大数据实验室建设方案

用团队 35 年来的沉淀与积累，

赋能高校人工智能专业、大数据专业学科建设、人才培养。

ZX人工智能

目录

一、 大数据实验室建设分析 3
(一) 企业对大数据人才的需求 3
(二) 大数据实验室建设项目目标 3
(三) 大数据实验室建设方案 4
二、 大数据实验室教学资源实验实训平台 5
(一) 覆盖整个专业教学流程的课程体系 6
(二) 完备方便的师生信息管理 6
(三) 方便敏捷的深度学习调度系统 8
(四) 功能强大的实验平台 11
三、 大数据科学与大数据技术核心课程 19
(一) 从事大数据技术的主要课程 19
(二) 大数据专业核心课程 19
(三) 大数据课程部分实验实训内容 22
四、 大数据实验实训资源库介绍 23
(一) 大数据实验室实验实训资源建设 23
(二) 《机器学习》课程实训环节（部分） 24
(三) 《HADOOP 大数据》综合实训项目（部分） 24
(四) 大数据综合实训项目示例—电子商务网站商品精准推荐（部分） 24
(五) 大数据综合实训项目示例二-某互联网金融产品网络舆情监测 25
(六) 大数据综合实训项目示例三-教育平台日志分析 25

一、 大数据实验室建设分析

(一) 企业对大数据人才的需求

为深入的了解用人单位对求职者的要求，更有针对性的建设大数据实验室，ZX人工智能调研部分市场对大数据人才的需求情况。发现互联网公司、数据分析、人工智能类公司对基于 Python 编程、数据清洗、数据挖掘和获取信息的 Python 爬虫需求较大。因此ZX人工智能设计大数据实验室方案时，充分考虑上述问题，提出满足企业需求和学校培养定位的建设方案，缩短高校教学和企业需求之间的关系。

地区	北京				
关键词	1、python	2、数据挖掘	3、数据清洗	4、python爬虫	
总数	6161	2053	209	9	8432
无经验	32	16	3	0	
1-3年经验	190	58	12	0	311
占比	3.60%	3.60%	7.18%	0	3.69%
地区	上海、成都、广州、深圳、杭州				
关键词	1、python	2、数据挖掘	3、数据清洗	4、python爬虫	
总数	21291	7276	632	85	29284
无经验	145	74	17	8	
1-3年经验	1598	531	49	13	2435
占比	8.19%	8.32%	10.44%	24.71%	8.32%
地区	桂林				
关键词	1、python	2、数据挖掘	3、数据清洗	4、python爬虫	
总数	440	154	20	3	617
无经验	6	1	0	0	
1-3年经验	42	13	1	0	63
占比	10.90%	9.09%	5.00%	0	10.21%
职位总数	38950				
可达标总数	2809				
占比	7.21%				

数据来自前程无忧网

(二) 大数据实验室建设项目目标

2015 年 8 月 31 日，国务院印发《促进大数据发展的行动纲要》，成为中国发展大数据产业的战略性指导文件。2016 年，继国家发后，环保部、国务院办公厅、国土资源部、国家林业局、煤工委、交通运输部、农业部均推出大数据发展意见和方案，大数据政策从全面、总体规划逐渐朝各大产业、各细分领域延伸，大数据产业发展也在逐步从理论研究走向实际应用之路。2017 年，大数据产业的发展正从理论研究加速进入应用时代，大数据产业相关的政策内容已经从全面、总体的指导规划逐渐向各大行业、细分领域延伸，渗透至物联网、云计算、人工智能、5G 技术等领域。

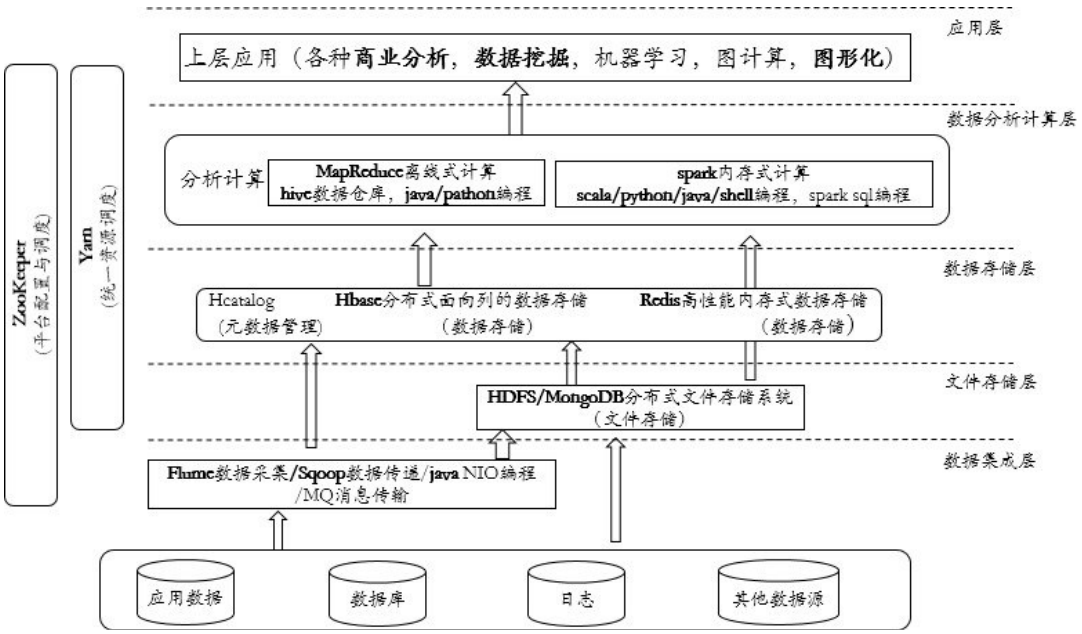
据前瞻产业研究院发布的《大数据产业发展前景与投资分析报告》数据显示，2016 年，全球大数据产业市场规模为 1403 亿美元，预计到 2020 年将达到 10270 亿美元，2014-2020 年间 CAGR 高达 49%;2015 年，我国大数据产业市场规模为 1692 亿元，预计到 2020 年将达到 13626 亿元，2014-2020 年间复合年增长率高达 53%。

围绕 XX 大学计算机和电子类专业建设，推动科教结合、产教融合协同育人的模式创新，多渠道培养大数据领域创新创业人才，依托ZX人工智能的优势产业资源和教育市场布局，XX 大学拟建设“大数据实验室”项目，创新产教融合校企合作机制，培养大数据应用领域技术技能人才，建设院校专业人才梯队,增强学校办学特色和学生就业创新能力，努力推动 XX 大学综合实力、区域竞争力、服务创新发展能力跃上新台阶，实现 XX 大学综向办学水平高、专业特色鲜明、与区域产业发展高度融合的应用型高校转变。具体没建设目标如下。

- 建设大数据实验室，主打复合专业特色，创新 XX 大学人才培养模式，打造成具有鲜明特色的电子信息工程和计算机工程专业，树立 XX 大学品牌、扩大在全国的知名度。
- 创新人才培养模式，优化人才培养方案，改革专业课程体系，建设优质课程资源，通过案例教学实施，提高就业竞争力。

(三) 大数据实验室建设方案

大数据与处理是人工智能方向的核心，是经典机器学习和深度学习的必然应用领域。需要部署以计算集群为主的硬件实验环境、多种流行计算框架，之后是智能科学与技术、大数据科学与技术的相关知识。由于涉及到多门学科交叉，该专业知识体系比较庞大，常规情况下学生需要投入相当多的时间从事学习和实验，同时也对教师提出较高的要求。



专业特色与市场需求紧密结合的大数据方向培养体系

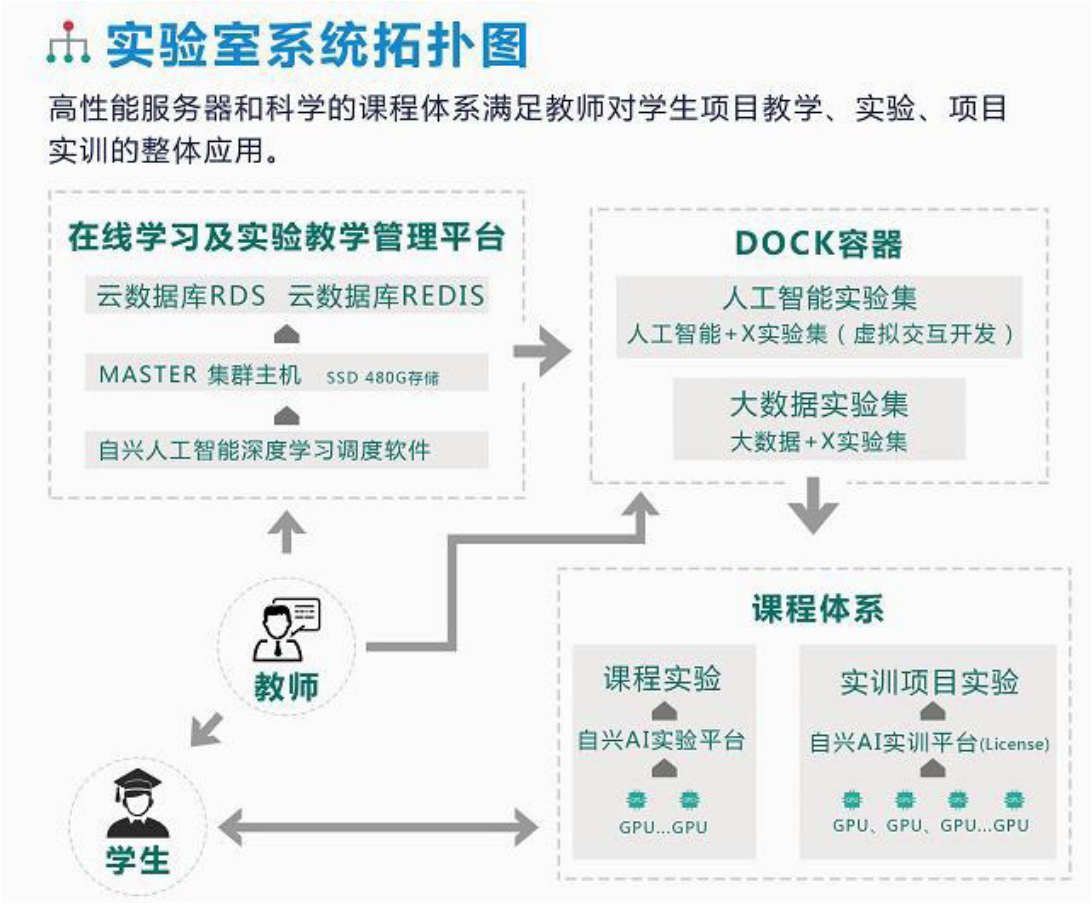
ZX人工智能认为，服务地方经济发展的应用型大学应侧重培养学生的实际应用能力，应该让学生掌握该课程体系最为核心关键的知识。为此，我们提出

大数据实验室建设方案=数据科学与大数据技术+专业特色+市场需求

的建设方案。就 XX 大学的现状，ZX人工智能学院设计了一条从大数据采集、存储、分析计算层到应用层面的最优解决方案，囊括大数据科学技术方向必需的、最满足企业需求的核心课程。同时ZX人工智能学院提供课程 ppt、实验、课程论文、学年论文和毕业论文等资料，降低教师工作强度，增加学生实验实训比重，提升与企业的契合度。

二、 大数据实验室教学资源实验实训平台

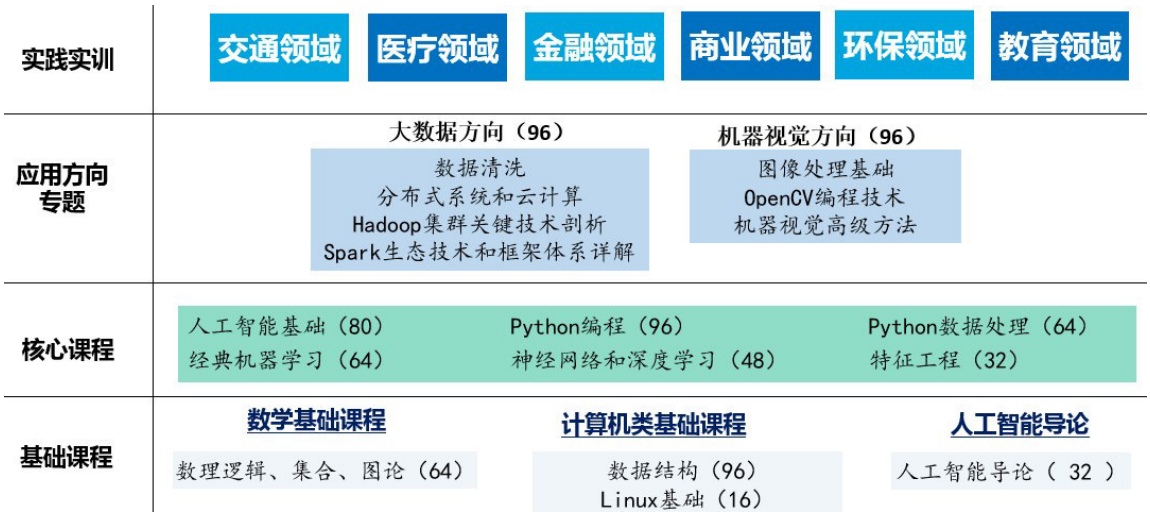
ZX-AI 教学实践资源平台是基于高等教育实践、课程资源重用、专业实训资源共享的概念而产生的，简称ZX-AI 实验室，是为高等教育事业提供课程实验和专业实训服务的平台。以大数据科学与大数据技术核心课程实验和综合实训为核心、通过对课程环节的梳理，设置满足教学需要的课程实验实训。系统由课程实验、专业实训、教师和学生的管理信息系统、负责服务器集群计算能力调度的算力调度系统四部分构成。具体如下图所示。



大数据实验室拓扑结构如上图所示。师生管理信息系统负责教师和学生角色的分配调度，实现课程、学生和实验内容的管理。调度软件负责 master 集群主机计算能力的调度，均衡各服务器的计算负载，提高系统整体计算效率。DOCK 容器封装不同专业的实验集，具有优秀的拓展性，满足不同专业的实验实训需求。课程体系包含课程实验和实训实验两部分，前者提供专业骨干课程实验，后者以公共云的形式提供学科的大型实训项目，缩短学校教育和企业需求之间的差距。

(一) 覆盖整个专业教学流程的课程体系

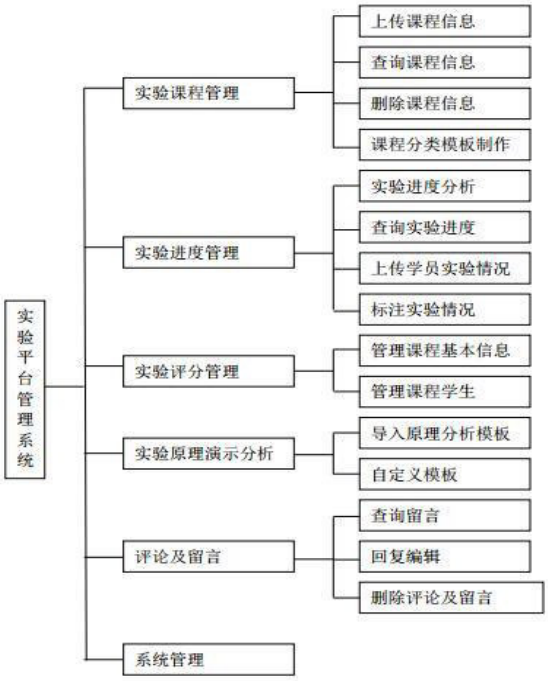
在平台中所有的应用都以资源为核心的，ZX-AI 教学实践资源共享平台涵盖了人工智能+X 专业的多个方向的课程实验实训。平台以建设专业课程集群为目标， 服务专业教学过程整个环节，提供课程 ppt、课程实验指导书、课程大纲、教案、习题库这些基础资源库为核心，然后与任务实训、在线课程、作业、测评考试等应用之间相互交互。如下图展示。



综上所述，基于ZX-AI 教学实践资源共享平台以资源为核心，其他应用之间通过基础资源进行交互，重点突出资源在平台中的重要性，促进大数据科学与大数据技术的向应用层面的快速发展。下一阶段ZX人工智能学院将在大数据科学与大数据技术层面做出更多努力，对接商业、工程等具体的行业需求；接下来实现大数据与社会科学的深度融合，促进数据科学在人文社会科学中的广泛深入应用，促进相关学科的科学化和计量化发展。

(二) 完备方便的师生信息管理

系统分为管理员、老师、学生三个角色。系统功能模块分为原理演示、基础练习、 项目实验三个进程，涉及计算机视觉、图形信息处理、数据库系统原理和计算机图形学等课程体系。其中用户原理功能演示，用户可以调用系统已有演示模板，也可以自己创建新的模板。具体功能如下图所示。



该系统为 B/S 三层结构，它的运行环境分客户端、应用服务器端和数据库服务器端三部分。以下是系统的软件环境。

（1） 客户端

操作系统：Windows 7 或更高版本操作系统。

浏览器： IE6 以上，其它常见浏览器如 FireFox，Google Chrome 等。

（2）应用服务器端

操作系统： Windows2008 Server 或更新版本。

应用服务器： Tomcat 7.0 或更新版本。

数据库访问： JDBC。

（3）数据库服务器端

操作系统： Windows2008 Server 或更新版本。

数据库系统： SQLServer 2000 或更新版本。

本系统主要用于系统管理员、教师和学生三类人员。系统管理员，完成系统管理与维护，例如，维护学生、教师等账号。教师，上传实验课程，查询分析实验课程进度，导入实验课程基础知识和实验课程项目。学生，可查询自己的实验课程进度，编写自己的实验和基础课程模板以及评论和留言。

学生管理基本事件流

1、 用户进入批量导入学生界面，本用例开始；

2、系统显示导入文件类型、格式说明、并提供导入的模板文件下载。

3、用户按照导入文件格式要求填写或生成对应文件，上传文件，点击确定。

4、系统检查文件的合理性，如果文件格式有误或有数据冲突，给出详细提示列表（错误所在行、错误原因），用户修改文件后再上传，如果上传文件合理，系统将学生信息导入系统。本用例结束。

教师管理基本事件流：

1、用户进入批量导入教师界面，本用例开始；

2、系统显示导入文件类型、格式说明、并提供导入的模板文件，供下载。

3、用户按照导入文件格式要求填写或生成对应文件，上传文件，点击确定。

4、系统检查文件的合理性，如果文件格式有误码或有数据冲突，给出详细提示列表（错误所在行、错误原因），用户修改文件后再上传，如果上传文件合理，系统将教师信息导入系统。本用例结束。

实验课程管理基本事件流：

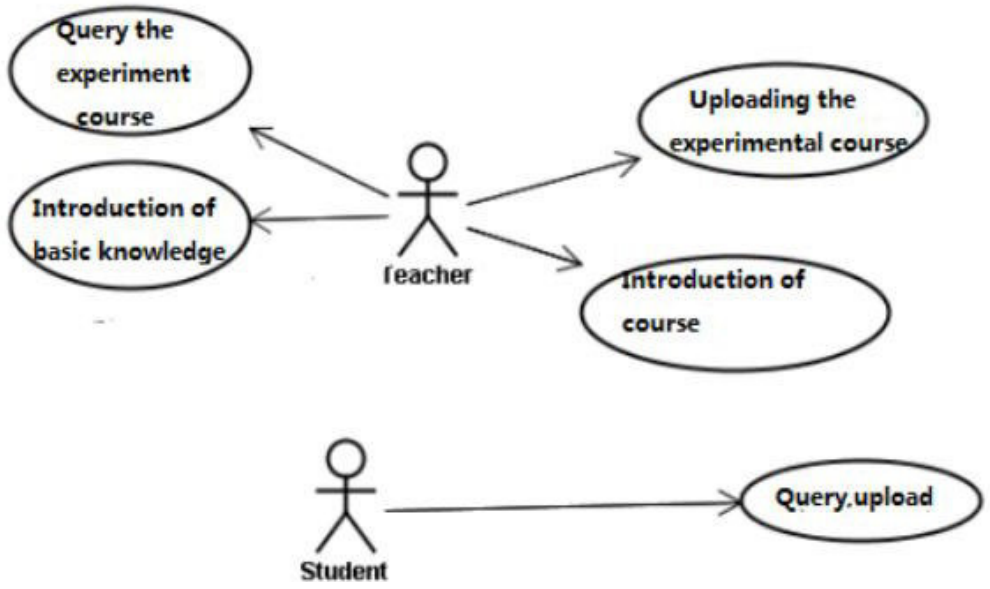
1、用户维护课程学生信息界面，本用例开始；

2、系统显示课程列表，用户选择要加入学生的课程，系统显示该课程已存在的学生；

3、用户选择加入新学生，系统显示学生列表，用户可通过搜索列表显示班级下的学生，用户选择要加入课程的学生。

4、系统将所选学生加入前面选定的课程，本用例结束。

关于教师和学生用户具体事例如下图所示。

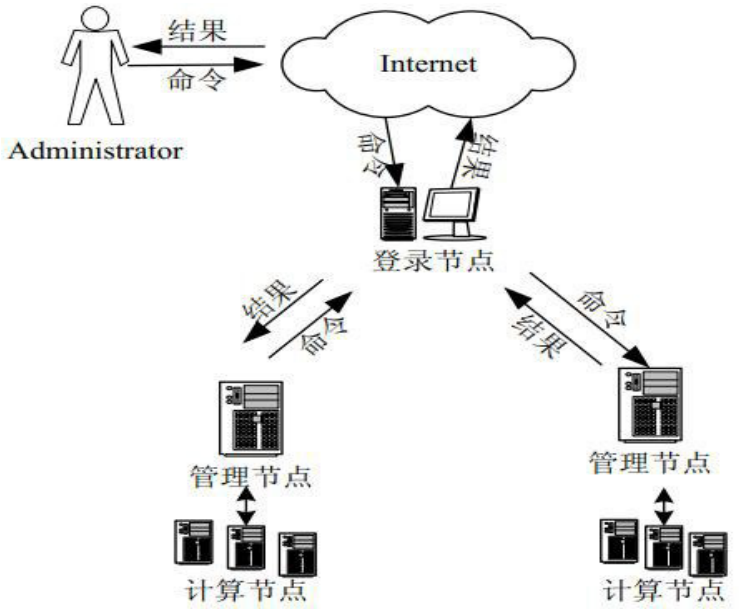


(三)　方便敏捷的深度学习调度系统

深度学习调度系统管理集群系统中的软硬件资源和用户提交的作业，根据集群中的资源使用情况来合理的调度用户提交的作业，从而达到提高资源的利用率和作业的执行效率的作用。调度软件是集群管理系统中非常重要的一部分，它负责管理用户提交的作业， 合理地给各个作业分配资源， 从而确保充分利用集群系统计算能力，并尽可能快地得到运算结果，与集群管理监控软件、部署软件形成一套完整的集群管理系统。

系统具有（1）统一的集群使用与管理平台，通过该平台可以完成对集群的使用

与管理工作，无需借助其他工具。（2） 模块化设计，每个模块完成相对独立的功能，方便用户操作集群，提高软件易用性；可定制以及动态添加新模块。（3）.权限控制：可以控制每个用户可使用的模块，方便进行管理。同时用户可以定制自己的首页面。本软件运行于计算集群之上，可以管理 1 个或者多个集群，拓扑图如下所示。



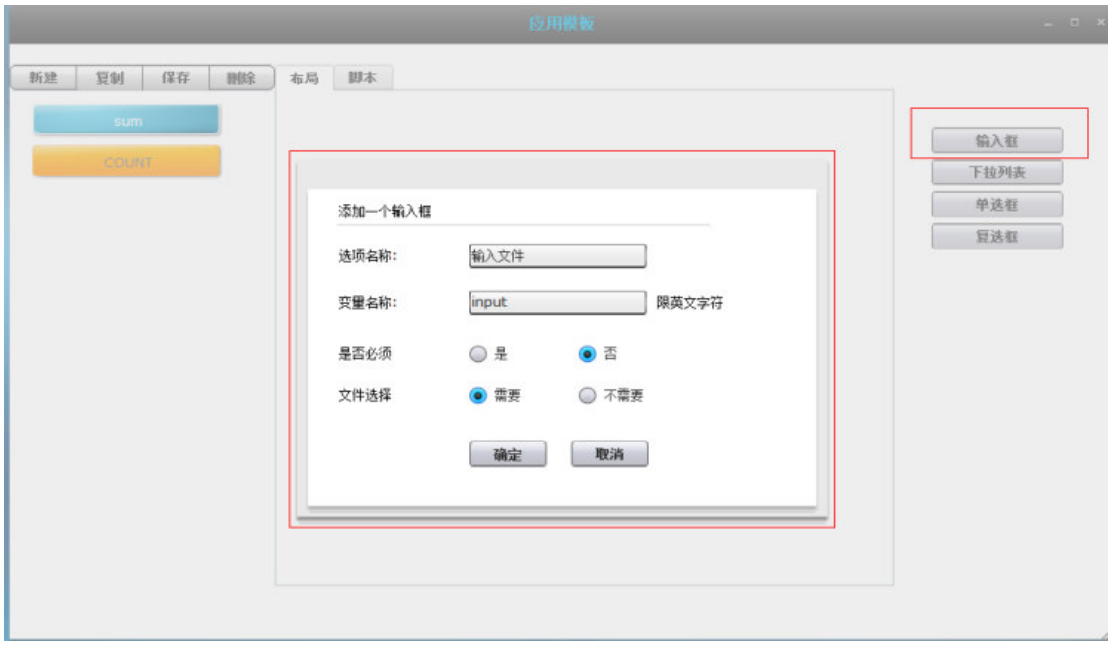
调度软件能够增加、删除和查看集群状态，如下图所示。

用户可以对特顶节点进行操作，如远程登录、查看节点监控视图、节点配置和远程开关机等操作。

用户通过应用模板模块，可以针对各种应用自定义不同的作业提交页面。应用模板的设置包含两个部分：页面布局及作业脚本。用户通过页面布局确定该应用的作业提交选项及输入方式；通过作业脚本设置应用的运行及如何使用作业提交页面中输入的参数。新建成功的应用模板,将会显示在作业提交模块的作业列表中. 用户在作业提交页面中，选择已设置好的作业模板，输入模板中设定的作业参数，即可将输入参数传递给对应的应用脚本，并提交该应用脚本。具体功能为新建作业模板

、保存作业模板、复制作业模板、删除作业模板和编辑作业模板。

在应用模板页面右侧操作栏中点击“输入框”，将弹出添加输入框页面，在该页面中输入选项名称，变量名称，选择是否是必须选项，是否需要文件选择功能，点击确定按钮，将给指定应用模板添加一个输入框选项。



作业提交页面

作业查看页面

本调度软件实现到了对计算集群全面的管理，方便的调度各个计算集群/服务器的计算能力，降低了单个服务器的计算负载，均衡整个计算集群计算负载，提高了系统的计算能力。

(四)　功能强大的实验平台

ZX人工智能开发的ZX人工智能实验实训平台是是一个基于浏览器/服务器的开源工程技术教学和实战平台，可以为高校、企业、协会提供覆盖全 IT 技术栈的教育服务，已经与多家知名高校建立长期战略合作，处于国际领先水平，为湖南省人工智能学会实训指定平台。目前该平台已经为人工智能的核心课程，如人工智能基础、Python 编程、Python 数据处理、经典机器学习、神经网络和深度学习建立起完善的实验实训体系，解决部分院校教学师资不足、与实际市场脱节的问题。同时应用层面也已经搭建起大数据和机器视觉两个方向，大数据方向建立了数据清洗、分布式系统和云计算、Hadoop 集群关键技术剖析、spark 生态技术和框架结构详解；机器视觉已经搭建起图像处理技术、OpenCV 编程技术、机器视觉高级方法。

在该平台上我们已经部署了部分课程的实验实训内容，满足教学环节实验、实训的要求。

ZX-AI 平台已经部署的部分课程实验实训内容。

可以方便的查看、批改学生提交的实验实训作业

本系统能够支持教师的全部教学活动，如学生导入、课程实验、课程实训项目设计、期末考试、毕业设计等。

学生导入系统后，按照学号和班级排序，并且能够显示上次作业的提交状态、作业的完成情况和最终的评分；教师可以调整学生分数和查看学生成绩。

本系统提供了良好的毕业设计管理功能，整个学院中每位教师指导学生毕业论文的进度都在这里展示。另外系主任可以发布信息、通知等给每位教师。下面以《面向对象的编程语言—Python》给出课堂实验和实训的案例。

实例一课堂实验：下面以《面向对象的编程语言—Python》为例说明该系统的使用过程。实验系统按照 python 语言的学习过程设计实验和实训过程，教师可以在授课后安排相应的实验作业给学生，学生提交后系统自动批改学生作业。下面以“Python 图像创建”说明实验过程。

如上图所示，右边部分是学生的编程环境，在学生提交代码后，系统会自动判定所提交程序，确定学生是否可以进入下一个实验任务。若学生编程出现错误，系统会给出提示，学生也可以查看左边的相关知识点。左边是过关任务、背景知识、参考答案和评论四个部分。参考答案可以教师控制是否让学生观看，评论区中学生能够能够上传对本次实验的思考和总结。

过关任务：任务描述部分给出学生在本次实验中应完成的任务，包括本关必读、本关任务和测试说明三部分。

本关必读

Python 扩展库 pillow 提供了大量用于图像处理的对象，其中 Image 用来创建或读取图像文件，ImageDraw 用来创建可绘制的图像对象。Image 对象的 getpixel()方法可以读取图像指定位置的像素颜色值，而 putpixel()方法可以设置指定位置的像素颜色值。

本关任务

本关的编程任务是补全 step1/step1.py 中的 generateCircle(width, height)函数，要求实现创建图像的功能。具体要求如下：

在当前文件夹 step1 中创建一个图像文件 circle.png，该图像文件的尺寸由函数参数 width 和 height 确定，要求在该图像中绘制一个椭圆，该椭圆的外接矩形左上角坐标为(0,0)，右下角坐标为(width-1,height-1)。

具体输入输出请参见后续测试样例

本关涉及的代码文件 step1.py 的代码框架如下：

```
from PIL import Image, ImageDraw


def generateCircle(width, height):
    # 请在这里补充代码，完成本关任务
    #-----------Begin----------


    #------------End-----------
```

其中：width 和 height 是传递给函数 generateCircle 的参数。

测试说明

本关的测试文件是 step1/testStep1.py，负责对你写的实现代码进行测试。该测试文件代码如下：（注意：测试程序 testStep1.py 中的代码不能被修改！）

```
import sys
from PIL import Image, ImageDraw
import step1

# 获取测试输入，设置图像宽度和高度
width = int(input())
height = int(input())

# 调用学生编写的代码，生成图像
step1.generateCircle(width, height)

# 由测试程序生成同样的图像
```

```
    img = Image.new('RGB', (width,height), (255,255,255))
    imgDraw = ImageDraw.Draw(img)
    imgDraw.arc((0,0,width-1,height-1), 0, 360, fill=(0,0,0))

    # 读取学生代码生成的图像，与测试程序生成的图像进行对比
    imgStudent = Image.open('circle.png')

    for w in range(width):
        for h in range(height):
            if img.getpixel((w,h)) != imgStudent.getpixel((w,h)):
                print(False)
                break
    else:
        print(True)
```

以下是平台测试样例集：测试输入：

50

30

预期输出：True

左边栏第二部分是背景知识。由于该部分很长，我们只截取部分内容，并用楷体与正文部分加以区分。

Python 语言介绍

Python 是目前最受欢迎的程序设计语言之一，由 Guido van Rossum 发明，1991 年发布了第一个版本。Python 是一种解释型语言，非常强调代码可读性，语言表达能力极强，很多非常复杂的功能只需简单几行代码，这在 C++或 Java 等编程语言中是不可想象的。1989 年圣诞节期间，Guido 在阿姆斯特丹的家里启动了 Python 项目，据说是为了打发圣诞节的无趣，但实际上是想完成孕育了多年的愿望：设计一种对广大 Unix/C 黑客们更具吸引力的脚本语言。Python 是对 ABC 语言的继承和发展，以彻底解决 ABC 语言的封闭性问题。

ABC 是由 Guido 参与设计的一种教学语言。ABC 非常优美和强大，专门为非专业程序员设计，但 Guido 认为其封闭性导致了 ABC 语言的失败。因此，Guido 决心在 Python 中避免这一错误。同时，Python 还结合了 Unix shell 和 C 的一些习惯。Python 使用 C 语言开发。

自 2004 年以后，Python 进入快速发展阶，在国外用 Python 做科学计算的研究机构日益增多，卡耐基梅隆大学、麻省理工学院已采用 Python 来教授程序设计课程。

Python 提供了极其丰富的标准库和扩展库，大量开源科学计算的软件包都提

供了 Python 的调用接口。经过多年的快速发展，Python 已经形成了一个覆盖科学计算、工程开发、数据分析、图表制作等众多功能支持的非常强大的软件技术生态。

……

我们在选择一种开发语言时，不仅要考虑语言本身的特性，很多时候还需要考虑开发团队的特点，已有软件资产的编程语言和架构，当地程序员人才的分布特点等各种因素。

参考文献

[1]. 董付国编著.Python 程序设计基础[M].北京：清华大学出版社,2015

[2]. 董付国编著.Python 程序设计[M].北京：清华大学出版社,2015

[3]. 董付国编著.Python 程序设计（第 2 版）[M].北京：清华大学出版社,2016

[4]. 董付国著.Python 可以这样学[M].北京：清华大学出版社,2017

[5]. 董付国著.Python 程序设计开发宝典[M].北京：清华大学出版社,2017

[6]. Mark Lutz 著，李军、刘红伟译，Python 学习手册：第 4 版，机械工业出版社，2011 年 4 月第 1 版 http://book.51cto.com/art/201104/257562.htm

[7]. 张颖，赖勇浩著.编写高质量代码——改善 Python 程序的 91 个建议[M].北京：机械工业出版社,2014

[8]. 杨佩璐，宋强等编著.Python 宝典[M].北京：电子工业出版社,2014

[9]. Python(programming language), Wikipeida, https://en.wikipedia.org/wiki/Python_(programming_language), 2017.5

[10].Python，百度百科， http://baike.baidu.com/item/Python, 2017.5

参考代码：给出本次实验的正确代码，教师有权限决定该部分是否启用。本关任务对应参考代码实现如下：

```
from PIL import Image, ImageDraw

def generateCircle(width, height):
# 请在这里补充代码，完成本关任务
#----------Begin----------
img = Image.new('RGB', (width,height), (255,255,255))
imgDraw = ImageDraw.Draw(img)
imgDraw.arc((0,0,width-1,height-1), 0, 360, fill=(0,0,0))
img.save('circle.png')
#-----------End----------
```

实例二课程实训：实训部分是 Python 课程知识的综合运用，考察学生对知识的综合应用能力、解决实际问题的能力和编程技巧。本次实训应用 Python 模拟财富的

分配过程。我们通常说“穷的人变的更穷，富的人越来越富”，这就表达了财富分配的不均匀。财富分配问题是经济学中的一个非常热门且经典的问题,至今已有一百多年的历史,得到了经济学家的广泛关注。

规则:在一个封闭的房间里，有一百人，每个人有 100 元，每个人拿出一元钱，随机给另一位人（包括自己），所有人获得这个一块钱的概率相等，反复进行实验，最后这一百个人的财富分布到底会是怎样的？

环境和系统要求

编程环境：Python 3.5，使用到的库 matplotlib 库，运行环境 windows7 windows10 等命令窗口下运行。

初始化：给 100 人每人 100 元，用 Python 的字典数据结构来表示，编号 0-99，作为 key，初始每个人的 value 为 100.

```
people={}
#初始财富分配，每人 100 元
for i in range(0,100):
    people[i]=100
```

一次完整分配：接下来写一个财富分配函数，进行每次的随机资金流动，对每个人都进行一次财富值减一操作，并且随机分配给其他人。我们暂且先设定，当一个人财富值为 0 时候，不再进行财富支出，但不影响他获得其他人财富的概率

```
def assign():
    for key in people.keys():
        #如果财富值为 0，不再流出，但是同样可以和其他人有同样的机会获取财富
        if people[key]<=0 :
            continue
        n=random.randint(0,99) #在 0-99 编号中随机选取一位
        people[key]=people[key]-1 #每人从自己财富中-1
        people[n]=people[n]+1 #每个人自己财富+1 的机会相等
        print(key,":",people[key])
```

理想状态进行模拟

假设个人从 20 岁开始上班 60 岁退休的理想状态

```
for i in range(365*(60-20)):
    #从 20 岁开始到 60 岁退休，每天进行一次资金流动
    assign()
```

统计结果

```
#高于，等于，低于初始值人数
a,b,c=0,0,0
```

```
#高于，等于，低于初始值人财富值总和
avalue,bvalue,cvalue=0,0,0
for value in people.values():
    if value>100:
        a=a+1
        avalue=avalue+value
    elif value==100:
        b=b+1
        bvalue=bvalue+value
    elif value<100 :
        c=c+1
        cvalue=cvalue+value
print("高于初值人数：", a,"\n 保持不变人数：", b, "\n 低于初值人数：
",c )
print("\n 高于初值人财富总值",avalue,"\n 等于初值财人富总值
",bvalue,"\n 低于初值人财富总值",cvalue)
number=list(people.keys())
wealth=list(people.values())
#wealth.sort()
print("\n 个人最高：",max(wealth),"个人最低：",min(wealth))
```

使用 matplotlib 进行画图显示

```
plt.fill_between(number,wealth,0,color='green')
plt.xlabel('player')
plt.ylabel('wealth')
plt.show()
```

实验结果

学生将实验结果填充在这里

总结:

学生将实训总结填充在这里

模型拓展

在经济学中社会财富分配问题一直是非常热门且经典的问题，至今已有 100 多年的研究历史，仍然是众多经济学家广泛关注的问题。目前有多中方法研究该问题， 如个体模拟方法、经典博弈理论等。感兴趣的同学可以查阅相关资料，用 Python 语言实现上述两种财富分配方法，并做比较。

三、 大数据科学与大数据技术核心课程

(一) 从事大数据技术的主要课程

（1） 操作系统与 linux 应用

（2） 程序设计（Java）

（3） 程序设计与应用（python）-python/scrapy

（4） 数据采集与迁移技术-sqoop/flume

（5） 大数据存储与处理技术-hadoop2.x

（6） 大数据缓存技术-Redis

（7） 数据仓库技术与应用-Hive

（8） 数据挖掘与分析

（9） 内存高性能计算-spark

（10） 数据图形化-sparkR

（11） 机器学习-SparkML lib

(二) 大数据专业核心课程

课程分类	课程名称	课程介绍
程序设计（Java）	Java 高级特性	● Java 多线程； ● 并发包线程池； ● 并发包消息队列； ● JMS； ● 动态代理
操作系统与 linux 应用	Linux 基础及 shell 编程	● Linux 系统概述； ● 系统安装及相关配置； ● Linux 网络基础 ； ● OpenSSH 实现网络安全连接 ； ● vi 文本编辑器； ● 用户和用户组管理；磁盘管理 ； ● Linux 文件和目录管理 ； ● Linux 终端常用命令 ； ● linux 系统监测与维护； ● shell 编程-基本语法； ● shell 编程-流程控制； ● shell 编程-函数；
程序设计应用（python）	Python 基础与网络爬虫	● Python 语言开发要点详解； ● Python 数据类型 ； ● 函数和函数式编程； ● 面向对象编程； ● 网络爬虫；
大数据存储与处理技术	HDFS	● 讲解 hadoop 的 HDFS 分布式存储文件系统； ● 其 Java API 对文件的处理技术。

	MongoDB	● 简介与安装 ● 用户、访问控制 ● shell 操作 ● 查询、索引、聚合 ● 性能优化 ● Java API
大数据 计算与分析	Hadoop2.x	● 讲解大数据应用发展、前景； ● Hadoop 2.x 概述及生态系统； ● Hadoop 2.x 环境搭建与测试 ● YARN 架构、集群管理、应用监控； ● MapReduce 编程模型、Shuffle 过程、编程调优 ● 分布式部署 Hadoop 2.x； ● 分布式协作服务框架 Zookeeper； ● HDFS HA 架构、配置、测试； ● HDFS 2.x 中高级特性； ● YARN HA 架构、配置； ● Hadoop 主要发行版本（CDH、HDP、Apache）
数据仓库技术 与应用	Hive	● Hive 功能、体系结构、使用场景； ● Hive 环境搭建、初级使用； ● Hive 原数据配置、常见交互方式； ● 内部表、外部表、分区表； ● 数据迁移； ● 常见查询； ● 内置函数和 UDF 编程 ● Hive 数据的存储和压缩； ● Hive 常见优化（数据倾斜、压缩等）
数据采集与迁 移	Sqoop	● Sqoop 功能、使用原则； ● 将 RDBMS 数据导入 Hive 表中（全量、增量）； ● 将 HDFS 上文件导出到 RDBMS 表中
	Flume	● Flume 设计架构、原理； ● flume 初步使用，实时采集数据； ● 如何使用 Flume 监控文件夹数据，实时采集录入 HDFS 中。

大数据存储与处理技术	Hbase	● Schema、表的设计； ● 环境搭建、shell 初步使用（CRUD 等） ● 数据存储模型； ● Java API 使用（CRUD、SCAN 等）； ● 架构深入剖析； ● 与 MapReduce 集成、数据导入导出。 ● 表及其预分区设计； ● 设置表属性； ● admin 操作。
大数据缓存系统	redis	● redis 和 nosql； ● 客户端连接； ● 数据结构 string、list、hash、set、sortedset 操作。
程序设计（scala）	scala	● 基础语法； ● 条件控制； ● 函数；数组；Map；Tuple； ● 面向对象； ● 函数式编程； ● 模式匹配； ● 类型参数； ● 隐式转换及参数； ● 多线程 Actor
内存高性能计算	Spark	● Spark 生态系统及 mapreduce 比较； ● 基本知识； ● 安装部署； ● RDD 及特性、常见操作、缓存策略； ● RDD Dependency、Stage 常、源码分析； ● 核心组件； ● Spark on YARN； ● historyServer
	Spark Streaming	● Spark Streaming 原理及流式计算； ● Dstream 设计； ● input 及 out
	Spark 高级应用	● Scala 编程； ● Hadoop 与 Spark 集群搭建； ● Spark 核心编程； ● Spark 性能调优； ● Spark 源码剖析；
图形化	SparkR	● Spark R 安装测试 ● 使用 spark-submit 向集群提交 R 代码文件 dataframe.R ● Eclipse 下 R 语言环境搭建 ● R 语言函数调用（跨文件调用） ● Sparklyr 学习

机器学习	SparkMLlib	● Spark MLlib 介绍 ● Spark MLlib 矩阵向量 ● Spark MLlib 线性回归算法 ● Spark MLlib 逻辑回归算法 ● Spark MLlib 贝叶斯分类算法 ● Spark MLlib 决策树算法 ● Spark MLlib KMeans 聚类算法 ● Spark MLlib FPGrowth 关联规则算法 ● Spark MLlib 协同过滤推荐算法 ● Spark MLlib 神经网络算法
数据挖掘与分析	数据挖掘	● 数据仓库原理及联机分析技术介绍； ● 数据仓库设计与开发； ● 基于数据仓库的决策支持系统； ● 数据仓库案例剖析； ● 数据挖掘与知识发现； ● 统计分析、方差分析、关联分析、回归分析、因子分析、聚类分析算法 ● 决策树算法、神经网络算法

(三) 大数据课程部分实验实训内容

课程分类	课程名称	实验内容
操作系统与 linux 应用	Linux 基础	● 免密登陆配置
程序设计与应用（python）	Python 基础与网络爬虫	● Scrapy 框架的安装 ● Scrapy 爬取某体育网站的信息 ● Scrapy 爬取某岗位招聘信息
大数据存储与处理技术应用	HDFS	● 开发 shell 采集脚本 ● HDFS Java API
大数据 计算与分析	Hadoop	● Hadoop 伪分布式模式 ● Hadoop 完全分布式模式 ● MapReduce-去重、排序、平均值 ● MapReduce-二次排序、倒排索引 ● MapReduce-Join ● MapReduce-社交好友推荐 ● MapReduce-互联网精准广告推送 ● Zookeeper 安装 ● Zookeeper 集群安装
数据仓库技术与应用	Hive	● Hive 环境安装 ● Hive 基本查询 ● Hive 排序 ● Hive 定义函数
数据采集与迁移	Sqoop/Flume	● Sqoop 安装 ● Sqoop 增量、全量导入 ● Flume 环境安装 ● Flume 源、通道、沉槽

大数据存储与处理技术	Hbase	● Hbase 环境安装 ● Hbase Java API ● Hbase 比较过滤器、专用过滤器
大数据缓存系统	Redis	● 任务调度队列 ● 购物车 ● 排行榜
Spark 内存高性能计算	Spark 基础和 Spark 高级应用	● Spark 单机模式安装 ● Spark standalone 模式安装 ● Spark shell 操作 ● Spark Java/Scala API ● Spark SQL 操作
数据挖掘与分析	spss	● 方差分析 ● 相关分析 ● 回归分析 ● 聚类分析 ● 因子分析

四、 大数据实验实训资源库介绍

(一) 大数据实验室实验实训资源建设

学生在实训过程中需要有足够资源支撑。大数据实验室实验实训资源建设基本包括该专业全部核心课程，实训资源所涉及到的案例均由企业中真实项目构成，经经验丰富的算法专家和教育专家分析、总结，融入实训资源体系中，并设置不同阶段、不同难易程度的教学目标。

校内实践应该贯穿整个教学环节，从低到高，由浅入深让学生具备扎实的实践动手能力。建立综合开发资源库，为师生提供教学、学习中所需要的各类资源；实训环节中，减少教师实训工作量的同时，给学生更为清晰的实训体验以及更为详尽的实训内容；在内容方面，提供更为全面的技术方向，使得学生能够人工智能领域的基础理论和实现方法。实训平台升级计划将会严格遵循“教训一体化”的原则， 以打造“教、训、测”全方位立体化平台为目标。平台将坚持以“实训”为中心， 统筹安排教学、实训过程；以“新工科”为导向，提供更多市场需要的主流开发技术教学、实训内容。

专业综合实训是教学环节最关键的环节。项目实训过程整体应该包括：项目计划、需求分析、设计、编码、测试、文档工作及部署工作。相应的资源应该包括项目详细需求，工期要求，建议分工，工作任务书相应的内容。为了提高学生综合知识运用能力与团队合作能力，ZX人工智能实训平台在公有云中部署了采集自真实企业环境下的海量数据，供学生综合运用专业知识完成大数据环境的开发实训。ZX-AI平台中部署了《Python 开发》、《数据挖掘》、《机器学习》、《神经网络与深度学习》、《Python 与数据分析》等基础课程实验和实训。在大数据技术层面部署了《Flume 数据采集》《Sqoop》、《HDFS 分布式文件存储》《Hive 数据仓库》《Shell

编程》等完全覆盖该专业的基础和核心课程。实训环节中，设置了能够给学生模拟智能企业的工作环境，极大缩短学生就业的适应期，能够快速胜任人工智能类相关工作。为了保持算法和实训数据的时代特征，平台将会定期更新实训案例。下面将以几门课程为例说明实验和实训内容。

(二)《机器学习》课程实训环节（部分）

（1） 课程实训一：鸢尾花识别
（2） 课程实训二：鲍鱼年龄预测
（3） 课程实训三：乳腺癌病情分类
（4） 课程实训四：隐形眼镜类型
（5） 课程实训五：病马死亡率预测 〞
（6） 课程实训六：鸢尾花分类
（7） 课程实训七：心血管病特征降维
（8） 课程综合实训项目：城市房价预测

(三)《Hadoop 大数据》综合实训项目（部分）

（1） 实训 1、二度人脉与好友推荐
（2） 实训 2、网站分析
（3） 实训 3、雾霾统计分析

(四) 大数据综合实训项目示例—电子商务网站商品精准推荐（部分）

随着电子商务和社会经济的发展，人们的生活习惯有了较大变化，越来越多的企业通过电子商务进行交易、结算等商务活动，在网上进行消费、投资活动的人群也逐年增多，人们的日常活动，包括食品、衣物、旅行、票务预订、教育等活动都可以通过网络得到满足，电子商务越来越紧密的和消费者结合。截至 2017 年6 月，中国电子商务交易额超过 7 万亿元，互联网用户超 7 亿，网购用户数量超过 3.1 亿人。但是虽然每天都有数以亿计的消费者在网络中进行购物活动，但网络中亦有数以万计的商家在网络中从事商业活动，对于每一个商家而言，如何对市场进行划分， 精准的进行产品、市场进行定位，抓住老客户，发现新客户，在众多的商家中脱颖而出成为摆在每个商家面前的问题。

实际研究认为精准推荐是通过定量和定性相结合的方法对目标市场的不同消费者进行细致分析，根据他们不同的消费心理和行为特征，采用有针对性的现代技术、方法和指向明确的策略，实现对目标市场不同消费者群体强有效性、高投资汇报的营销沟通。在电子商务网站中，经常需要针对不同的客户进行商品推荐，缩减客户搜索成本，提高客户体验，提高网站流量的转化率，提高营收。根据商品推荐的对

象来分，可以分为面向浏览用户的推荐和面向登录用户的推荐两种，面向浏览用户的推荐往往是常规推荐，其指的是符合常规商品关联逻辑的一些推荐，面向登录用户的推荐往往是个性化推荐，是指基于购买行为间关联性归纳出的商品推荐。针对的用户在电子商务网站上的商品评分行为进行分析，依据海量数据，研究用户兴趣偏好，分析用户的需求和行为；从商品的角度出发，分析用户对商品的购买偏好。设计协同过滤算法，将商品准确推荐给所需用户。项目的业务系统底层主要采用 Python 架构，大数据主要采用 Hadoop+Spark 框架以及各类协同推荐算法。

(五) 大数据综合实训项目示例二-某互联网金融产品网络舆情监测

网络爬虫技术、大数据存储技术、大数据计算技术在商业活动中的应用，对金融产品而言，大数据技术的应用将助力金融产品实现四个“最了解”，即“最了解自身”、“最了解客户”、“最了解竞争对手”和“最了解经营环境”，具体应用场景如：网络舆情监测、客户全景画像、竞争对手分析、行业垂直搜索等。

网络舆情是当前社会主流舆论的表现方式之一，它主要搜集和展示经互联网传播后大众对部分社会焦点和热点问题的观点和言论。对于金融产品而言，对网络舆情进行监测，是对自身品牌管理和危机公关的重要技术手段，从而以网络作为一面“镜子”，构建“最了解自身的个体”。

网络舆情作为当前社会的主流信息媒介之一，具有传播快、影响大的特点，对于金融产品而言，创建自动化的网络舆情监控系统十分必要，一方面可以使金融产品获得更加精准的社会需求信息，另一方面可以使金融产品在新的舆论平台上传播自身的服务理念和服务特色，提升自身的业务拓展水平。大数据技术的使用使得网络舆情监测更加方便、及时、全面。依据本金融产品产品相关主题信息，使用多线程网络爬虫技术 scrapy，爬取互联网社会焦点、社会热点等中的和本产品相关信息，爬取的信息存放到分布式文件存储系统 HDFS，使用 yarn 协调资源，运用 spark 内存计算框架，实时分析计算和本产品相关的舆论信息，使用分布式面向列数据库 Hbase，和缓存数据库 redis。实时的分析统计相关的舆论信息指标，如果舆论对自己的品牌有异议，那么就会及时做出公关应对；如果对自己的品牌比较认同，那么就会做出加大投入资源的决策。

通过运用大数据技术对网络舆情进行监测，可以更加全面深入地了解客户对个体的态度与评价，洞察个体自身经营的优势与不足，同时可以起到防御声誉风险、增强品牌效应的作用。

(六)　大数据综合实训项目示例三-教育平台日志分析

日志在计算机系统中是一个非常广泛的概念，任何程序都有可能输出日志：操作系统内核、各种应用服务器等等。日志的内容、规模和用途也各不相同。互联网应用在运营的时候会产生大量的系统日志。每条日志通常代表着用户的一次访问行

为，例如下面的 apache 日志：

211.87.152.44 – - [18/Mar/2005:12:21:42 +0800] “GET / HTTP/1.1″ 200 899 “http://www.baidu.com/” “Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Maxthon)”

从上面这条日志中，我们可以得到很多有用的信息，例如访问者的 IP、访问的时间、访问的目标网页、来源的地址以及访问者所使用的客户端的 UserAgent 信息等。如果需要更多的信息，则要用其它手段去获取：例如想得到用户屏幕的分辨率， 一般需要使用 js 代码单独发送请求；而如果想得到诸如用户访问的具体新闻标题等信息，则可能需要 Web 应用程序在自己的代码里输出。

随着日志量的增大，企业利用大数据技术处理这些日志，可以从中获取大量信息中分析出各种有用的信息比如：平台访问时间段、平台课程访问对比、平台课程综合活跃度、平台课程反馈意见分析、网络爬虫分析等等，为企业的互联网应用决策提供数据依据。首先是日志采集器 Flume 把日志采集到文件存储系统 HDFS，然后通过 Spark RDD,Spark SQL,MapReduce，Hive 对数据进行处理，把处理的数据传递到 Mysql，最后 web 展现给用户。