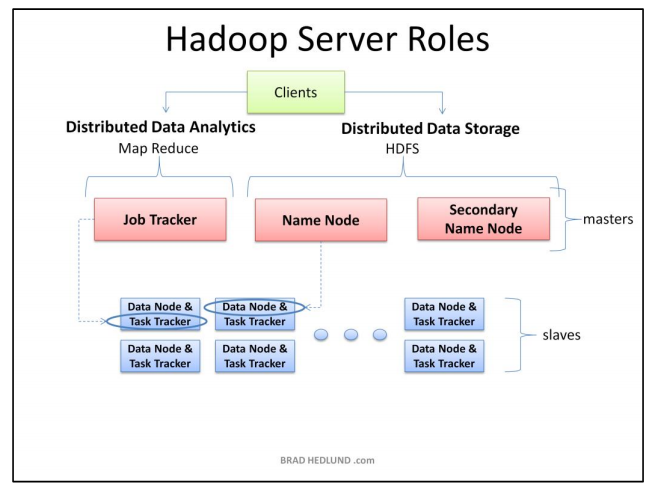
深入理解hadoop 集群和网络  
原文 [http://bradhedlund.com/2011/09/1 ... rs-and-the-network/](http://bradhedlund.com/2011/09/10/understanding-hadoop-clusters-and-the-network/)



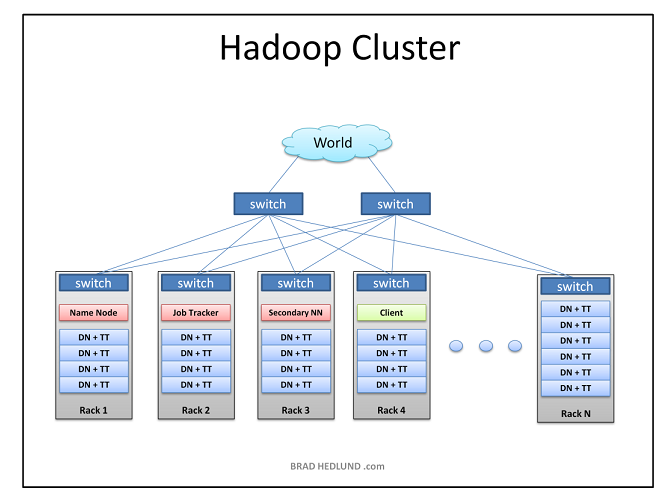
通过上图可以，hadoop架构中主要由三种角色组成：client machine（用户机）、master node（主控节点）、salve node（从节点）。

Client machine的主要职责：向集群中传数据，提交mapreduce任务和接收任务处理结果。

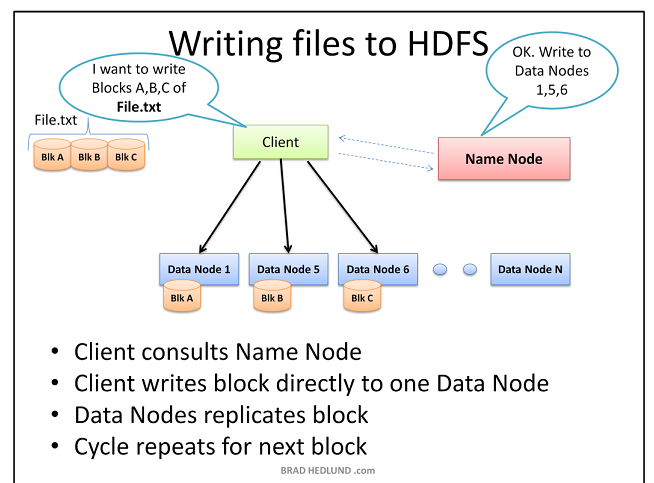
Master node的主要职责：负责HDFS系统中元数据的管控和处理mapreduce任务（JobTracker负责）

Salve node的主要职责：负责实际的计算任务

在小型的系统集群中，可能会用一个物理机器来运行多个服务器职能，但是在大一点的则要单独架设物理机器，做到专职专责。



上图就是一个典型的hadoop集群的拓扑图。在同一个rack（机架）里的datanode传输素的要快于不同rack的速度。每隔rack都有一个switch（交换机），这些交换机和上级交换机互联，构成了hadoop的集群拓扑。注意slave nodes通常有更大的存储空间和较小的DRAM以及CPU，而Master node则相反。这是根据不同任务的需求所做的合理配置。



向HDFS文件系统写文件流程：

①　客户端告知namenode他要写文件

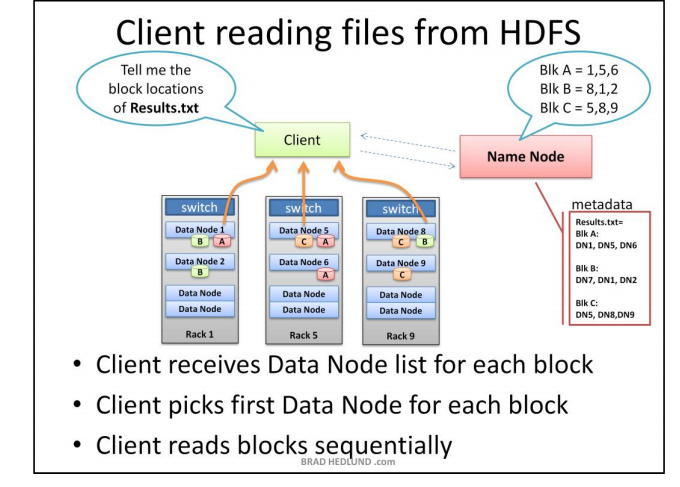
②　Namenode回复客户端写到datanode1,5,6中

③　客户端把文件组成之一A块写到datanode1，5,6中

④　接下来把剩下的B,C块写到datanode1,5,6中

组成文件的每个块要复制多少份可以通过hdfs-site.xml中的dfs.replication

读数据拓扑图：



读数据的步骤：

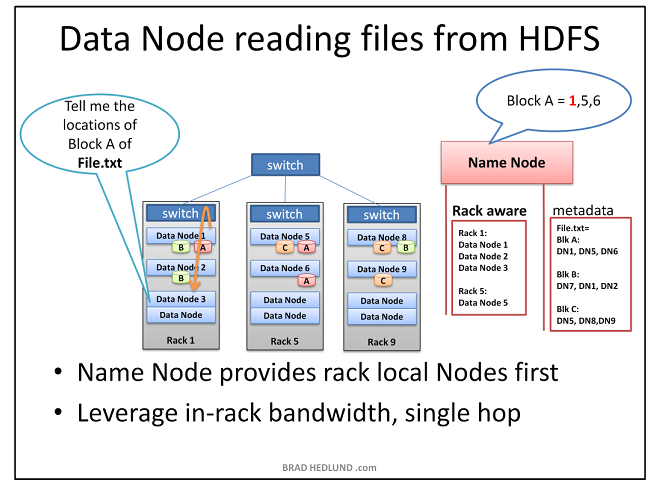
①　客户端向namenode询问block列表

②　Namenode提供给客户端一份关于每个block的列表

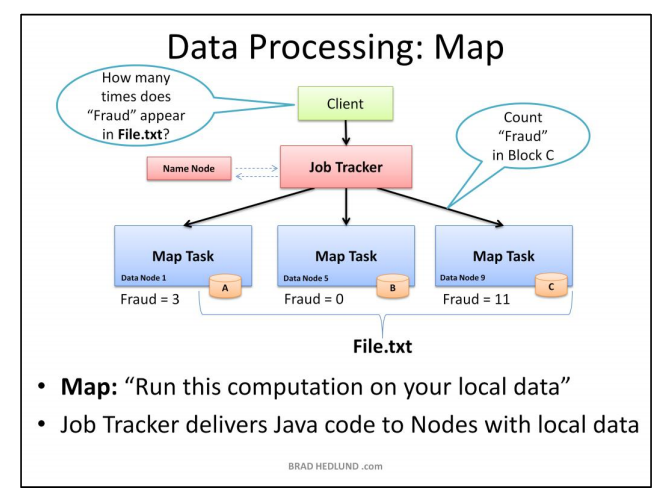
③　客户端选择在每份block列表的第一个datanode

④　块被顺序读取

⑤　如果当前block没有被读取完就不会进行下一次block读取



在某些情况下，datanode本身需要处理自身不具有的block，那么在这种情况下，他就会想namenode询问block的地址。在上图中，namenode中存储了我们需要的block A，分别位于datanode1，datanode5，datanode6，但是显而易见如果datanode3需要取block A，那么从位于同一个机架中的datanode1取无疑更加迅速（因为他们在同一个机架里）。所以namenode就会把datanode1中存储的block地址告知datanode3.这是机架感知的一个案例



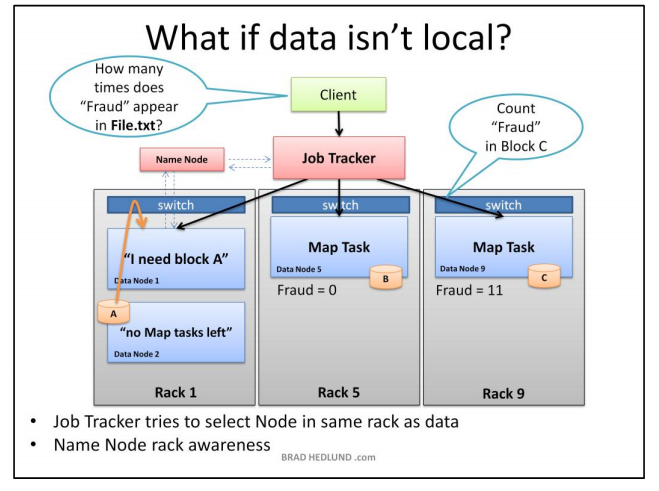
以上的图讲解了map reduce的过程：

（一）首先由客户端向Job Tracker提交一个计算Fraud在File.txt中出现了多少次的请求。

（二）Job Tracker向namenode询问在那些datanode中存储着Fraud文件的块

（三）Job Tracker告知Task Tracker这些块的信息，然后由Task Tracker来执行map计算

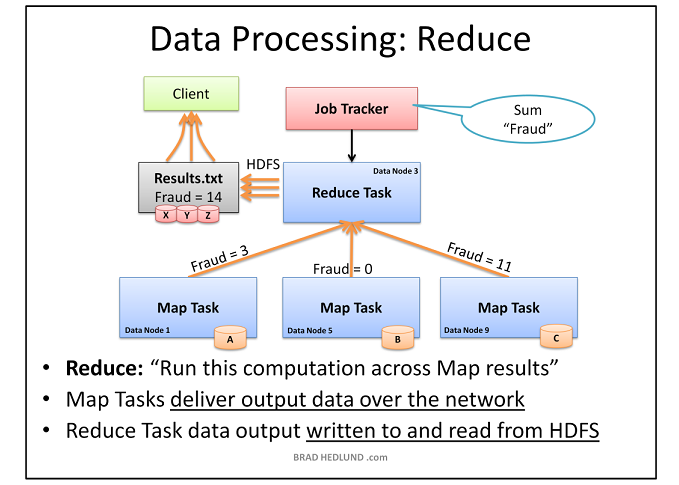
（四）最后把map计算的结果汇总到Reduce进程中去得出结果



如果datanode上已经没有运行map的空间时候的处理办法：

①　Job Tracker会向namenode通过机架感知机制向namenode咨询同一机架里哪一个节点可以执行map任务

②　找到这样的节点后把他原来datanode的数据拷贝到新datanode里面去



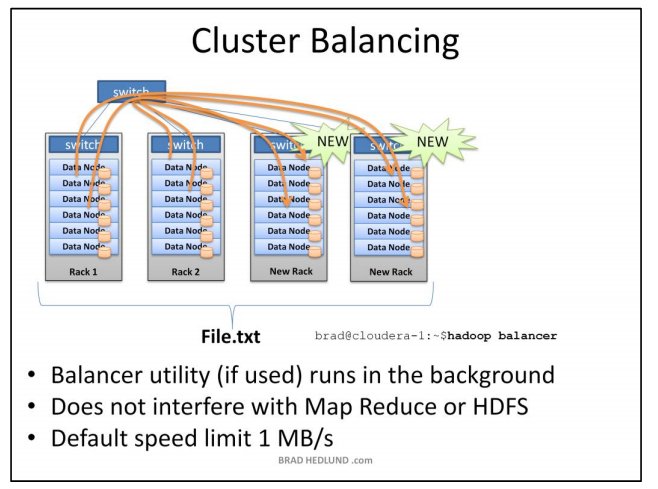
Reduce工作流程：

1) Job Tracker进程会在集群中的任意一个datanode中开启Reduce进程

2) Reduce进程所有map进程生成的中间结果，这些结果通过network被汇总到reduce节点

3) Reduce节点会把结果汇总为一个最终文件，这些文件会被复制到HDFS系统中

4) 返回结果给用户



①　Balancer是一个很有用的工具，他是用来做负载均衡的

②　Balancer这个工具除非你手工启动，否则他是不会自动运行的

③　有一个阀值用来确定什么时候开始做负载均衡

④　Balancer可以用的网络带宽可以通过Hdfs-site.xml中的dfs.balance.bandwidthPerSec参数来配置，默认情况下为1MB/s

[图片3.png](http://f.dataguru.cn/forum.php?mod=attachment&aid=NDg1OTZ8YTgyNWQ0OGV8MTM1MDgwODkzMXwwfDI0NDU1&nothumb=yes)*(77.37 KB, 下载次数: 0)*

