# 学习心得

1813003 李昱龙

HDFS

HDFS是Hadoop分布式文件系统（Hadoop Distributed File System）的缩写，为分布式计算存储提供了底层支持。采用Java语言开发，可以部署在多种普通的廉价机器上，以集群处理数量积达到大型主机处理性能。

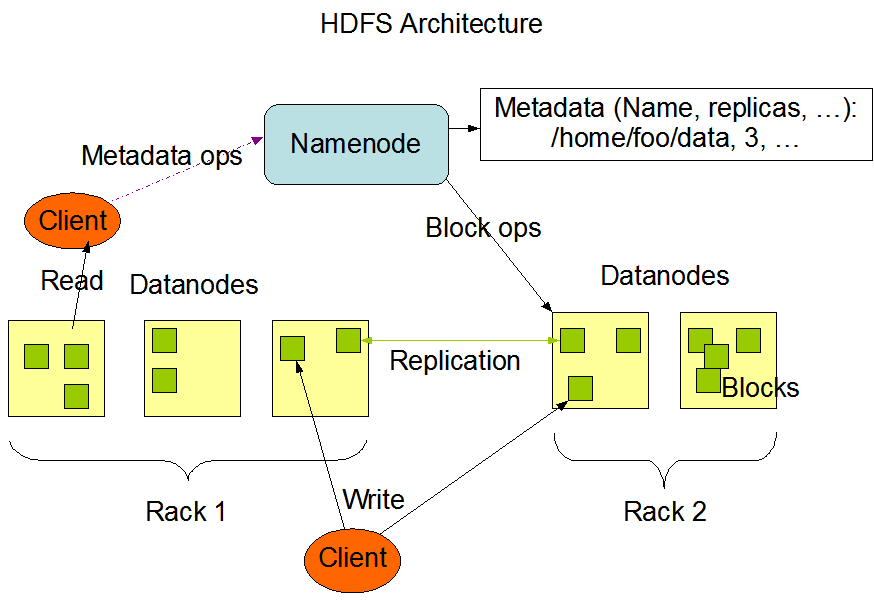
HDFS 架构原理

HDFS采用master/slave架构。一个HDFS集群包含一个单独的NameNode和多个DataNode。

NameNode作为master服务，它负责管理文件系统的命名空间和客户端对文件的访问。NameNode会保存文件系统的具体信息，包括文件信息、文件被分割成具体block块的信息、以及每一个block块归属的DataNode的信息。对于整个集群来说，HDFS通过NameNode对用户提供了一个单一的命名空间。

DataNode作为slave服务，在集群中可以存在多个。通常每一个DataNode都对应于一个物理节点。DataNode负责管理节点上它们拥有的存储，它将存储划分为多个block块，管理block块信息，同时周期性的将其所有的block块信息发送给NameNode。

下图为HDFS系统架构图，主要有三个角色，Client、NameNode、DataNode。



Hadoop 处理流程

在描述Hadoop处理流程之前，先提一个分布式计算最为重要的设计原则：Moving Computation is Cheaper than Moving Data。意思是指在分布式计算中，移动计算的代价总是低于移动数据的代价。本地计算使用本地数据，然后汇总才能保证分布式计算的高效性。

下图所示Hadoop处理流程：

