得之我幸，失之我命；
不以物喜，不以己悲;
努力争取，靠向目标!

昵称：大牛笔记
园龄：7年
粉丝：268
关注：23
+加关注

【Hadoop】HDFS的运行原理

博文已转移，请借一步说话

http://www.daniubiji.cn/archives/596

简介

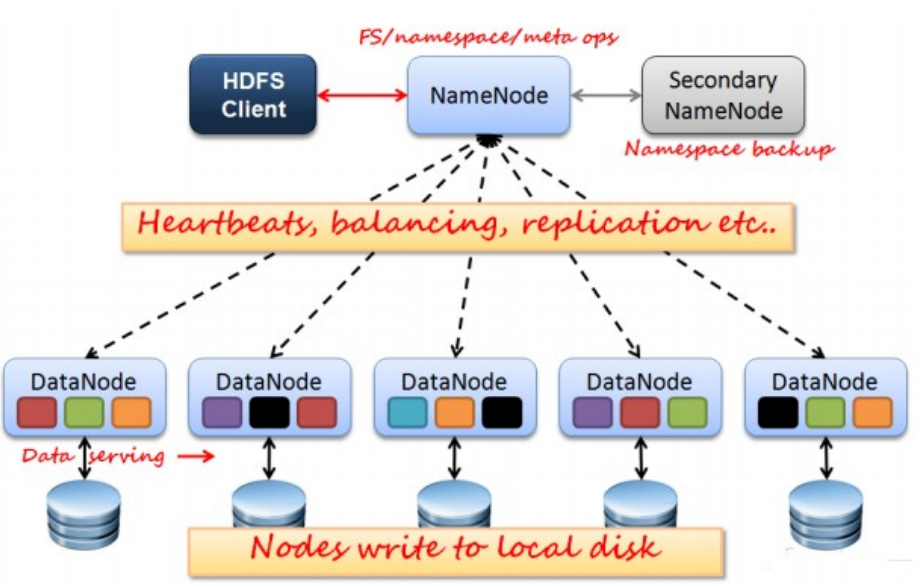
HDFS（Hadoop Distributed File System ）Hadoop分布式文件系统。是根据google发表的论文翻版的。论文为GFS（Google File System）Google 文件系统（中文，英文）。

HDFS有很多特点：

① 保存多个副本，且提供容错机制，副本丢失或宕机自动恢复。默认存3份。

② 运行在廉价的机器上。

③ 适合大数据的处理。多大？多小？HDFS默认会将文件分割成block，64M为1个block。然后将block按键值对存储在HDFS上，并将键值对的映射存到内存中。如果小文件太多，那内存的负担会很重。



如上图所示，HDFS也是按照Master和Slave的结构。分NameNode、SecondaryNameNode、DataNode这几个角色。

NameNode：是Master节点，是大领导。管理数据块映射；处理客户端的读写请求；配置副本策略；管理HDFS的名称空间；

SecondaryNameNode：是一个小弟，分担大哥namenode的工作量；是NameNode的冷备份；合并fsimage和fsedits然后再发给namenode。

DataNode：*Slave*节点，奴隶，干活的。负责存储client发来的数据块block；执行数据块的读写操作。

热备份：b是a的热备份，如果a坏掉。那么b马上运行代替a的工作。

冷备份：b是a的冷备份，如果a坏掉。那么b不能马上代替a工作。但是b上存储a的一些信息，减少a坏掉之后的损失。

fsimage:元数据镜像文件（文件系统的目录树。）

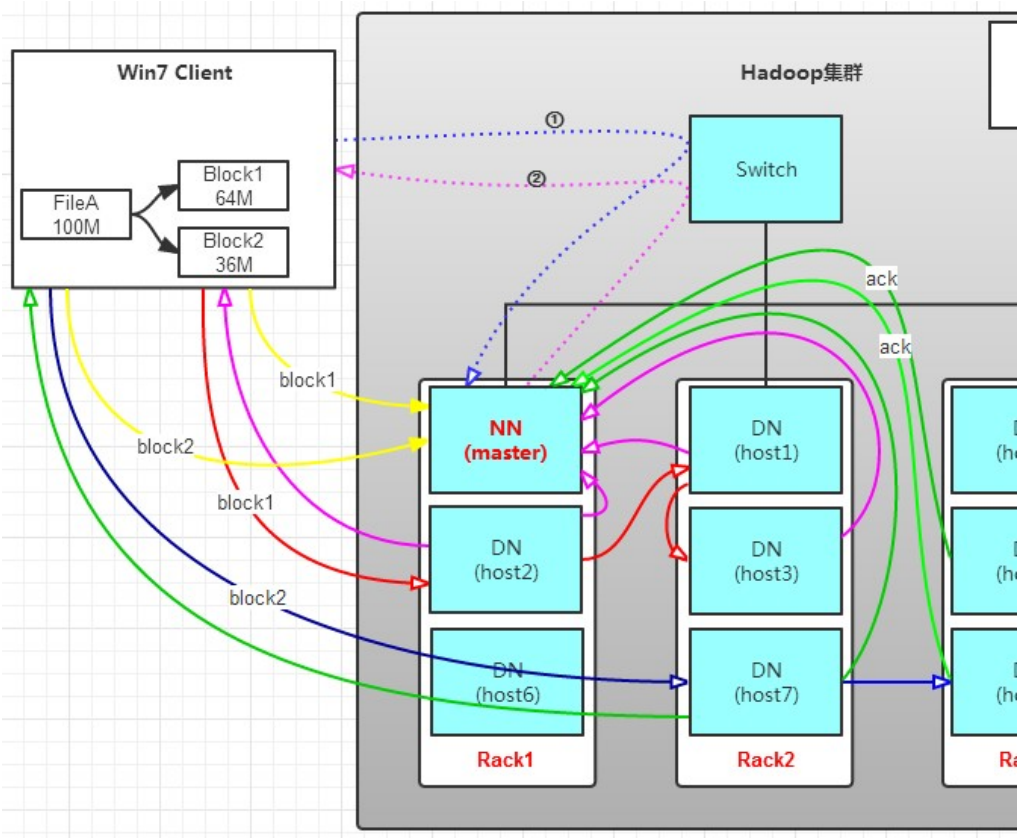
edits：元数据的操作日志（针对文件系统做的修改操作记录）

namenode内存中存储的是=fsimage+edits。

SecondaryNameNode负责定时默认1小时，从namenode上，获取fsimage和edits来进行合并，然后再发送给namenode。减少namenode的工作量。

工作原理

写操作：



有一个文件FileA，100M大小。Client将FileA写入到HDFS上。

HDFS按默认配置。

HDFS分布在三个机架上Rack1，Rack2，Rack3。

a. Client将FileA按64M分块。分成两块，block1和Block2;

b. Client向nameNode发送写数据请求，如图蓝色虚线①------>。

c. NameNode节点，记录block信息。并返回可用的DataNode，如粉色虚线②--------->。

Block1: host2,host1,host3

Block2: host7,host8,host4

原理：

NameNode具有RackAware机架感知功能，这个可以配置。

若client为DataNode节点，那存储block时，规则为：副本1，同client的节点上；副本2，不同机架节点上；副本3，同第二个副本机架的另一个节点上；其他副本随机挑选。

若client不为DataNode节点，那存储block时，规则为：副本1，随机选择一个节点上；副本2，不同副本1，机架上；副本3，同副本2相同的另一个节点上；其他副本随机挑选。

d. client向DataNode发送block1；发送过程是以流式写入。

流式写入过程，

1>将64M的block1按64k的package划分;

2>然后将第一个package发送给host2;

3>host2接收完后，将第一个package发送给host1，同时client想host2发送第二个package；

4>host1接收完第一个package后，发送给host3，同时接收host2发来的第二个package。

5>以此类推，如图红线实线所示，直到将block1发送完毕。

6>host2,host1,host3向NameNode，host2向Client发送通知，说“消息发送完了”。如图粉红颜色实线所示。

7>client收到host2发来的消息后，向namenode发送消息，说我写完了。这样就真完成了。如图黄色粗实线

8>发送完block1后，再向host7，host8，host4发送block2，如图蓝色实线所示。

9>发送完block2后，host7,host8,host4向NameNode，host7向Client发送通知，如图浅绿色实线所示。

10>client向NameNode发送消息，说我写完了，如图黄色粗实线。。。这样就完毕了。

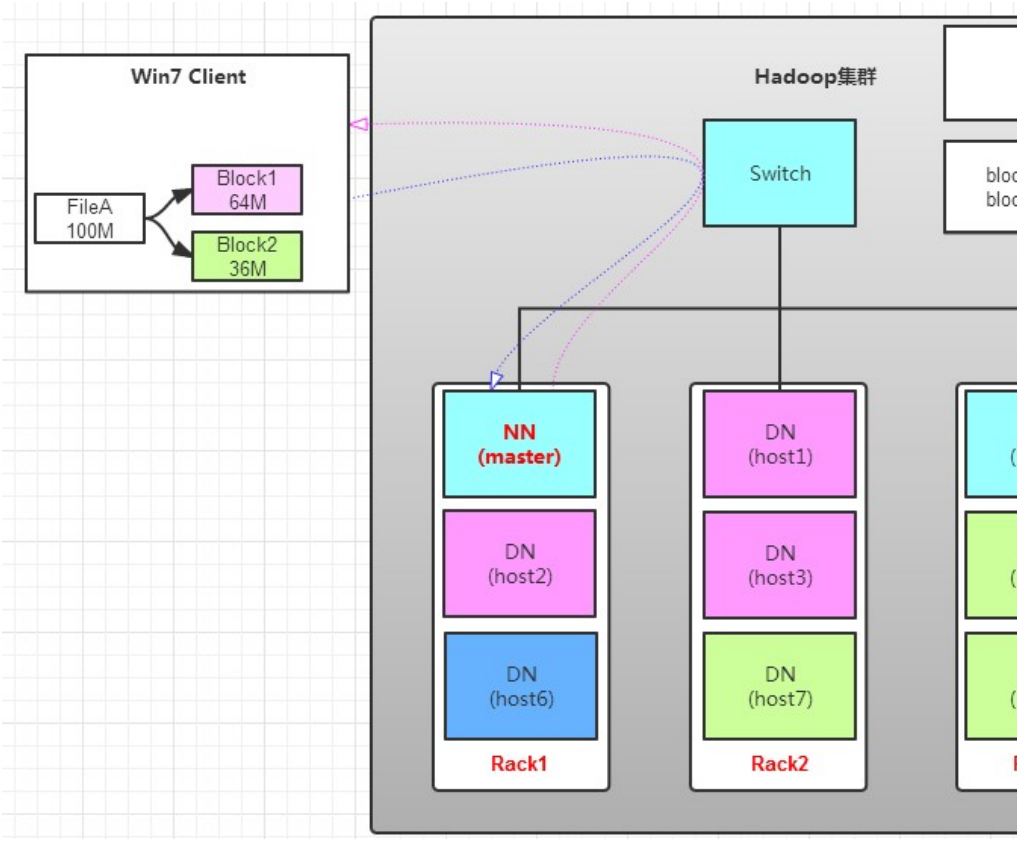
分析，通过写过程，我们可以了解到：

①写1T文件，我们需要3T的存储，3T的网络流量贷款。

②在执行读或写的过程中，NameNode和DataNode通过HeartBeat进行保存通信，确定DataNode活着。如果发现DataNode死掉了，就将死掉的DataNode上的数据，放到其他节点去。读取时，要读其他节点去。

③挂掉一个节点，没关系，还有其他节点可以备份；甚至，挂掉某一个机架，也没关系；其他机架上，也有备份。

读操作：



读操作就简单一些了，如图所示，client要从datanode上，读取FileA。而FileA由block1和block2组成。

那么，读操作流程为：

a. client向namenode发送读请求。

b. namenode查看Metadata信息，返回fileA的block的位置。

block1:host2,host1,host3

block2:host7,host8,host4

c. block的位置是有先后顺序的，先读block1，再读block2。而且block1去host2上读取；然后block2，去host7上读取；

上面例子中，client位于机架外，那么如果client位于机架内某个DataNode上，例如,client是host6。那么读取的时候，遵循的规律是：

优选读取本机架上的数据。

HDFS中常用到的命令

1、hadoop fs

```
+ View Code
```

2、hadoop fsadmin

```
+ View Code
```

3、hadoop fsck

4、start-balancer.sh

分类: Hadoop

大牛笔记
关注 - 23
粉丝 - 268
+加关注

« 上一篇：【Hadoop】用web查看hadoop运行状态
» 下一篇：【JAVA】配置JAVA环境变量，安装Eclipse

posted @ 2013-11-26 16:57 大牛笔记 阅读(127406) 评论(9) 编辑 收藏

评论列表

#1楼 2013-11-26 19:54 fo0ol

新手学习hadoop最好是在linux平台吧，是不是一般都用centos的版本

支持(1) 反对(0)

#2楼[楼主] 2013-11-26 21:12 大牛笔记

@ fo0ol
hadoop运行平台可以用linux虚拟机运行。mapreduce可以在win7下编写。
linux版本，可以是Ubuntu系统，也可以是centos系统。

支持(0) 反对(0)

#3楼 2013-12-13 11:39 拜门求学

大神，向你致敬

支持(0) 反对(0)

#4楼 2015-02-10 09:44 bossdk

拜读！ 请问楼主用什么软件画的图?

支持(0) 反对(0)

#5楼[楼主] 2015-02-10 14:24 大牛笔记

@ bossdk
一个挺不错的web端的工具，强烈推荐
http://www.processon.com/invitation/526f30c10cf22f64f6308bbf
（此url包含我的邀请链接，不会对你产生影响。）
此工具可以画uml图，流程图，还有简单的界面ui设计等。

支持(1) 反对(0)

#6楼 2016-03-24 11:01 rocky_24

这个图画的。。。。 有点别扭。 怎么还用win7 做client。。。。 不用linux

支持(1) 反对(0)

#7楼 2016-08-26 19:49 杉枫

画图工具收藏了

支持(0) 反对(0)

#8楼 2016-10-26 18:15 J123456

牛牲犇掰

支持(0) 反对(0)

#9楼 2017-08-01 21:11 Try_It

大牛，学习了，讲的很透彻明白，你是我看了这么多头一个写这么明白的，膜拜，

支持(2) 反对(0)

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

相关博文：
· HDFS运行原理
· Hadoop之HDFS及NameNode单点故障解决方案
· HDFS 详解以及运行原理
· hadoop之hdfs学习
· 一、Hadoop初识

最新新闻：
· 你每个月10GB流量，6GB都被广告偷走了！
· 高通骁龙855发布了，明年Android手机会有什么变化？
· .NET Core 2.2正式发布，有你喜欢的特性吗？
· 科学家第一次测量宇宙全部星光，证实令人不安的结论
· 猎豹应用的小烦恼，出海绕不过去的本地化
» 更多新闻...