**0.什么是大数据？**

大数据（big data），指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

**1.什么是Hadoop？**

Hadoop是适合大数据的分布式存储与计算平台。两个功能：存储和计算。存储：HDFS；计算：MapReduce。

**2.HDFS？**

Hadoop Distributed File System 分布式文件系统

**3.HDFS的架构**

**主从结构：**

主节点，只有一个: namenode（管理datanodes）

从节点，有很多个: datanodes

**namenode负责：**

接受用户操作请求

维护文件系统的目录结构

管理文件与block之间关系，block与datanode之间关系

**datanodes负责：**

存储文件

文件被分成block存储在磁盘上

为保证数据安全，文件会有多个副本

**4.MapReduce？**

并行计算框架

**5.MapReduce架构**

**主从结构：**

主节点，只有一个: JobTracker（管理TaskTrackers）

从节点，有很多个: TaskTrackers

**JobTracker负责：**

接收客户提交的计算任务

把计算任务分给TaskTrackers执行

监控TaskTracker的执行情况

**TaskTrackers负责：**

执行JobTracker分配的计算任务

**6.Hadoop的特点**

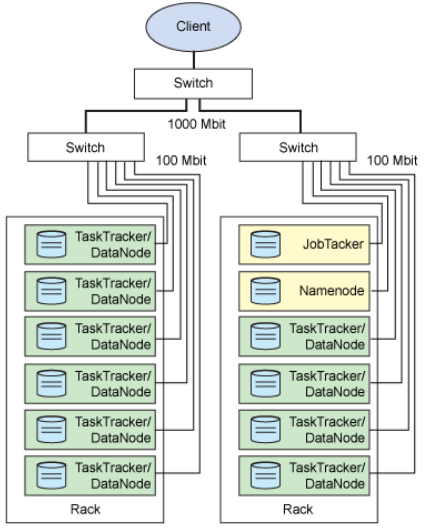
**扩容能力（Scalable）：**能可靠地（reliably）存储和处理千兆字节（PB）数据。

**成本低（Economical）：**可以通过普通机器组成的服务器群来分发以及处理数据。这些服务器群总计可达数千个节点。

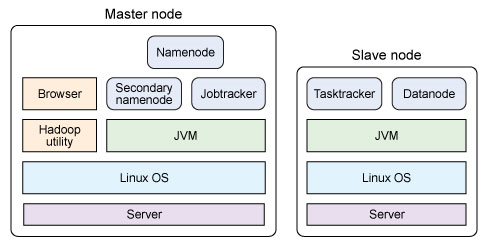
**高效率（Efficient）：**通过分发数据，hadoop可以在数据所在的节点上并行地（parallel）处理它们，这使得处理非常的快速。

**可靠性（Reliable）：** hadoop能自动地维护数据的多份副本，并且在任务失败后能自动地重新部署（redeploy）计算任务。

**7.Hadoop集群节点的物理分布**



**8.单节点物理结构**



**9.Hadoop部署方式**

本地模式

伪分布模式

集群模式

**10.要点：**

主从节点之间的通信是跨虚拟机(服务器)通信。