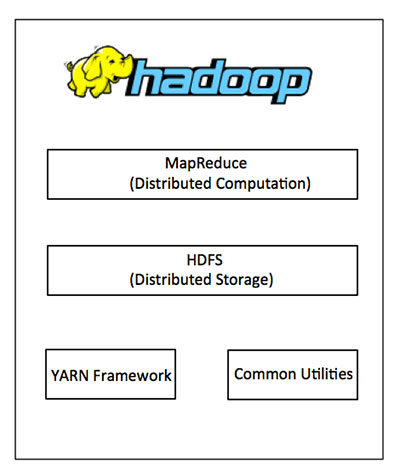
# Hadoop是什么？

Hadoop是使用Java编写，允许分布在集群，使用简单的编程模型的计算机大型数据集处理的Apache的开源框架。 Hadoop框架应用工程提供跨计算机集群的分布式存储和计算的环境。 Hadoop是专为从单一服务器到上千台机器扩展，每个机器都可以提供本地计算和存储。

## **一、Hadoop的架构**

在其核心，Hadoop主要有两个层次，即：

* 加工/计算层(MapReduce)，以及存储层(Hadoop分布式文件系统)。



## 二、MapReduce

MapReduce是一种并行编程模型，用于编写普通硬件的设计，谷歌对大量数据的高效处理(多TB数据集)的分布式应用在大型集群(数千个节点)以及可靠的容错方式。 MapReduce程序可在Apache的开源框架Hadoop上运行。

## 三、Hadoop分布式文件系统

Hadoop分布式文件系统（HDFS）是基于谷歌文件系统（GFS），并提供了一个设计在普通硬件上运行的分布式文件系统。它与现有的分布式文件系统有许多相似之处。来自其他分布式文件系统的差别是显著。它高度容错并设计成部署在低成本的硬件。提供了高吞吐量的应用数据访问，并且适用于具有大数据集的应用程序。

除了上面提到的两个核心组件，Hadoop的框架还包括以下两个模块：

Hadoop通用：这是Java库和其他Hadoop组件所需的实用工具。

Hadoop YARN ：这是作业调度和集群资源管理的框架。

## **四、Hadoop如何工作？**

建立重配置，处理大规模处理服务器这是相当昂贵的，但是作为替代，可以联系许多普通电脑采用单CPU在一起，作为一个单一功能的分布式系统，实际上，集群机可以平行读取数据集，并提供一个高得多的吞吐量。此外，这样便宜不到一个高端服务器价格。因此使用Hadoop跨越集群和低成本的机器上运行是一个不错不选择。

Hadoop运行整个计算机集群代码。这个过程包括以下核心任务由 Hadoop 执行：

* 数据最初分为目录和文件。文件分为128M和64M（128M最好）统一大小块。
* 然后这些文件被分布在不同的群集节点，以便进一步处理。
* HDFS，本地文件系统的顶端﹑监管处理。
* 块复制处理硬件故障。
* 检查代码已成功执行。
* 执行发生映射之间，减少阶段的排序。
* 发送排序的数据到某一计算机。
* 为每个作业编写的调试日志。

## **五、Hadoop的优势**

Hadoop框架允许用户快速地编写和测试的分布式系统。有效并在整个机器和反过来自动分配数据和工作，利用CPU内核的基本平行度。

Hadoop不依赖于硬件，以提供容错和高可用性（FTHA），而Hadoop库本身已被设计在应用层可以检测和处理故障。

服务器可以添加或从集群动态删除，Hadoop可继续不中断地运行。

Hadoop的的另一大优势在于，除了是开源的，因为它是基于Java并兼容所有的平台。