28-HDFS

定义与特点

Hadoop分布式文件系统(HDFS)是指被设计成适合运行在通用硬件(commodity hardware)上的分布式文件系统（Distributed File System）。

HDFS在LinuxFS系统的基础上自己搭建了一个可以存储大量数据的结构

适合存储大文件，建立在认为“write-once, read many-times”是最高效的数据读写方式的基础上，对硬件的要求不高。

支持用户配额和访问权限，不支持链接和软链接

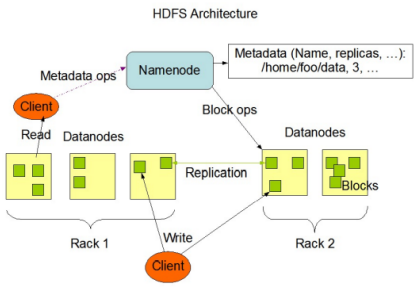
但是HDFS不太适用于以下场景：

1、低延迟数据访问：HDFS是为提供高吞吐量的数据而优化的，这可能导致时延较大

2、有许多小文件：由于namenode在内存中保存文件系统元数据，因此文件系统中文件数量受namenode内存量的限制

3、多个写入，任意文件修改：HDFS中的文件可以由单个写入程序写入。写入总是在文件末尾进行。不支持多个写入程序，也不支持在文件中以任意偏移量进行修改。

NameNode&DataNodes



主从结构

HDFS exposes a file system namespace and allows user data to be stored in files.
- Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes.
- The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes.
- The DataNodes are responsible for serving read and write requests from the file system's clients.
- The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

The existence of a single NameNode in a cluster greatly simplifies the architecture of the system.
- The NameNode is the arbitrator and repository for all HDFS metadata.
- The system is designed in such a way that user data never flows through the NameNode.

- **Blocks**：block size是硬盘一次读写的最小数据大小

  HDFS把数据存在blocks里面，把一个block当作一个文件放在LinuxFS里

- **NameNode**可以告诉你要找一个数据去哪台机器上找，当**NameNode**告诉你在哪里后你就直接和DataNode交流

    - DataNode不知道实际数据的全貌，会给NameNode发**心跳和block report**；namenode在分发请求的时候就不会往崩了的数据发
    - 不能用单独一个进程或者线程来同Timer发心跳，不能保证运行程序的进程或者线程也是正常的，要用守护进程调用循环来发心跳。
- 由于存的数据可能会很大，datanode对存储的结构进行优化（存成树形结构）
- NameNode&DataNodes难说谁是从谁是主，就近读取

持久化

Namenode保存了HDFS的namespace：将所有文件系统元数据的变化写到EditLLog以实现持久化，FsImage存了整个文件系统的namespace（包括各个block映射到的文件、文件系统的属性），FsImage文件存在本地文件系统里。

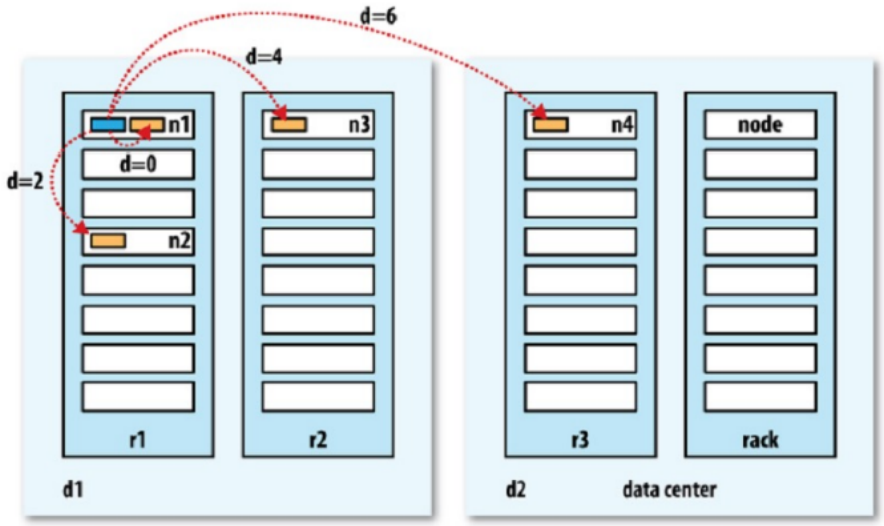
EditLog不能存在HDFS里面，存在本地文件系统里，FsImage也是。先写EditLog再写FsImage。

一旦写成功，就把旧的数据写到FsImage后清空，相当于一个checkpoint。启动的时候有可能会出现FsImage和EditLog不一致的情况，（EditLog是准确的），重新执行一遍EditLog里的命令并保持一致

数据副本

HDFS不同replica实现不同等级的备份，要保证不同请求尽量平均地分发给不同的机器

作用：让用户在读数据时可以访问离自己最近的数据副本所在机器，可以最大限度地减少整体带宽消耗和读取延迟；如果HDFS群集跨越多个数据中心，则驻留在本地数据中心的副本优于任何远程副本。



Datanode挂了：NameNode支持不同文件在LinuxFS里面副本数量不同，但是同一个文件的副本不能同时出现在一个机器里。如果datanode挂了，里面存的副本丢失，就会导致对应的文件的副本数不符合设定，namenode就需要重新复制这个文件到其它datanode，另外，namenode在分发请求的时候也不会往这datanode发了

安全模式 (Safemode)

Namenode在启动时先进入安全模式，直接去获取每个datanode的基本数据，这段时间里不会发生data blocks复制，当一部分configuration符合的时候就可以开始接受请求了

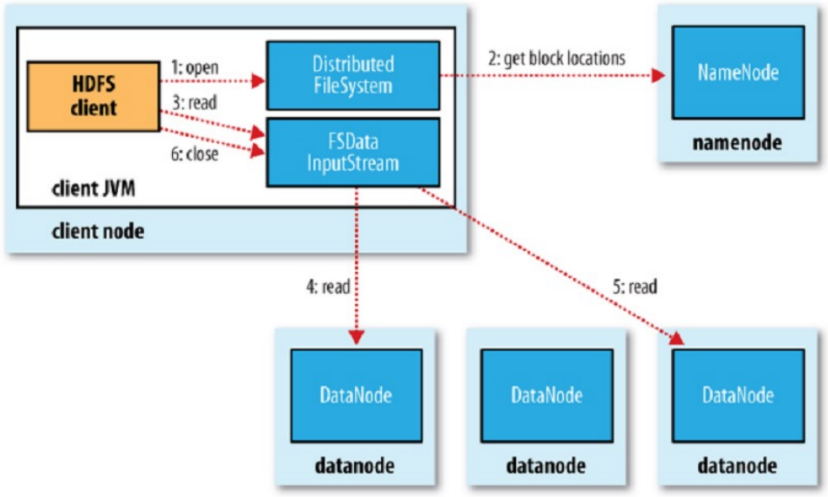
鲁棒性

HDFS的主要目标是即使在出现故障时也能可靠地存储数据。三种常见的故障类型是NameNode故障、DataNode故障和网络分区。

- Cluster Rebalancing：HDFS体系结构与数据再平衡的思想兼容
- Data Integrity：从DataNode获取的数据块可能已损坏。HDFS客户端软件会对HDFS文件的内容执行校验和检查。当客户端创建HDFS文件时，它会计算文件每个块的校验和，并将这些校验和存储在同一HDFS命名空间中的单独隐藏文件中。当客户端检索文件内容时，它会验证从每个DataNode接收的数据是否与存储在关联校验和文件中的校验和匹配。如果没有，则客户端可以选择从具有该块副本的另一个DataNode检索
- Metadata Disk Failure：FsImage和EditLog是HDFS的中心数据结构

- Snapshots：快照支持在特定时刻存储数据副本。快照功能的一个用途可能是将损坏的HDFS实例回滚到以前已知的好的版本。

读数据：



写数据：