8、HDFS 的文件读取过程

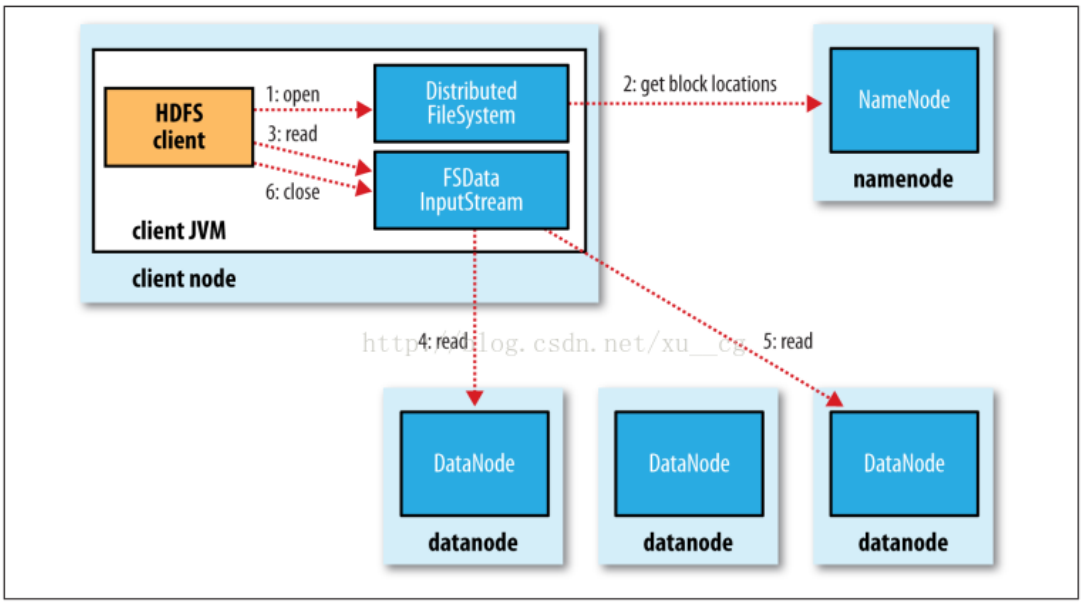

Figure 3-2. A client reading data from HDFS

详细步骤解析

1、 Client 向 NameNode 发起 RPC 请求，来确定请求文件 block 所在的位置；

2、NameNode 会视情况返回文件的部分或者全部 block 列表，对于每个 block，NameNode 都会返回含有该 block 副本的 DataNode 地址； 这些返回的 DN 地址，会按照集群拓扑结构得出 DataNode 与客户端的距离，然后进行排序，排序两个规则：网络拓扑结构中距离 Client 近的排靠前；心跳机制中超时汇报的 DN 状态为 STALE，这样的排靠后；

3、 Client 选取排序靠前的 DataNode 来读取 block，如果客户端本身就是 DataNode,那么将从本地直接获取数据(短路读取特性)；

4、 底层上本质是建立 Socket Stream（FSDataInputStream），重复的调用父类 DataInputStream 的 read 方法，直到这个块上的数据读取完毕；

5、 当读完列表的 block 后，若文件读取还没有结束，客户端会继续向 NameNode 获取下一批的 block 列表；

6、 读取完一个 block 都会进行 checksum 验证，如果读取 DataNode 时出现错误，客户端会通知 NameNode，然后再从下一个拥有该 block 副本的 DataNode 继续读。

7、 read 方法是并行的读取 block 信息，不是一块一块的读取；NameNode 只是返回 Client 请求包含块的 DataNode 地址，并不是返回请求块的数据；

8、 最终读取来所有的 block 会合并成一个完整的最终文件。