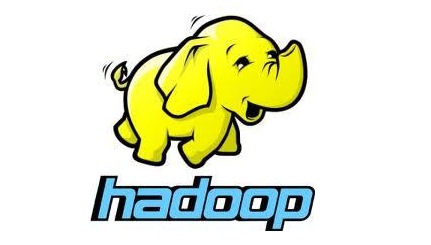
传说中的Hadoop，我终于来对着你唱“征服”了，好可爱的小象！☺

总的来说，hadoop的思路比较简单(map-reduce)，就是将任务分开进行，最后汇总。但这个思路实现起来，比较复杂，但相对于Intel等硬件公司提出的网格运算等方式，显得更加开放。

Tip：实践应用是核心



提到列式(Column Family)数据库，就不得不提Google的BigTable，其开源版本就是我们熟知的HBASE。BigTable建立在谷歌的另两个系统GFS和Chubby之上，这三个系统和分布式计算编程模型MapReduce共同构成Google云计算的基础，Chubby解决主从自动切换的基础。接下来通过一个表格对比来引入Hadoop。

|  |  |
| --- | --- |
| Google云计算 | Hadoop中的对应 |
| 分布式文件系统GFS | HDFS |
| 分布式所服务Chubby | Zookeeper |
| 分布式计算框架MapReduce | MapReduce |
| 分布式数据库BigTable | HBase |

Hadoop是有Apache Lucene的作者Boug Cutting开发的，其主体结构如下图所示。



* HDFS

NameNode:整个文件系统的大脑，提供整个系统的目录信息并管理各个数据服务器。

DataNode:分布式文件系统中每一个文件被切割为若干数据块，每个数据块存储在不同服务器，这些就是数据服务器。

Block:每个被切分的数据块就是一段文件内容，其是基本的存储单位，被称为数据块，典型大小为64MB。

Tip:由于硬件错误是常态，HDFS是很多台Server的集合，因而错误检测和恢复是核心功能；其以流式读为主，做批量操作，关注数据访问的高吞吐量。

HDFS采用**master/slave架构**，一个HDFS集群由**一个NameNode和若干DataNode组成**，中心服务器NameNode负责管理文件系统的namespace和客户端对文件的访问。DataNode一般一个节点一个，负责管理节点上附带的存储。在内部，一个文件被分成一个或多个block，这些block存储在DataNode集合中。NameNode和DataNode均可运行在廉价的linux机器上，HDFS由java语言开发，跨平台好，总体结构示意图如下所示。



复制：采用rack-aware策略改进数据可靠性和网络带宽的利用；NameNode决定每个DataNode的Rack id；大多数情况，**replication因子是3**，简单来说就是将一个副本放在本地机架节点，一个副本放在同一机架另一个节点，最后一个放在不同机架；在读取时，会选择最近的副本；NameNode启动时会进入SafeMode状态，该状态时，NameNode不会进行数据块的复制，这是会检测DataNode的副本数量，如果满足要求则认为安全。

NameNode用于存储元数据，任何修改均被Editlog记录，通讯协议基于TCP/IP，可以通过java API调用。

* MapReduce模型
* Zookeeper
* HBase

参考资料：

1. 皮雄军. NoSQL数据库技术实战[M]. 北京:清华大学出版社, 2015.