7-hdfs存储原理

----------成都尚学堂-mr-zeng----------

HDFS概念

（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）。

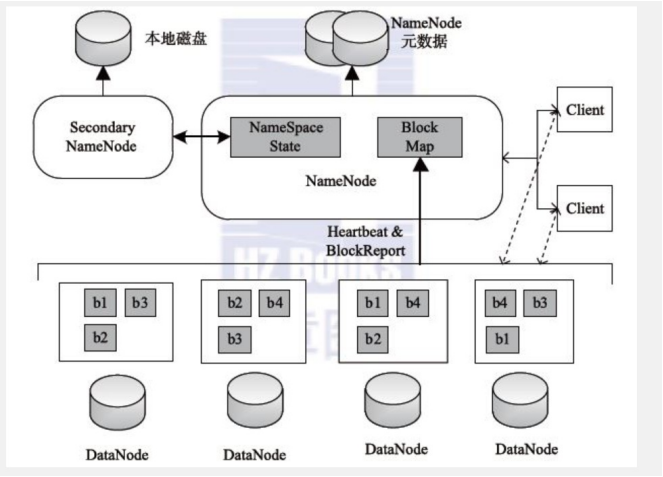
HDFS特点

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，v1默认64M，v2默认128m，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。

HDFS存储数据的架构



如上图所示，HDFS也是按照Master（主）和Slave（从）的结构。

分NameNode、SecondaryNameNode、DataNode这几个角色。

表示：一个大文件分割为b1，b2，b3，b4；4个block块。

NameNode：存储元信息，提供hdfs访问服务入口：文件名，大小，块位置等

SecondaryNameNode：存储元信息备份

DataNode：存储块（数据）信息

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。

冷备份（hadoop1提供冷备份SecondaryNameNode）：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。

热备份（hadoop2提供热备份配合zookeeper）：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。

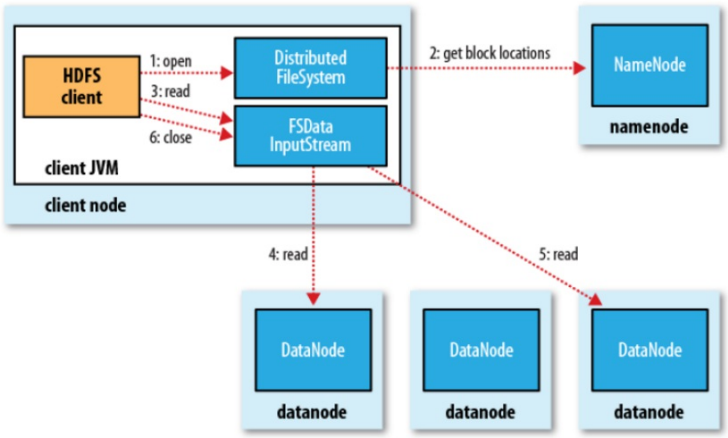
fsimage:元数据镜像文件（文件系统的目录树。）

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。

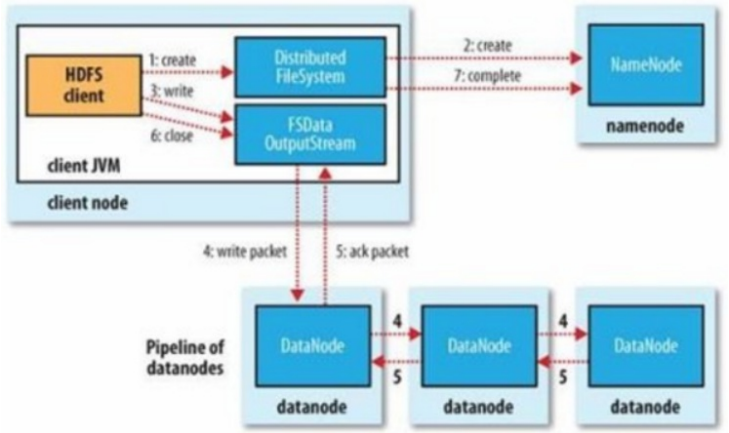
HDFS读取数据流程



在读取的时候，如果client与datanode通信时遇到一个错误，那么它就会去尝试对这个块来说下一个最近的块。它也会记住那个故障节点的datanode，以保证不会再对之后的块进行徒劳无益的尝试。client也会确认datanode发来的数据的校验和。如果发现一个损坏的块，它就会在client试图从别的datanode中读取一个块的副本之前报告给namenode。

这个设计的一个重点是，client直接联系datanode去检索数据，并被namenode指引到块中最好的datanode。因为数据流在此集群中是在所有datanode分散进行的。所以这种设计能使HDFS可扩展到最大的并发client数量。同时，namenode只不过提供块的位置请求（存储在内存中，十分高效），不是提供数据。否则如果客户端数量增长，namenode就会快速成为一个“瓶颈”。

HDFS写入数据流程



复本的布局：需要对可靠性、写入带宽和读取带宽进行权衡。Hadoop的默认布局策略是在运行客户端的节点上放第1个复本（如果客户端运行在集群之外，就随机选择一个节点，不过系统会避免挑选那些存储太满或太忙的节点。）第2个复本放在与第1个复本不同且随机另外选择的机架的节点上（离架）。第3个复本与第2个复本放在相同的机架，且随机选择另一个节点。其他复本放在集群中随机的节点上，不过系统会尽量避免相同的机架放太多复本。
总的来说，这一方法不仅提供了很好的稳定性（数据块存储在两个机架中）并实现很好的负载均衡，包括写入带宽（写入操作只需要遍历一个交换机）、读取性能（可以从两个机架中选择读取）和集群中块的均匀分布（客户端只在本地机架上写入一个块）。