Hadoop NFS

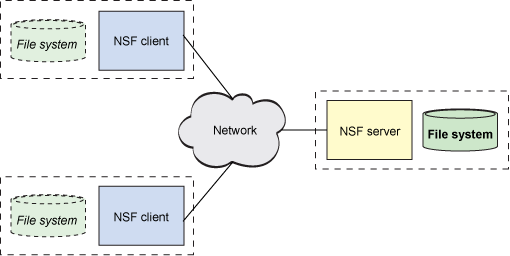
NFS(Network File Systen)，网络文件系统，是FreeBSD支持的文件系统中的一种，运行网络中的计算机之间通过TCP/IP网络共享资源。在NFS的应用中，本地NFS客户端应用可以透明的读写位于远端NFS服务器上的文件，就像访问本地文件一样。在Hadoop中，支持NFS Gateway，将HDFS作为客户端本地文件系统的一部分挂载到本地文件系统，通过NFS Gateway:

* 用户可以在基于NFSv3客户端兼容的操作系统上的本地文件系统上浏览HDFS文件系统
* 用户可以从挂载到本地文件上的HDFS上下载文件
* 用户可以从本地文件系统上直接上传文件到HDFS文件系统
* 用户可以通过挂载点直接将数据流写入到HDFS，目前支持文件append

# 1. NFS（Network File System）

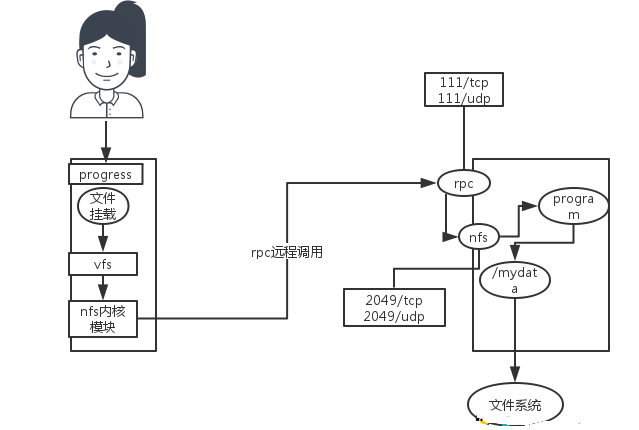
## 1.1 NFS简介

NFS使不同计算机之间通过网络进行文件共享的网络协议，通过网络让不同的服务器之间可以共享文件或者目录。NFS的客户端一般是应用服务器，可以通过挂载的方式将NFS服务器端共享的目录挂载到NFS客户端本地目录下，如下所示：



NFS采用客户/服务器工作模式，在NFS服务器上将共享的文件目录设置为共享目录后，其他客户端就可以将这个目录挂载到自己系统中某个目录下，这个目录如同本地磁盘分区一样，可以使用cp、mv等与磁盘相关的命令透明得访问服务器上的文件系统。

NFS可以在网络中进行文件共享，但是NFS协议本身并没有提供数据传输功能，必须借助与RPC协议来实现数据的传输，可以将NFS服务器看成是RPC服务器，NFS客户端看做为RPC客户端，这样NFS服务器和NFS客户端之间可以通过RPC协议进行数据传输。流程如下图所示：



## 1.2 NFS的安装和启动

NFS依赖RPC通信，使用NFS服务至少依赖3个系统守护进程：

* rpc.nfsd，NFS守护进程，管理客户端登入服务器
* rpc.mountd，RPC安装守护进程，主要功能是管理NFS的文件系统，客户端通过nfsd登录NFS服务器后，必须通过文件使用权限的验证，读取nfs的配置文件/etc/exports来对比客户端的权限
* portmap，进行端口映射，当客户端尝试连接并使用RPC服务器提供的NFS服务时，portmap会将所管理的与服务对应的端口号提供给客户端，从而使客户端可以通过该端口向服务器请求服务。

要部署NFS服务，需要安装下面的软件包：

* nfs-utils，NFS服务的主程序
* rpcbind，RPC的主程序，在启动之前，做好端口和功能的对应映射工作。

1. 安装

*$yum -y install nfs-utils rpcbind*

安装后如下：

*# rpm -qa|grep nfs*

*libnfsidmap-0.25-15.el7.x86\_64*

*nfs-utils-1.3.0-0.33.el7\_3.x86\_64*

*# rpm -qa|grep rpcbind*

*rpcbind-0.2.0-38.el7\_3.1.x86\_64*

1. 启动NFS服务

* 先启动rpc服务

*#rpcbind start*

*#netstat -anp|grep rpcbind*

*udp 0 0 0.0.0.0:111 0.0.0.0:\* 9971/rpcbind*

*udp 0 0 0.0.0.0:818 0.0.0.0:\* 9971/rpcbind*

*udp6 0 0 :::111 :::\* 9971/rpcbind*

*udp6 0 0 :::818 :::\* 9971/rpcbind*

启动RPC服务

* 启动nfs服务

*$systemctl start nfs.service*

查看状态：

*# systemctl status nfs.service*

*● nfs-server.service - NFS server and services*

*Loaded: loaded (/usr/lib/systemd/system/nfs-server.service; disabled; vendor preset: disabled)*

*Active: active (exited) since Mon 2018-02-05 17:07:25 CST; 27min ago*

*Main PID: 10675 (code=exited, status=0/SUCCESS)*

*Memory: 0B*

*CGroup: /system.slice/nfs-server.service*

* 配置exports，配置内容如下：

*# vim /etc/exports*

*/tmp/nfs 10.139.4.82/24(rw,sync,no\_subtree\_check)*

使用下面的命令使其生效：

*$exportfs –a*

通过下面的命令查看共享的目录：

*# exportfs*

*/tmp/nfs 10.139.4.82/24*

http://blog.topspeedsnail.com/archives/4109

注：exports配置详解

* rw: read-write，可读写
* ro:read-only，只读
* sync，文件同时写入硬盘和内存
* async，文件暂存于内存，而不是直接写入内存
* no\_root\_squash，nfs客户端连接服务端时如果使用的是root的话，那么服务器端分析的目录来说也拥有root全新
* root\_squash，nfs客户端连接时如果使用root，对于服务端分享的目录来说，拥有匿名用户权限，通常使用nobody或nfsbody身份
* all\_squash，不论NFS客户端连接时使用什么用户，对服务端分享的目录来说都拥有匿名用户权限
* anonuid，匿名用户UID，可自行设定
* anogid，匿名用户GID值
* fsid=0，将挂载目录包装成根目录

## 1.3 NFS的使用

客户端安装nfsutils，命令如下：

*$yum -y install nfs-utils*

创建本地目录：

*# mkdir /mnt/nfs*

修改本地的/etc/exports文件

*# vim /etc/exports*

*/tmp/nfs 10.139.4.82/24(rw,no\_root\_squash,no\_all\_squash,sync,no\_subtree\_check)*

挂载目录：

*mount -t nfs -o nolock,nfsvers=3,vers=3 10.139.4.82:/tmp/nfs /mnt/nfs*

查看如下：

*# ll /mnt/nfs/*

*total 4 ----rwxrwx 1 root root 9 Feb 5 17:54 hellonfs*

hellonfs文件为nfs server上的/tmp/nfs目录中的文件

修改hellonfs文件：

*test nfs*

*add content*

在其他nfs客户端查看文件内容如下：

*test nfs*

*add content*

文件已经进行了同步。

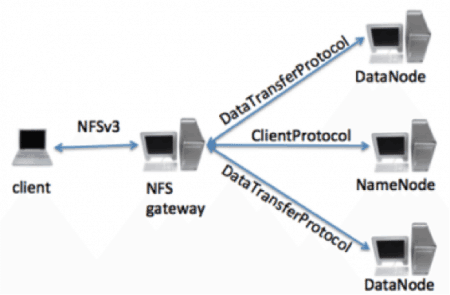
http://www.tozzofrios.com.br/gerenciador/nfs-linux

# 2. NFS GateWay（HDFS）

HDFS是分布式文件系统，只能通过hadoop shell或者java api等工具访问hdfs的文件，可以将hdfs文件通过nfs挂载到本地的一个目录上，像操作本地文件系统一样使用hdfs文件。

## 2.1简介

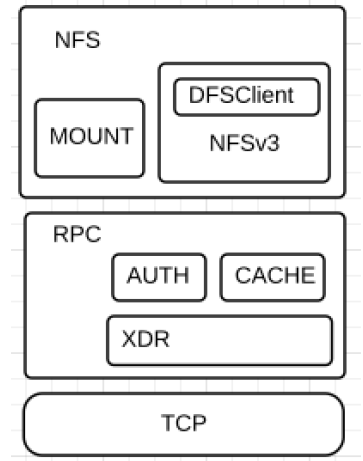
在hadoop中，实现了NFS Gateway，将NFS协议转换成HDFS访问协议，如下图所示：



NFSGateway是一个无状态进程，可以为客户端启动多个NFS Gateway以达到高吞吐量。Gateway主要提供下面功能：

* 支持访问HDFS的NFSv4及其他协议
* NFS Gateway的高可用性
* 支持Keberos认证

下图是实现的功能架构图：



* TCP，仅支持基于TCP通信的NFS
* XDR， External Data Representation，基于RFC 4506的数据类型描述
* ONCRPC，RFC 1831，Open network Computing Remote Procedure call
* AUTHR，NFC Client使用UID/GID与Server进行交互，但是HDFS使用User /Group name
* CACHE
* NFSv3，处理所有的NFS请求，并使用DFSClient来访问HDFS
* MOUNT，处理所有的MOUNT请求

## 2.2.NFS Gateway的配置

1）设置mount用户，运行NFS-Gateway的用户必须能够代理所有NFS挂载点的用户，在core-site.xml中配置参数如下：

*<property>*

*<name>hadoop.proxyuser.nfsserver.groups</name>*

*<value>\*</value>*

*</property>*

*<property>*

*<name>hadoop.proxyuser.nfsserver.hosts</name>*

*<value>\*</value>*

*</property>*

2）默认情况下，export可以被任何客户端挂载，为了更好的控制访问，可以配置下面的属性，value包括机器名字和访问权限，用空格分开，例如：

*<property>*

*<name>dfs.nfs.exports.allowed.hosts</name>*

*<value>\* rw</value>*

*</property>*

其他配置如下：

|  |  |  |
| --- | --- | --- |
| 参数 | 描述 | 默认值 |
| nfs.aix.compatibility.mode.enabled | 从AIX系统访问HDFS NFS Gateway | True |
| nfs.superuser | 与NameNode进程配置相同的hdfs超级用于 |  |
| dfs.namenode.accesstime.precision | 客户端运行挂载时间允许挂载 | 1小时 |
| dfs.nfs3.dump.dir | 更新文件的转存目录，连续的写操作在存入到hdfs之前是保存乱序的写操作，超过阈值后被转存 | /tmp/.hdfs-nfs |
| nfs.rtmax | 读请求的最大请求数据量 | 1048576 |
| nfs.wtmax | 写请求的最大数据量 | 1048576 |
| nfs.allow.insecure.ports | False时运行使用安全端口（小于1024） | true |
| nfs.server.port | Hadoop NFS配置的端口 | 2049 |
| nfs.moutd.port | Hadoop mout进程端口 | 4242 |
| nfs.keytab.file|principal | 开启Keberize时的配置 |  |

## 2.3.NFS Gateway的使用

1）停止机器中的nfs/rpcbind服务

*# service rpcbind status*

*# service nfs status*

在hadoop中包含了portmap的启动程序

2）开启portmap进程

在root下执行：

*$nohup hdfs portmap 2>&1 &*

启动信息如下:

*# ps -ef|grep portmap*

*root 20885 6456 6 15:38 pts/1 00:00:00 /usr/jdk64/jdk1.7.0\_67/bin/java \*\*\* org.apache.hadoop.portmap.Portmap*

*# netstat -anp|grep 20885*

*tcp 0 0 0.0.0.0:111 0.0.0.0:\* LISTEN 20885/java*

*udp 0 0 0.0.0.0:111 0.0.0.0:\* 20885/java*

1. 启动nfs3进程，命令如下：

在hdfs用户下执行：

*&nohup hdfs nfs3 2>&1 &*

启动信息如下：

*$ ps -ef|grep nfs*

*root 17344 2 0 15:17 ? 00:00:00 [nfsiod]*

*hdfs 21282 21262 7 15:40 pts/0 00:00:04 /usr/jdk64/jdk1.7.0\_67/bin/java ..... org.apache.hadoop.hdfs.nfs.nfs3.Nfs3*

*$ netstat -anp|grep 21282*

*tcp 0 0 0.0.0.0:50079 0.0.0.0:\* LISTEN 21282/java*

*tcp 0 0 0.0.0.0:2049 0.0.0.0:\* LISTEN 21282/java*

*tcp 0 0 0.0.0.0:4242 0.0.0.0:\* LISTEN 21282/java*

*udp 0 0 0.0.0.0:4242 0.0.0.0:\* 21282/java*

1. 验证nfs启动

在启动节点查看命令如下：

*# rpcinfo -p localhost*

*program vers proto port service*

*100005 2 tcp 4242 mountd*

*100000 2 udp 111 portmapper*

*100000 2 tcp 111 portmapper*

*100005 1 tcp 4242 mountd*

*100003 3 tcp 2049 nfs*

*100005 1 udp 4242 mountd*

*100005 3 udp 4242 mountd*

*100005 3 tcp 4242 mountd*

*100005 2 udp 4242 mountd*

启动成功。

验证hdfs命名空间是否可被export和挂载，查看如下：

*# showmount -e localhost*

*Export list for localhost:*

*/ \**

1. 将/\*挂载到本地目录/tmp/nfs(用户创建)

*mount -t nfs -o vers=3,proto=tcp,nolock,noacl,sync localhost:/ /tmp/nfs*

查看/tmp/nfs目录如下：

*ll /tmp/nfs*

*total 6*

*drwxrwxrwx 8 yarn hadoop 256 Nov 6 09:34 app-logs*

*drwxr-xr-x 6 hdfs hdfs 192 Sep 12 12:28 apps*

*drwxr-xr-x 4 hdfs hdfs 128 Oct 19 15:04 benchmarks*

*drwxr-xr-x 3 hdfs hdfs 96 Aug 19 09:34 iothrottle*

*drwxr-xr-x 3 mapred hdfs 96 Aug 11 17:39 mapred*

*drwxrwxrwx 4 mapred hadoop 128 Aug 11 17:39 mr-history*

*drwxrwxrwx 4 slider hdfs 128 Aug 11 17:42 slider*

*drwxr-xr-x 2 hdfs hdfs 64 Aug 25 08:47 system*

*drwxr-xr-x 5 hdfs hdfs 160 Oct 17 16:43 terasort*

*drwxrwxrwx 16 hdfs hdfs 512 Feb 6 15:18 tmp*

*drwxrwxrwx 10 hdfs hdfs 320 Nov 6 09:34 user*

挂载成功。

1. 验证在挂载目录进行文件操作，在root下执行命令如下

*$cd /tmp/nfs/tmp*

*$mkdir hinfs*

查看hdfs文件，创建的目录如下：

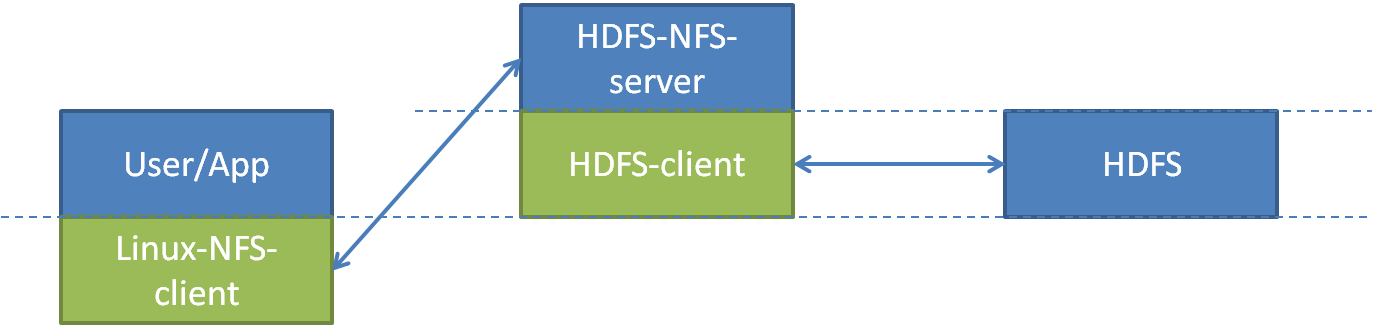
*hdfs dfs -ls /tmp*

*drwxr-x--- - root root 0 2018-02-06 15:20 /tmp/hinfs*

可以确认在hdfs中的/tmp目录下创建了hinfs目录，由于在root下执行因此属性为：root:root

# 3. Hadoop NFS源码分析

在NFS中所有的文件操作，包括文件/目录API和读写文件数据，都是通过RPC调用来实现的。客户端使用本地操作系统提供的系统调用文件系统，当发现需要访问NFS时，虚拟机文件系统会将操作传递给NFS客户端组件，由该组件通过RPC访问相应的NFS服务器。数据访问路径如下图：



用户或程序通过Linux自带的NFS Client访问HDFS NFS服务，然后再由NFS Gateway作为客户端访问HDFS。对源码进行分析，设计思路理解如下：



NFS RpcClient和Server端是基于Netty实现通信，核心Handler为RpcProgramNfs3。

1. Nfs3，作为HDFS 与NFS之间的Gateway，将NFS Client的请求通过Server内维护的DFSClient转发给HDFS集群
2. RpcProgrameNfs3， NFS Server的核心处理Handler，继承SimpleChannelUpstreamHandler

接收NFS 客户端的文件操作请求，对HDFS进行操作。

1. WriteManager，当需要从HDFS中读取或者写入为数据时，通过WriteManager来进行操作，其中维护了文件流及rpc resposne缓存。不涉及到文件流时，例如修改文件权限，创建目录等，直接通过DFSClient与NameNode通信进行操作即可。
2. DFSClientCache，缓存DFSClient
3. RpcResponseCache，有一些文件操作，例如获取文件状态时，这些操作具有幂等函数，因此将这些结果缓存，下次相同操作时直接从缓存中读取。
4. OpenFileCtx，对某个文件操作的上下文封装，包括维护文件流，写入的文件数据等

下面是对HDFS NFS的类进行的分析：



# 附录（参考文献）

1. NFS的配置（中文）：http://blog.csdn.net/xichenguan/article/details/38732537

https://hadoop.apache.org/docs/r2.8.0/hadoop-project-dist/hadoop-hdfs/HdfsNfsGateway.html

https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.6.2/bk\_hdfs-nfs-gateway-user-guide/content/user-guide-hdfs-nfs-instructions.html

1. HDFS NFS: http://www.codeclip.com/2285.html
2. hdfs nfs的介绍：

https://hortonworks.com/blog/simplifying-data-management-nfs-access-to-hdfs/

1. 相关JIRA:

https://issues.apache.org/jira/browse/HDFS-4750