HDFS 初识

1. 什么是 HDFS？

HDFS 是 Hadoop Distributed File System（分布式文件系统）的简称，是 hadoop 项目的一部分。HDFS 有着高容错性的特点，并且设计用来部署在低廉的硬件上，以流的形式访问文件系统中的数据。它可以提供高吞吐量访问应用程序的数据的能力，适合那些有着超大数据集的应用程序。它所具有的高容错、高可靠性、高可扩展性、高吞吐率等特征为海量数据提供了不怕故障的存储，为超大数据集的应用处理带来了很多便利。

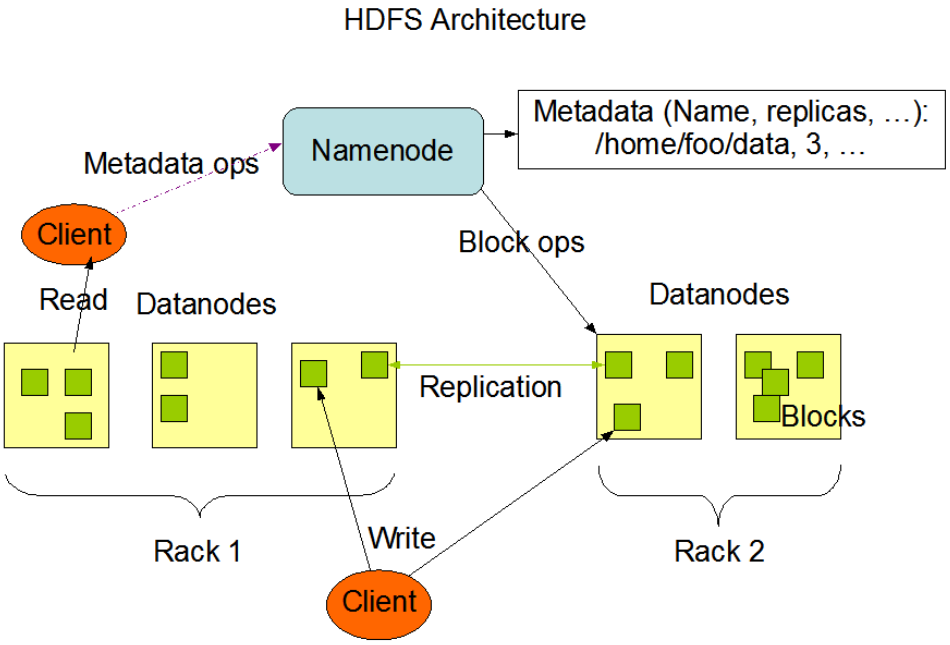
运行在 HDFS 之上的程序有很大量的数据集。典型的 HDFS 文件大小是 GB 到 TB 的级别。所以，HDFS 被调整成支持大文件。它应该提供很高的聚合数据带宽，一个集群中支持数百个节点，一个集群中还应该支持千万级别的文件。

2. HDFS 结构

HDFS 是一个的主从结构，一个 HDFS 集群是由一个名字节点，它是一个管理文件命名空间和调节客户端访问文件的主服务器，当然还有一些数据节点，通常是一个节点一个机器，它来管理对应节点的存储。HDFS 对外开放文件命名空间并允许用户数据以文件形式存储。

内部机制是将一个文件分割成一个或多个块，这些块被存储在一组数据节点中。名字节点用来操作文件命名空间的文件或目录操作，如打开，关闭，重命名等等。它同时确定块与数据节点的映射。数据节点来负责来自文件系统客户的读

写请求。数据节点同时还要执行块的创建，删除，和来自名字节点的块复制指令。



名字节点和数据节点都是运行在普通的机器之上的软件，机器典型的都是 GNU/Linux，HDFS 是用 java 编写的，任何支持 java 的机器都可以运行名字节点或数据节点，利用 java 语言的超轻便型，很容易将 HDFS 部署到大范围的机器上。典型的部署是由一个专门的机器来运行名字节点软件，集群中的其他每台机器运行一个数据节点实例。体系结构不排斥在一个机器上运行多个数据节点的实例，但是实际的部署不会有这种情况。

集群中只有一个名字节点极大地简单化了系统的体系结构。名字节点是仲裁者和所有 HDFS 元数据的仓库，用户的实际数据不经过名字节点。

HDFS 的这种结构简单、轻便，可以快速地扩展新的数据节点，从而扩充整个系统的存储能力。但是，由于其显而易见的弱点——单节点的名字节点，使其存在该节点失效时存在整个系统不可使用的风险，若能增加一个名字节点作为热备，则可大大提升 HDFS 的系统可用性，若一个节点的故障率为 0.1，则两个节点的故障率可降低到 0.01，提升的效果会极其明显，进一步提升系统的可靠

性。

3. 高容错性保障

3.1. 心跳

心跳是用来保持 HDFS 数据节点与名字节点联系的重要方式，数据节点不断地向名字节点发送数据包，表明自己还是活动的，从而达到在名字节点需要数据时，向该数据节点发送读取请求。

名字节点负责处理所有的块复制相关的决策。它周期性地接受集群中数据节点的心跳和块报告。一个心跳的到达表示这个数据节点是正常的。一个块报告包括该数据节点上所有块的列表。

3.2. 复制因子

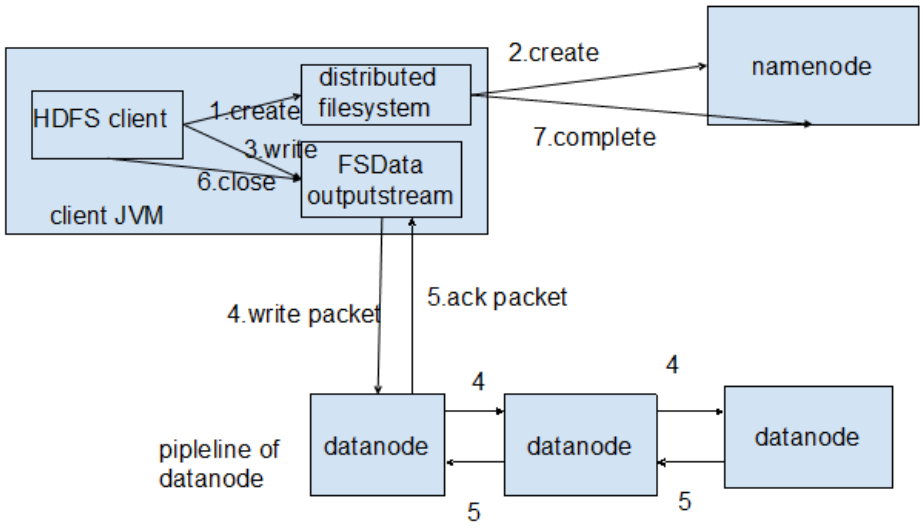
复制因子是保证 HDFS 数据可靠性的重要特征，简单地可以理解为数据块被复制的次数，通常为 3，同时，系统会按照一定的策略将各个数据块副本存储在不同的位置。一个合适的复制因子直接影响了 HDFS 的整体系统可靠性和存取效率。

HDFS 设计成能可靠地在集群中大量机器之间存储大量的文件，它以块序列的形式存储文件。文件中除了最后一个块，其他块都有相同的大小。属于文件的块为了故障容错而被复制。块的大小和复制数是以文件为单位进行配置的，应用可以在文件创建时或者之后修改复制因子。HDFS 中的文件是一次写的，并且任何时候都只有一个写操作。

块副本存放位置的选择严重影响 HDFS 的可靠性和性能。副本存放位置的优化是 HDFS 区分于其他分布式文件系统的的特征，这需要精心的调节和大量

的经验。机架敏感的副本存放策略是为了提高数据的可靠性，可用性和网络带宽的利用率。目前副本存放策略的实现是这个方向上比较原始的方式。短期的实现目标是要把这个策略放在生产环境下验证，了解更多它的行为，为以后测试研究更精致的策略打好基础。

HDFS 运行在跨越大量机架的集群之上。两个不同机架上的节点是通过交换机实现通信的，在大多数情况下，相同机架上机器间的网络带宽优于在不同机架上的机器。



为了尽量减小全局的带宽消耗读延迟，HDFS 尝试返回给一个读操作离它最近的副本。假如在读节点的同一个机架上就有这个副本，就直接读这个，如果 HDFS 集群是跨越多个数据中心，那么本地数据中心的副本优先于远程的副本。

一个数据节点周期性发送一个心跳包到名字节点。网络断开会造成一组数据节点子集和名字节点失去联系。名字节点根据缺失的心跳信息判断故障情况。名字节点将这些数据节点标记为死亡状态，不再将新的 IO 请求转发到这些数据节点上，这些数据节点上的数据将对 HDFS 不再可用，可能会导致一些块的复制

因子降低到指定的值。

名字节点检查所有的需要复制的块，并开始复制他们到其他的数据节点上。重新复制在有些情况下是不可或缺的，例如：数据节点失效，副本损坏，数据节点磁盘损坏或者文件的复制因子增大。

3.3. 安全模式

在启动的时候，名字节点进入一个叫做安全模式的特殊状态。安全模式中不允许发生文件块的复制。名字节点接受来自数据节点的心跳和块报告。一个块报告包含数据节点所拥有的数据块的列表。

每一个块有一个特定的最小复制数。当名字节点检查这个块已经大于最小的复制数就被认为是安全地复制了，当达到配置的块安全复制比例时（加上额外的 30 秒），名字节点就退出安全模式。它将检测数据块的列表，将小于特定复制数的块复制到其他的数据节点。

3.4. 数据块校验

从数据节点上取一个文件块有可能是坏块，坏块的出现可能是存储设备错误，网络错误或者软件的漏洞。HDFS 客户端实现了 HDFS 文件内容的校验。当一个客户端创建一个 HDFS 文件时，它会为每一个文件块计算一个校验码并将校验码存储在同一个 HDFS 命名空间下一个单独的隐藏文件中。当客户端访问这个文件时，它根据对应的校验文件来验证从数据节点接收到的数据。如果校验失败，客户端可以选择从其他拥有该块副本的数据节点获取这个块。

4. 小结

HDFS 被设计为一次写多次读的模式，它简单化了数据一致的问题和并使高

吞吐量的数据访问变得可能，因此，不适合作为日常业务数据系统的存储，它更适合作为归档类数据系统、大数据系统分析的原数据存储。

由于 HDFS 是设计用于大吞吐量数据的，这是以一定延时为代价的，它不太适合于那些要求低延时（数十毫秒）访问的应用程序。HDFS 是单 Master 的，所有的对文件的请求都要经过它，当请求多时，肯定会有延时。对于那些有低延时要求的应用程序，HBase 更适合。