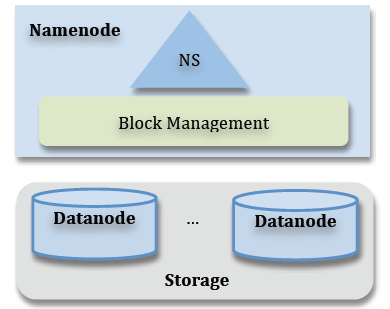
Hadoop 2.0 Federation的实现

# 1、HDFS 1.0架构

当前的HDFS架构有两层，如图所示：



Namespace管理层，管理Namespace中的directories,文件和blocks。提供文件和目录的creation/modification/deletion/listing

Block管理层，主要分为两个部分:

1) Block管理，管理Datanodes，提供block相关的操作(creation/deletion/modifactin/

replicaPlacement/blockReplication等)

2)Storage的管理：物理存储管理，访问block数据

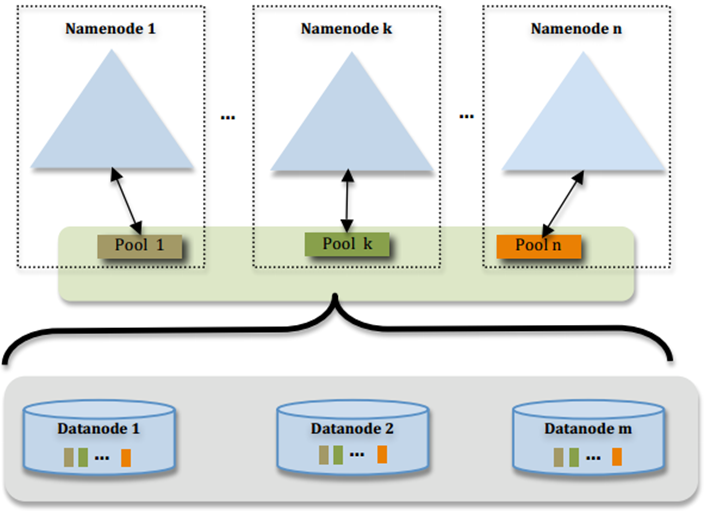
实现的架构如上图所示：

1) NN实现Namespace和Block的管理，DN提供物理的存储和访问BLK数据。DN注册到NN的blk管理层，为HDFS提供Storage层。

2) DN仅仅与NN的blk管理层通信而不会参与的Namespace管理。Blk管理层一部分在DN中，一部分在NN中，现在的实现及Java API都没有提供完全的隔离。

# 2、Hadoop 2.0 federation框架

Federation的实现方式如下图所示：



从图中可以看出：

1)多个Namenode共用一个集群里DN上的存储资源，每个Namenode都可以单独对外提供服务

2)每个Namenode都会定义一个存储池(pool)，有单独的ID，每个Datanode都为所有存储池提供存储

3) DN会按照存储池向其对应的NN汇报块信息，同时DN会向所有的Namenode汇报本地存储可用资源情况

4) 如果需要客户端方便的访问若干个NN个资源，可以使用客户端挂载表，把不同的目录映射到不同的Namenode上，但Namenode上必须存在相应的目录。

以上的设计可以分离命名空间关了和块存储管理，提供良好的扩展型的同时允许其他文件系统或应用直接使用块存储池，同一的块存储管理保证了资源的利用率，可以只通过防火墙配置达到一定的文件访问隔离，而无需使用复杂的Kerberos认证。

# 3、Hadoop 2.0 federation的安装与配置

搭建的hadoop2.0 Federation 集群，共有三个节点：host1,host2,host3，其中host1,host2作为federation的namenode，host1-host3作为datanode节点。

配置如下：

hdfs-site.xml:

*<configuration>*

*<property>*

*<name>dfs.federation.nameservices</name>*

*<value>host1,host2</value>*

*</property>*

*<property>*

*<name>dfs.namenode.rpc-address.host1</name>*

*<value>host1:9000</value>*

*</property>*

*<property>*

*<name>dfs.namenode.http-address.host1</name>*

*<value>host1:50070</value>*

*</property>*

*<property>*

*<name>dfs.namenode.rpc-address.host2</name>*

*<value>host2:9000</value>*

*</property>*

*<property>*

*<name>dfs.namenode.http-address.host2</name>*

*<value>host2:50070</value>*

*</property>*

*<property>*

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

*<value>/home/hadoop/name</value>*

*</property>*

*<property>*

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

*<value>/home/hadoop/data</value>*

*</property>*

*<property>*

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

*<value>host1:2181,host2:2181,host3:2181</value>*

*</property>*

*</configuration>*

core-site.xml的配置如下：

host1:

*<configuration>*

*<property>*

*<name>fs.default.name</name>*

*<value>hdfs://host1:9000</value>*

*</property>*

*</configuration>*

host2:

*<configuration>*

*<property>*

*<name>fs.default.name</name>*

*<value>hdfs://host2:9000</value>*

*</property>*

*</configuration>*

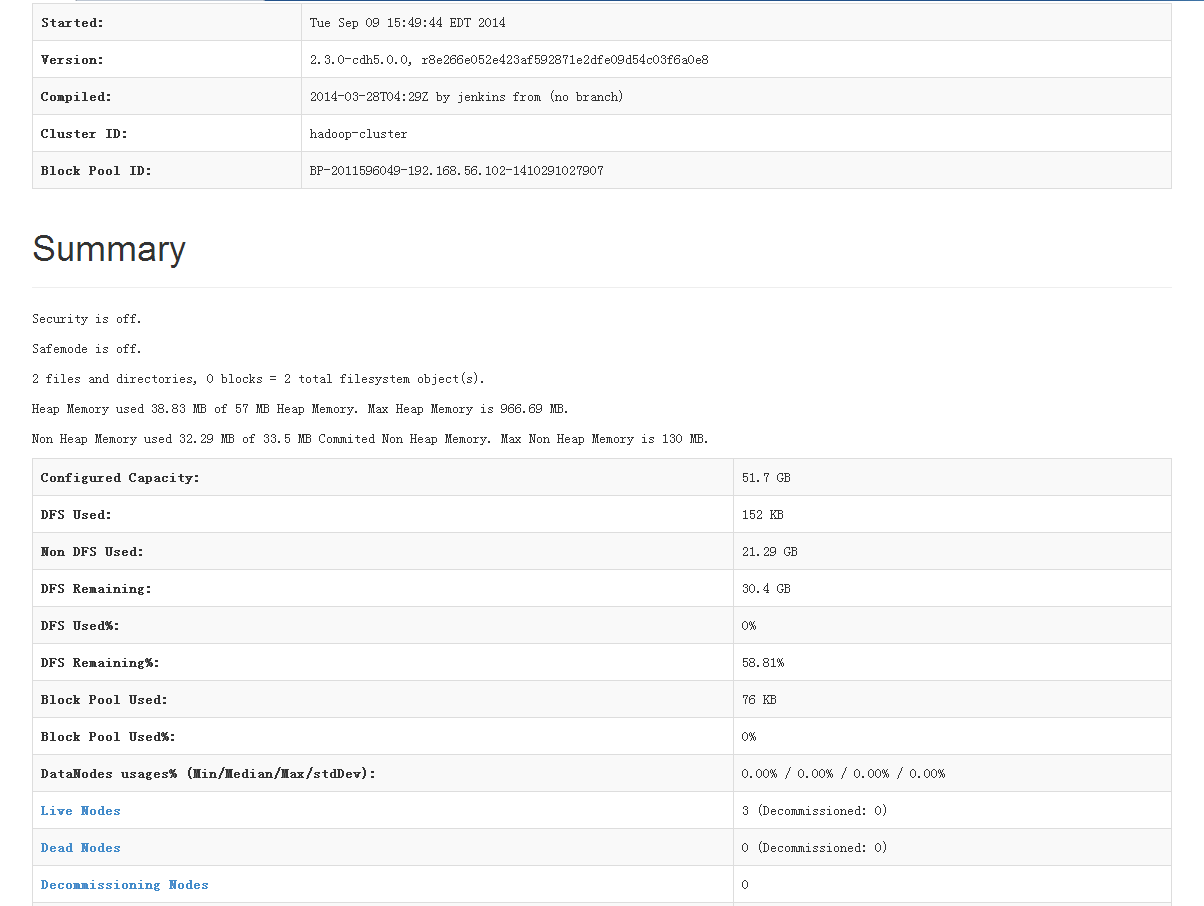
格式化命令：

*$bin/hdfs namenode -format -clusterId hadoop-cluster*

启动命令：

*$sbin/start-dfs.sh*

通过http://192.168.56.101:50070/及http://192.168.56.102:50070/访问namenode的信息，如下所示：



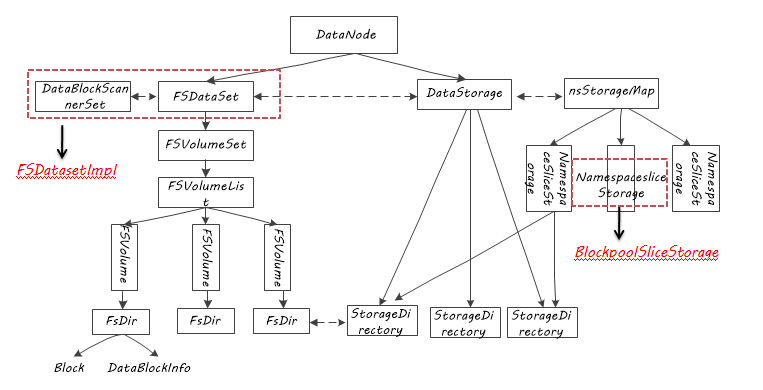
页面中有ClusterId，BlockPoolID及versiond等信息，host1和host2的poolId是不同的，且共有三个datandode节点。

# 4、Hadoop 2.0 federation源码分析

1) Block

为了支持federation，实现类ExtendBlock，是Block的子类。在类中增加了成员变量：blockPool

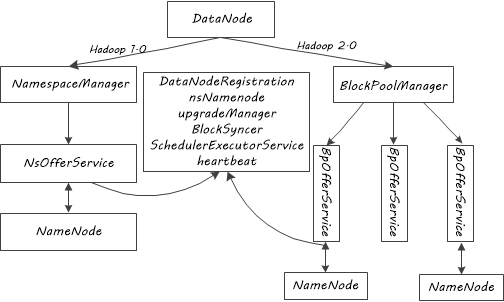
1. DataStorage



如上图所示，原图是Hadoop 1.0的DataStorage信息，红色是替换后的Datanode元数据。

在FSVolumeImpl中增加了成员变量Map<String,BlockPoolSlice> bpSlices。在datanode集群中，BlockPoolSlice的bpid相同则属于同一blockPool。

1. 与NameNode RPC通信过程



如上图所示，datanode中保存了多个Namenode管理的block，以blockpoolID作为区分，datanode为每个namenode创建一个BpOfferservice，BpOfferService完成datanode与该namenode的通信，注册、发送心跳信息及reportBlock等过程。该BpOfferService的通信内容至于该namenode相关。