## 讲稿1.1 Hadoop概述

为什么要学习Hadoop存储和运算

首先是人才需求量大，很多互联网公司都会招聘Hadoop大数据工程师。

然后，就是工资待遇比较高。

最后，我觉得比较重要的是，hadoop大数据技术的前景很好，随着云计算、人工智能技术的发展。hadoop大数据作为基础，已经成为互联网从业者的必备技能。

我们来看看网上的一些hadoop工程师的招聘信息。

可以看到几乎所有的招聘待遇都在15k以上。

职位要求包括熟悉大数据分析技术，包括HDFS、Hive、Spark、MapReduce等等。

像这些大数据相关技术都会在我们这门课程中学习到。

什么是Hadoop框架

Hadoop是Apache基金会组织的一个顶级项目，其核心为HDFS和MapReduce，HDFS为海量数据提供存储，MapReduce海量数据提供计算。

解决问题：海量数据的存储和计算

解决方法:

  海量数据的存储 -> 分布式存储

  海量数据的计算 -> 分布式并行计算

接下来，我们了解一下Hadoop的发展历史：

hadoop最原始的理论和架构的原型最早来源于于google的三驾马车，这三驾马车也就是Google公司的核心技术。

后来google通过三篇论文公布了这三项核心技术的原理。

后来经过doug cutting的山寨，出现了java版本的 hdfs   mapreduce 和 hbase

以上三个组件整合起来成为apache的一个顶级项目  hadoop

经过演化，hadoop的组件又多出一个yarn

最终形成， hdfs+mapreduce+ yarn 组成hadoop的三大核心组件。

并且，hadoop外围产生了越来越多的工具组件，形成一个庞大的hadoop生态体系。

介绍完hadoop的发展历史，我们来探讨一下什么是hadoop.

首先，hadoop是用于处理（运算分析）海量数据的，且是采用分布式集群的方式；

通俗来说，可以把hadoop理解为一个编程框架 （比如常见的structs、spring、hibernate）

从另一个角度，hadoop又可以理解为一个提供服务的软件（比如数据库服务oracle/mysql）

 总而言之，Hadoop与传统的软件并无本质的区别，依然是提供编程框架和数据服务。

具体来说，hadoop两个大的功能：海量数据的存储；海量数据的分析；

Hadoop有3大核心组件：

HDFS ---- hadoop分布式文件系统海量数据的存储(集群服务)，

MapReduce ----运算框架，海量数据运算分析

Yarn ----资源调度管理集群(可以理解为一个分布式的操作系统，集群服务)

介绍完了hadoop的基本概念，下面了解一下hadoop的发行版。

hadoop目前的发行版主要有以下几个，分别是CDH， HDP，MapR

CDH是一个较为成熟的Hadoop发行版，由Cloudera公司发布。

HDP是Hortonworks开发的，它提供的Hadoop和其他软件是完全免费的，只对技术支持和培训收费。

其他的一些发行版，如 MapR、DKhadoop我们先不做了解。

了解完了Hdoop的发行版，接下来认识一下hadoop生态圈技术相关的具体应用，其中最为重要的应用之一就是云计算。

/

首先，什么是云计算呢？

一首先让大家明白什么是云端，所谓云端需要两层理解

（1）服务不在本地，这一层可以理解为服务器

（2）它和普通的服务器是不一样的，这些云端的服务器的资源是共享的，一旦一个服务器不能承受，将会把任务分配给其他机器。

二、云技术与其他技术的区别：

可以用比如Java语言开发。与其他技术，最显著的区别，不是在开发上，而是在于架构上，最显著的特点是分布式。

云计算的部署方式

从部署方式来说，总共有两类云计算：

·         私有云：数据中心部署在企业内部，由企业自行管理。。

·         公共云：数据中心由第三方的云计算供应商提供，供应商帮助企业管理基础设施（例如硬件，网络，等等）。

企业将自己的软件及服务部属在供应商提供的数据中心，并且支付一定的租金。像Windows Azure正是这样一个公共云平台。

从云计算的运营方式来说，总共有三类云计算：

·         软件即服务（SaaS）：云计算运营商直接以服务的形式供应软件，供最终用户使用。

有些服务还提供了SDK，从而使得第三方开发人员可以进行二次开发。

在这种运营模式下，开发人员通常只能针对现有的产品开发插件，而无法充分挖掘平台和操作系统的特点，

不过他们可以在现有产品的基础上添加新的功能，而不必从头开始实现。

微软的Bing，Windows Live，Microsoft Business Productivity Online等产品就属于这一类型。

·         第二种是平台即服务（PaaS）：

云计算运营商将自己的开发及部署平台提供给第三方开发人员，第三方开发人员在这个平台上开发自己的软件和服务，供自己或其它用户使用。

在这种运营模式下，开发人员有了更多的自由，可以发挥出平台的强大功能，而不受现有产品的束缚。

Windows Azure正是这样一个产品。

·         基础设施即服务（IaaS）：

云计算运营商提供但不管理基础设施，第三方开发人员将开发好的软件和服务交给自己公司的IT管理员，由IT管理员负责部署及管理。

在这种运营模式下，开发人员和IT管理员有最大限度的自由，然而由于必须自行管理部分基础设施，因此成本通常也会较大，对管理员的要求也会较高。目前微软尚未提供IaaS的云计算运营模式，不过我们正在考虑如何给予开发人员和IT管理员更多的自由。

总结一下，云计算指的就是把你的软件和服务统一部署在数据中心，统一管理，从而实现高伸缩性。

从部署方式来说，云计算可以分为私有云和公共云。从运营方式来说，云计算可以分成SaaS，PaaS，IaaS三类。

那么有同学会问，像我们国内的阿里云和腾讯云是部署方式中那一层呢？

实际上阿里云已经构建了涵盖IaaS、PaaS、SaaS，具有从弹性计算、存储、安全到异构计算、大数据、垂直行业解决方案的完备的[云计算](https://www.baidu.com/s?wd=%E4%BA%91%E8%AE%A1%E7%AE%97&tn=SE_PcZhidaonwhc_ngpagmjz&rsv_dl=gh_pc_zhidao)产品体系。

所以，他们提供综合性的云计算业务和服务。

## 讲稿1.2 安装Hadoop前的准备

上一节，我们初步接触了Hadoop的概念、框架、发展历史和云计算应用。这一节开始我们介绍Hadoop软件的安装。

首先，在安装之前要对所需要的机器、资源、环境等最一些基本了解和具体操作。

一般我们用Hadoop进行学习和测试可以采用两种部署模式：

第一是单机部署（伪分布式）

伪分布模式这种模式是在一台单机上运行，但用不同的Java进程模仿分布式运行中的各类结点。

第二是集群部署

  可以准备三台虚拟机，模拟分布式集群部署。

接下来讲解一下机器和资源的具体要求：

1. 准备三台linux服务器（系统建议Ubuntu或者centos 64位）

•解压centos镜像压缩包到某个目录，并用vmware打开

•准备操作系统环境（主机名，ip地址配成static，域名和ip的本地映射hosts）

•关闭图形界面的启动   修改/etc/inittab中的启动级别为3

2. Linux 基础

鸟哥的linux私房菜：基础学习篇（第四版）

熟悉http、https、tcp等网络协议，对CDN有所了解

ok，准备好机器资源之后，开始着手安装java环境。

1.针对Hadoop安装Java8

登录网址：<http://www.oracle.com/technetwork/java/javase/downloads/index.html>

下载安装包，比如64位的JDK1.8版本。

jdk-8u231-linux-x64.tar.gz

2. 将安装包移动到/opt目录并，解压gz文件

tar –zxvf jdk-8u231-linux-x64.tar.gz

创建软连接

ln –s jdk1.8.0\_231 java

3. 相关安装包准备

准备java环境，安装jdk，配置环境变量等

•解压安装包，添加环境变量：  JAVA\_HOME   PATH

vim ~/bashrc

环境变量生效：source ~/bashrc

验证：java –version

如果输出jdk1.8的信息表示安装正确

下载Hadoop安装包：

登录网址：

<https://archive.apache.org/dist/hadoop/core/hadoop-3.0.0/>

下载 hadoop-3.0.tar.gz

解压：

tar –zxvf hadoop-3.0.tar.gz

1. 配置防火墙（关闭）

service iptables stop

2. 为hadoop软件准备一个专门的linux用户

（比如，创建新用户hadoop，密码：hadoop），为hadoop用户设置sudo权限

在/etc/sudoers  配置文件中设置

设置SSH免密码登录：

各个节点之间进行数据交互。

测试是否为无密码登录localhost：$ssh localhost

设置命令：

最后再进行测试，验证效果。

ok，我们总结一下本小节所讲的内容：

1. 第一个部分是，基础资源配置，介绍伪分布式集群部署的概念。

2. 第二部分讲了，如何Java环境准备

3. 第三部分介绍了， 如何把Hadoop包下载下来，准备安装。

4. 最后讲了，Hadoop安装的其他准备，如防火墙设置、用户设置和SSH免密码登录。