameNode NameNode at master/192.168.88.88:8020

17/12/11 20:43:48 INFO zookeeper.ZooKeeper: Client environment:zookeeper.version=3.4.6-1569965, built on 02/20/2014 09:09 GMT

17/12/11 20:43:48 INFO zookeeper.ZooKeeper: Client environment:host.name=kerberos.hadoop.com

17/12/11 20:43:48 INFO zookeeper.ZooKeeper: Client environment:java.version=1.8.0\_151

17/12/11 20:43:48 INFO zookeeper.ZooKeeper: Client environment:java.vendor=Oracle Corporation

17/12/11 20:43:48 INFO zookeeper.ZooKeeper: Client environment:java.home=/opt/jdk1.8.0\_151/jre

…

格式成功后，查看zookeeper中可以看到

[hadoop@master sbin]$ zkCli.sh

[zk: localhost:2181(CONNECTED) 0] ls /

[zookeeper, tmp, hadoop-ha]

[zk: localhost:2181(CONNECTED) 1] ls /hadoop-ha

[bigdata]

[zk: localhost:2181(CONNECTED) 2]

其中一个namenode节点格式化（master）

[hadoop@master hadoop]$ hadoop namenode -format

把刚刚格式化之后的元数据拷贝到另外一个namenode上

#启动刚刚格式化的namenode（master）

[hadoop@master sbin]$ hadoop-daemon.sh start namenode

starting namenode, logging to /opt/bigdata/hadoop/logs/hadoop-hadoop-namenode-master.out

#在没有格式化的namenode上执行：hdfs namenode –bootstrapStandby（slave1）

[hadoop@slave1 bin]$ /opt/bigdata/hadoop/bin/hdfs namenode -bootstrapStandby

#启动第二个namenode（slave1）

[hadoop@slave1 bin]$ /opt/bigdata/hadoop/sbin/hadoop-daemon.sh start namenode

starting namenode, logging to /opt/bigdata/hadoop/logs/hadoop-hadoop-namenode-slave1.out

通过start-dfs.sh 直接启动所有服务

[hadoop@master sbin]$ start-dfs.sh

17/12/11 22:37:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Starting namenodes on [master slave1]

slave1: starting namenode, logging to /opt/bigdata/hadoop/logs/hadoop-hadoop-namenode-slave1.out

master: starting namenode, logging to /opt/bigdata/hadoop/logs/hadoop-hadoop-namenode-master.out

slave1: starting datanode, logging to /opt/bigdata/hadoop/logs/hadoop-hadoop-datanode-slave1.out

slave2: starting datanode, logging to /opt/bigdata/hadoop/logs/hadoop-hadoop-datanode-slave2.out

Starting journal nodes [master slave1 slave2]

slave2: journalnode running as process 11948. Stop it first.

slave1: journalnode running as process 8974. Stop it first.

master: journalnode running as process 20285. Stop it first.

17/12/11 22:37:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Starting ZK Failover Controllers on NN hosts [master slave1]

slave1: starting zkfc, logging to /opt/bigdata/hadoop/logs/hadoop-hadoop-zkfc-slave1.out

master: starting zkfc, logging to /opt/bigdata/hadoop/logs/hadoop-hadoop-zkfc-master.out

查看进程

[hadoop@master sbin]$ jps

4084 JournalNode

4646 NameNode

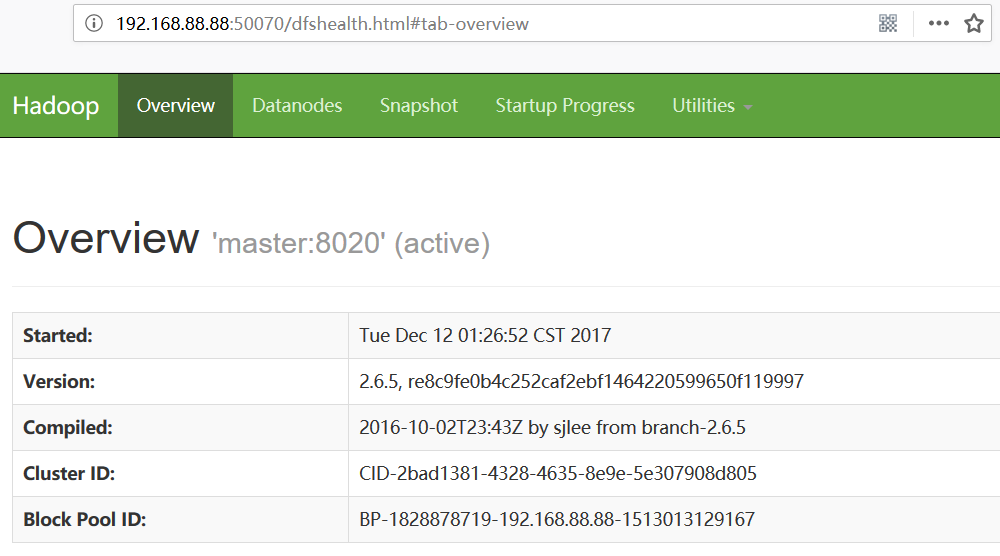
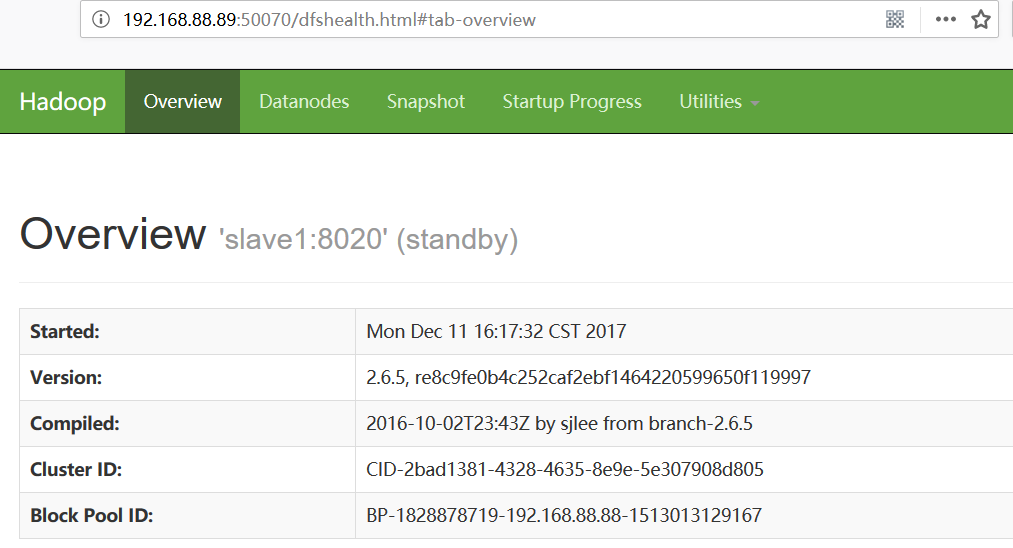
5336 Jps

3962 QuorumPeerMain

5215 DFSZKFailoverController

[hadoop@master sbin]$

使用浏览器访问192.168.88.88:50070会看到该节点已经成为active

然后kill掉master上的active NN进程，standby NN会成为active。

注意：手动切换时，会提示下面警告。所以一般在启动zkfc的情况下也无需进行切换。

[hadoop@master sbin]$ hdfs haadmin -transitionToActive nn1

Automatic failover is enabled for NameNode at slave1/192.168.88.89:8020

Refusing to manually manage HA state, since it may cause

a split-brain scenario or other incorrect state.

If you are very sure you know what you are doing, please

specify the forcemanual flag.

[hadoop@master sbin]$

**参考**

http://www.cloudera.com/content/cloudera-content/cloudera-docs/CDH4/latest/CDH4-High-Availability-Guide/cdh4hag\_topic\_2\_3.html

http://blog.csdn.net/u010967382/article/details/30976935

http://blog.csdn.net/chenpingbupt/article/details/7922089