Hadoop

[Hadoop](http://baike.baidu.com/view/908354.htm)是一个由Apache基金会所开发的[分布式系统](http://baike.baidu.com/view/991489.htm" \t "_blank)基础架构。

用户可以在不了解分布式底层细节的情况下，开发分布式程序。充分利用集群的威力进行高速运算和存储。

[1]  Hadoop实现了一个[分布式文件系统](http://baike.baidu.com/view/771589.htm" \t "_blank)（Hadoop Distributed File System），简称HDFS。HDFS有高[容错性](http://baike.baidu.com/view/2700299.htm" \t "_blank)的特点，并且设计用来部署在低廉的（low-cost）硬件上；而且它提供高吞吐量（high throughput）来访问[应用程序](http://baike.baidu.com/view/330120.htm" \t "_blank)的数据，适合那些有着超大数据集（large data set）的应用程序。HDFS放宽了（relax）POSIX的要求，可以以流的形式访问（streaming access）文件系统中的数据。

Hadoop的框架最核心的设计就是：HDFS和MapReduce。HDFS为海量的数据提供了存储，则MapReduce为海量的数据提供了计算。

特点：

Hadoop是一个能够让用户轻松架构和使用的[分布式计算](http://baike.baidu.com/subview/30655/11213586.htm" \t "_blank)平台。用户可以轻松地在Hadoop上开发和运行处理海量数据的[应用程序](http://baike.baidu.com/view/330120.htm" \t "_blank)。它主要有以下几个优点：

1. **高可靠性**。Hadoop按位存储和处理数据的能力值得人们信赖。
2. **高扩展性**。Hadoop是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。
3. **高效性**。Hadoop能够在节点之间动态地移动数据，并保证各个节点的[动态平衡](http://baike.baidu.com/view/1603996.htm" \t "_blank)，因此处理速度非常快。
4. **高容错性**。Hadoop能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。
5. **低成本**。与一体机、商用数据仓库以及QlikView、Yonghong Z-Suite等数据集市相比，hadoop是开源的，项目的软件成本因此会大大降低。

Hadoop带有用[Java](http://baike.baidu.com/view/29.htm" \t "_blank)语言编写的框架，因此运行在 Linux 生产平台上是非常理想的。Hadoop 上的[应用程序](http://baike.baidu.com/view/330120.htm" \t "_blank)也可以使用其他语言编写，比如 [C++](http://baike.baidu.com/view/824.htm)。

Hadoop Distributed File System（HDFS）：

对外部客户机而言，[HDFS](http://baike.baidu.com/view/3061630.htm" \t "_blank)就像一个传统的分级文件系统。可以创建、删除、移动或[重命名](http://baike.baidu.com/view/1175233.htm" \t "_blank)文件，等等。但是 HDFS 的架构是基于一组特定的节点构建的（参见图 1），这是由它自身的特点决定的。这些节点包括 NameNode（仅一个），它在 HDFS 内部提供元数据服务；DataNode，它为 HDFS 提供存储块。由于仅存在一个 NameNode，因此这是 HDFS 的一个缺点（单点失败）。

存储在 HDFS 中的文件被分成块，然后将这些块复制到多个计算机中（DataNode）。这与传统的 RAID 架构大不相同。块的大小（通常为 64MB）和复制的块数量在创建文件时由客户机决定。NameNode 可以控制所有文件操作。HDFS 内部的所有通信都基于标准的 [TCP/IP](http://baike.baidu.com/view/7729.htm) 协议。