**Hadoop**

**大数据**：

包括体积庞大，高速流和可扩展的各种数据

分为三种类型：

结构化数据：关系数据

半结构化数据：xml数据

非结构化数据：Word，PDF，文本，媒体日志等

**大数据技术：**

操作大数据：像MongoDB系统，提供业务实时的能力，主要是数据捕获和存储互动工作

分析大数据：大规模并行处理（MPP）数据库系统和MapReduce提供用于回顾性和复杂的分析

MapReduce（分布式计算系统）提供分析数据的基础上，可以按比例增加从单个服务器向成千上万的 高端和

低端机的互补sql提供的功能

这两个技术是互补的，并且经常一起部署。

**大数据的挑战：**

* + 采集数据
  + 策展
  + 存储
  + 搜索
  + 分享
  + 传输
  + 分析
  + 展示

**Hadoop:**

Hadoop使用MapReduce算法运行，其中数据在使用其他并行处理的应用程序。

MapReduce是谷歌的大数据的解决方案。这个算法将任务分成小份，并将它们非配到多台计算机，

并且从这些计算机手机结果并综合，形成结果数据集

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。是专为从单一服务器到上前台机器的扩展，每个机器都可以提供本地计算和存储。

**Hadoop架构：**

其核心，主要有两个层次：

* + - 加工/计算层（MapReduce）

MapReduce是一种并行编程模型，提供一个计算模型，基于分支策略，用于编写普通硬件的设计，MapReduce程序可在Apache的开源框架Hadoop上运行

* + - 存储层（Hadoop分布式文件系统）HDFS(Hadoop distributed file system)

Hadoop的分布式文件体统（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。它高度容错并设计成部署在低成本的硬件。提供了高吞吐两的应用数据访问，并适用于具有大数据集的应用程序

除了两个核心组件，Hadoop框架还包括以下两个模块：

Hadoop通用（common utilities）：这是Java库和其他Hadoop组件所需实用工具

Hadoop YARN（YARN framework）：这是作业调度和集群资源管理的框架

Hadoop环境安装设置：

Hadoop有GNU/Linux平台支持

建议创建一个单独的用户：Hadoop，以从Unix文件系统隔离Hadoop文件系统。

使用 su 命令开启root ：su root

创建新用户： useradd username

打开刚创建的用户： su username

**Hadoop操作模式：**

操作Hadoop集群的三种模式：

本地/独立模式：默认模式。它会被配置在一个独立的模式，用于运行Java程序。

模拟分布式模式：在单台机器的分布式模拟。Hadoop守护每个进程，如HDFS,YARN…

完全分布式模式：最小两台或者多台计算机的集群。

管理界面：http://localhost:8088

NameNode界面：http://localhost:50070

HDFS NameNode界面：http://localhost:8042