* + 1. **数据**

数据是事实或观察的结果，是对客观事物的逻辑归纳，适用于表示客观事物的未经加工的原始材料。数据可以是连续的值，比如声音、图像、称为模拟数据。也可以是离散的，如符号，文字，统称为数字数据。数据量最小的单位是bit，按顺序给出所有单位：bit、byte、KB、MB、GB、TB、PB（企业级数据级别）、EB、ZB（全球数据总量级别）、YB、BB、NB、DB。

1 btye = 8 bit

大数据，传统数据处理应用软件不足以处理（1.存储2.计算）它们的大而复杂的数据集。

处理：海量数据的存储和运算。

大数据的特点（4V）:

1. 容量大：数据的大小决定所考虑的数据的价值和潜在的信息
2. 种类：数据类型的多样性，包括文本，图片，视频，音频。

结构化数据：可以用二维数据库表来抽象，抽取数据规律

半结构化数据：介于结构化和非结构化之间，主要指XML，HTML等，也可称非结构化。

非结构化数据：不可以用二维表抽象，比如图片，图像，音频，视频等

1. 速度快

产生数据速度爆炸式增长：要求处理数据速度快。

运用分而治之的手段处理、加机器

数据量 处理效率 响应时间

100m 100m/s 1s

1000m 100m/s 10s

10000m 100m/s 100s

1. 价值高

整体价值高，单条记录无意义

Hadoop的核心组件：

1. Common（基础功能组件）工具包，RPC框架，JNDI和RPC
2. HDFS，Hadoop Distributed File System分布式文件系统
3. YARN运算资源调度系统
4. MapReduce，Map和Reduce分布式运算编程框架，跑在YARN系统上的

Apache Hadoop软件库是一个允许跨集群用简单的模型对于大数据的分布式处理的框架。它的目的是扩大从单一服务器到成千上万的机器，提供每个本地计算和存储。

发行版：

Apache hadoop：入门学习，简单

Cloudera hadoop：比Apache hadoop性能更强，企业使用较多

Hortonworks hadoop：有雅虎和风投公司，文档全面，2018年cloud和Hortonworks合并。

Apache hadoop：0.x 1.x 2.x 3.x

云计算：是指分布式计算、并行计算、网络计算、多核计算、网络存储、虚拟化、负载均衡等传统计算机技术和互联网技术融合发展的产物。借助laas(基础设施即服务)、PaaS(平台即服务)、SaaS（软件即服务）等业务模式，把强大的计算能力提供给终端用户。

1. 现阶段，云计算的两大底层支撑技术：“虚拟化”和“大数据技术”
2. 而Hadoop则是云计算的PaaS层的解决方案之一，并不等同与PaaS，更不等同于云计算本身。

集群：

集群 + 负载均衡：每个节点是可以独立完成作业，并发高，连接数多，需要多个节点处理请求，每个节点的负载尽可能的均衡。

分布式：1、该软件会划分成多个子系统或模块，各自运行在 不同的机器上，子系统或模块之间通过网络通信进行协作，实现最终的整体功能。

1. 比如分布式操作系统、分布式程序设计语言及其编译（解释）系统、分布式文件系统和分布式数据库系统等。

利用多个节点共同协作完成一项或多项具体业务能力的系统就是分布式系统。

负载均衡集群：每个节点都可以独立处理完一个业务，业务量较多，需要多个节点减轻压力。

hdfs架构：

1. 一主多从

namenode（nn）：存储文件的元数据，如文件名，文件目录结构，文件属性（生成时间、副本数、文件权限），以及每个文件的块列表和块所在的DataNode等。

1. datanode(dn)：在本地文件系统存储文件块数据，以及块数据的校验和
2. Secondary namenode(2nn)：辅助namenode进行元数据保存。

YARN架构：

1. 一主多从 --》1rm n个nm

ResourceManager(rm)：整个集群资源调度和管理

1. NodeManager(nm)：处理单个节点资源调度。处理rm的命令
2. ApplicationMaster：数据切分、为应用程序申请资源，并分配给内部任务、任务监控与容错。
3. Container：定义任务运行环境的抽象，封装了CPU、内存等多维资源及环境变量、启动命令等任务运行相关的信息。

MapReduce架构：

1. Map阶段并行处理输入数据
2. Reduce阶段对Map结果进行汇总

hadoop的集群搭建

1. 单节点
2. 伪分布
3. 完全分布式
4. HA（高可用）
5. 联邦模式