泰迪智能科技
服务热线：40068-40020

1+X 大数据应用开发（Python）职业技能等级证书实训基地建设方案

广东泰迪智能科技股份有限公司

2020 年 10 月

目录

1. 方案概述

1.1. 政策背景介绍

2019 年 1 月 24 日“国务院关于印发国家职业教育改革实施方案的通知（国发〔2019〕4 号）”，发布《国家职业教育改革实施方案》（职教 20 条），其中第六条为“启动 1+X 证书制度试点工作”，提出“深化复合型技术技能人才培养培训模式改革，借鉴国际职业教育培训普遍做法，制定工作方案和具体管理办法，启动 1+X 证书制度试点工作”。结合学历证书与职业技能等级证书，探索实施 1+X 证书制度，是职教 20 条的重要改革部署，也是重大创新。《方案》提到在职业院校、应用型本科高校启动“学历证书+若干职业技能等级证书”（即 1+X 证书）制度试点，旨在鼓励学生在获得学历证书的同时，积极取得多类职业技能等级证书。同年，《政府工作报告》进一步指出，“要加快学历证书与职业技能等级证书的互通衔接”。

2020 年 9 月 23 日，1+X 证书制度试点第四批职业教育培训评价组织和职业技能等级证书公示，其中广东泰迪智能科技股份有限公司申请的**“大数据应用开发（Python）”**赫然在列。

本标准由广东泰迪科技股份有限公司牵头发起，数十家大数据技术企业、行业协会、院校单位及专家学者的广泛参与、支持与推荐，具有较高的权威性和先进性。标准依据行业现状和岗位技能的要求，将本职业技能分为初级、中级和高级三个等级，**针对商务数据分析、大数据管理、大数据应用开发及人工智能等各方面的工作岗位需求**，提供以**广泛通用的的技术 Python 为核心**的基础和实践技能技术，便于培训、易于评价。直接面向新兴（大）数据职业岗位及岗位群：大数据分析师、数据分析专员、算法工程师、开发工程师、运维工程师等。

针对 1+X 证书制度，泰迪公司组织行业内专家和院校资深教师深度解析大数据应用开发（Python）职业技能等级技能标准要求，为应用型本科及职业院校学生获取 1+X 职业技能等级证书提供优质**完善的实训环境建设建议并配套丰富而精准的教学资源**。利用公司从事高校大数据教育行业多年的丰富教学经验及先进的教学理念，为学校提供符合行业人才技能需求的人才培养解决方案专业建议。

1.2. 行业背景介绍

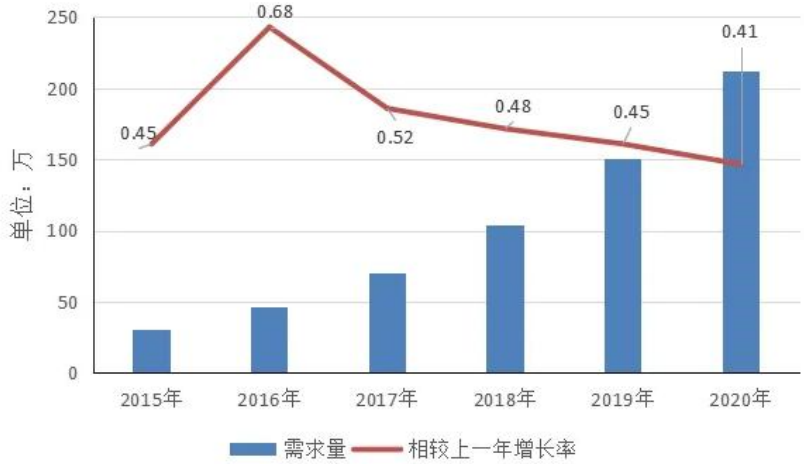
当前信息化对人类经济活动产生深刻影响，正渗透到生产生活方方面面，数据已经成为新的生产要素，大数据行业已成为人们按需使用信息处理、信息存储、信息交互资源的重要模式，也是进行大数据处理和深度挖掘的重要平台，大数据工程技术人员在我国现阶段及未来发挥的作用将日益凸显。

《大数据产业发展规划（2016-2020 年）》指出，目前大数据人才队伍建设亟需加强，大数据基础研究、产品

研发和业务应用等各类人才短缺，难以满足发展需要。要建设多层次人才队伍，建立适应大数据发展需求的人才培养和评价机制。加强大数据人才培养，整合高校、企业、社会资源，推动建立创新人才培养模式，建立健全多层次、多类型的大数据人才培养体系。

根据天府大数据国际战略与技术研究院(简称“天府大数据研究院”)《2018 全球大数据发展分析报告》数据，2018 年我国大数据产业人才占整体就业人口规模的 0.23%，大约 179.4 万人。

猎聘《2019 年中国 AI&大数据人才大数据人才就业趋势报告》指出，2019 年中国大数据人才缺口高达 150 万。另据中国商业联合会数据分析专业委员会统计，未来中国基础性数据分析人才缺口将达到 1400 万。



随着大数据、物联网、5G 等技术应用的不断发展，社会对该职业从业人员的需求日益增长。预计 2020 年中国大数据行业的人才需求规模将达到 210 万，2025 年前大数据人才需求仍将保持 30%-40%的增速，需求总量在 2000 万人左右。

1.3. 1+X 制度下院校任务与具体工作

对于院校而言，首先需要将 1+X 作为一项重大改革和制度设计，此过程的主要目标如下。

（1） 将 1+X 证书制度试点与专业建设、课程建设、教师队伍建设等紧密结合，推进“1”和“X”的有机衔接，提升职业教育质量和学生就业能力。

（2） 通过试点，深化教师、教材、教法“三教”改革，促进校企合作；

（3） 建好用好实训基地；

（4） 探索建设职业教育国家“学分银行”，构建国家资历框架。

其次，需要开展 1+X 职业技能教育的具体工作，工作内容主要包括如下内容。

（1） 选择有关职业技能等级证书，确定参与试点的专业。

（2） 统筹专业(群)资源，深入研究职业技能等级标准与有关专业教学标准，推进“1”和“X”的有机衔接，将证书培训内容及要求有机融入专业人才培养方案，优化课程设置和教学内容，加强专业教学团队建设，选派教师参加有关培训。

（3） 根据在校学生取证需要，对专业课程未涵盖的内容或者需要特别强化的实训，在培训评价组织支持下，组织开展专门培训，同时可面向社会成员开展培训。

（4） 符合条件的院校按程序申请设立为考核站点，配合培训评价组织实施证书考核。

（5） 管理和使用好有关经费。

1.4. 泰迪科技支持

作为 1+X 《大数据应用开发（Python）职业技能等级标准》评价组织，广东泰迪智能科技股份有限公司结合在专业及课程建设方面具有丰富的行业经验，可协助学校针对 1+X《大数据应用开发（Python）职业技能等级证书》培训内容进行人才培养方案修订、课证融通，提供专业课程资源，为学校 1+X 制度下的人才培养软实力提升提供帮助。

1.4.1. 专业及课程设置

（1） 制定人才培养方案：针对 1+X《大数据应用开发（Python）职业技能等级标准》要求，面向院校新一代信息技术领域专业制定人才培养方案，将标准要求的专业知识**融入到人才培养方案**中。

（2） 修订专业课程，融合职业技能等级标准：通过修订课程大纲，建设课程资源包形式，优化课程设置和教学内容，将 1+X 证书培训内容及要求有机融入课程教学中，实现**课证融通**。

（3） 教材研发：在职业技能等级证书的基础上，联合人民邮电出版社，开发 1+X《大数据应用开发（Python）职业技能等级标准》**配套教材**。

（4） 课程资源配套：在配套教材的基础上，泰迪公司联合业内专家，配置对应的课程资源，包括视频、PPT、实训指导书、代码、习题库等。

1.4.2. 实训基地建设

依托公司在大数据行业的积累，精心提供 1+X《大数据应用开发（Python）职业技能等级标准》实训基地建设意见，为学校进行 1+X 制度下的实训基地建设提供设备选型参考。

1.4.3. 师资建设

泰迪智能科技在百余个企业项目，特别是多年与高校合建大数据教学实训平台中，积累了丰富的项目经验和师资、人才培养经验，以总经理张良均为代表的大数据培训团队，既有项目实战经验，也有教学、培训资深经历，把大数据的知识、技术、项目经验等，以案例为导向，深入浅出、融会贯通，融入到教学 PPT、视频、使用指导等教学资源中，更多的是把大数据教学的方式、方法和手段，通过授课等方式进行智力传递。

结合 1+X《大数据应用开发（Python）职业技能等级标准》，泰迪科技更是制定了专业的师资培训方案，其主要内容包括了依托行业领先企业的产业优势、资源优势、技术优势，派驻高校教师用半年至一年的时间到企业学习先进技术，提升高校教师的实践能力和知识水平；或者以项目型驻校工程师团队，而非课程型工程师团队的组态形式进行教学。

2. 标准先进性与证书特色

《大数据应用开发（Python）职业技能等级标准》由广东泰迪科技股份有限公司牵头发起，得到华为技术有限公司、蓝盾信息安全技术股份有限公司、广州智能装备研究院有限公司、中国联合网络通信股份有限公司、人民邮电出版社有限公司、网宿科技股份有限公司、广州思迈特软件有限公司、广州粤嵌通信科技股份有限公司、广州佰聆数据股份有限公司、深圳市怡亚通供应链股份有限公司、电子工业出版社有限公司、广东省人才研究会、中山大学、深圳职业技术学院、广州番禺职业技术学院等数十家大数据技术企业、行业协会、院校单位及专家学者的广泛参与支持。

相比于其他的 1+X 职业技能等级证书，泰迪科技具有证书定位、课证融通、赛证融通三方面的先进性。

2.1. 证书定位及特色

首先是证书定位，《大数据应用开发（Python）》证书是以 Python 技术为主线，结合企业大数据应用开发场景制定的人才培养等级评价标准。证书主要面向中等职业院校、高等职业院校和应用型本科院校的数据分析、大数据、软件和计算机相关专业。通过学习，一方面可以掌握大数据平台与系统的搭建、配置、操作、监控、优化；另一方面可以掌握数据处理、数据分析与挖掘、数据可视化、文本挖掘、深度学习；同时还能够具备一定的项目管理能力。

泰迪科技大数据应用开发（Python）基于企业多年大数据挖掘项目实践制定标准，贴合企业实际应用，不空泛、重技能。

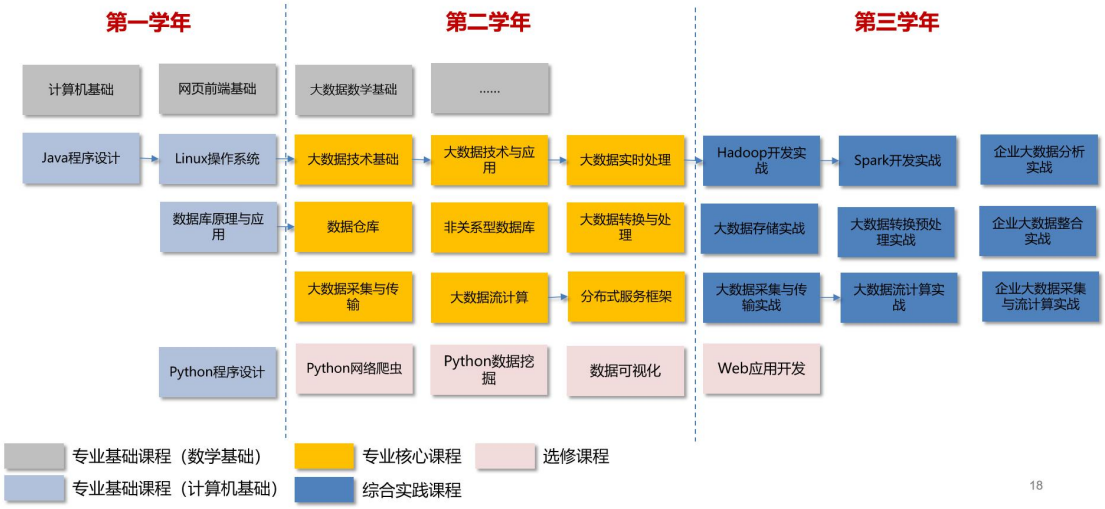
大纲内容与教学课程教材无缝对接，实现真正意义上的课证融通及书证融通，1+x 职业能力等级试点考点的建设不再只是为了考证，而是借助考证把师资队伍培养成为贴合实际企业应用的双师人才，学生的学习也与产业实际需求对接，极大提升未来就业竞争力。

图 2-1 1+x 职业等级认证系列图书教材

教材内容注重项目应用的实战，证书培训不讲空洞理论，用实际案例应用提升学生应用技能水平，强调“用以致学”证书应用理念。

2.2. 课证融通

课证融通指的是将 1+X 标准与专业课相结合，通过专业课的学习即可参加并通过 1+X 的考试。以高职院校的大数据技术与应用专业为例，其专业课程体系大体结构如下：



大数据应用开发（Python）职业技能等级中的中级对应的知识点与对应课程如下表所示。

表 2-1 大数据应用开发（Python）初中高对应课程清单

技能等级	课程类别	课程名称
初级	课程	Excel 数据分析基础与实战 Power BI 数据分析与可视化 Python 编程基础
	案例	新零售智能销售数据分析 航空公司客户价值分析 学生校园消费行为分析 疫情期间湘鄂省区物流数据分析 百货商场用户画像描绘与价值分析 浙江省区采购数据分析 福建省区酒饮退货数据分析 供应链商品经营数据分析 供应链销售数据分析 广东省区采购数据分析 黑吉省区家电销售发货数据分析 江浙两省日化销售数据分析 陕西省销售数据分析 豫冀两省母婴销售数据分析
中级	课程	Python 数据分析与应用 Python 数据可视化 Python 数据分析实训 Hadoop 大数据开发基础

	案例	市财政收入分析预测 Python 网络爬虫技术 热门电影影评数据爬取及分析 网络招聘数据采集与大数据人才需求分析 信用卡高风险客户识别 金融服务机构资金流量预测 运营商流失用户分析 基于基站定位数据的商圈分析 金融理财的广告牌精准投放 二手汽车销售售价预测 汽车用户销售投诉数据爬取 二手车交易数据爬取 人口增长与医疗需求预测
高级	课程	Python 数据分析与挖掘 特征工程实践 TensorFlow2 实战 深度学习原理及编程实现 文本挖掘实战
	案例	垃圾短信智能识别 利用循环神经网络（RNN）对路透社新闻进行分类 动态人脸智能识别 基于深度学习的推荐系统受众性别预测 020 优惠券使用预测 车牌智能识别 家用热水器用户行为分析与事件识别 城市公交站点设置的优化分析 电力窃漏电用户识别 广电大数据营销推荐项目实战 电子商务网站用户行为分析及服务推荐 电商产品评论数据情感分析 基于水色图像的水质识别 消费者投诉举报信息意见挖掘 P2P 网络信贷获贷结果预测 中医证型的关联规则挖掘 血管三维重构

大数据技术与应用专业的学生通过相关课程的学习，就可以考取大数据应用开发（Python）相应等级证书。

2.3. 赛证融通

赛证融通指的“泰迪杯”大数据竞赛与证书知识点相结合，通过参与比赛模拟大数据应用开发（Python）职业技能的考核。

目前泰迪科技的“泰迪杯”大数据竞赛分为数据挖掘挑战赛和数据分析技能赛（详见：www.tipdm.org）两个子赛项，其中数据挖掘挑战赛所需的技能要求对应《大数据应用开发（Python）职业技能等级标准》中的高级，数据

分析技能赛所需的技能要求对应《大数据应用开发（Python）职业技能等级标准》中的中级和初级。

3. 标准解读

3.1. 对应专业

（1） **中等职业学校：**大数据应用与技术、软件开发技术、计算机网络技术、数据商务、数据管理、商务数据分析与应用、软件与信息服务、计算机应用、通信技术、电子信息技术、云计算技术。

（2） **高等职业学校：**大数据技术与应用、商务数据分析与应用、人工智能技术服务、智能产品开发、信息统计与分析、统计与会计核算、软件与信息服务、移动应用开发、云计算技术与应用、移动互联应用技术。

（3） **应用型本科学校：**数据科学与大数据技术、大数据管理与应用、人工智能、计算机科学与技术、应用统计学、数学与应用数学、信息与计算科学、软件工程。

3.2. 面向工作岗位（群）

大数据应用开发（Python）职业技能等级分为三个等级：初级、中级、高级。三个级别依次递进，高级别涵盖低级别职业技能要求。

（1） 【大数据应用开发（Python）】（初级）：主要面向互联网企业以及向互联网转型的政府、企事业单位的基础设施管理、应用软件开发部门，从事数据分析师、爬虫工程师、数据可视化工程师等工作岗位，能根据业务要求完成数据采集、数据处理、数据分析、数据可视化等工作任务。

（2） 【大数据应用开发（Python）】（中级）：主要面向互联网企业以及向互联网转型的政府、企事业单位的大数据应用软件开发部门，从事数据分析师、数据可视化工程师、大数据开发工程师、项目经理等工作岗位，能根据业务要求完成数据挖掘、数据可视化、基础项目管理等工作任务。

（3） 【大数据应用开发（Python）】（高级）：主要面向互联网企业以及向互联网转型的政府、企事业单位的大数据应用软件开发部门，从事算法工程师、大数据开发工程师、文本挖掘工程师、项目经理等工作岗位，能根据业务要求完成数据挖掘、文本挖掘、项目管理等工作任务。

3.3. 教学实训形式及内容

3.3.1. 初级

3.3.1.1.理论教学

工作领域	工作任务	职业技能要求	教学目标
1.平台管理	1.1 软件安装	1.1.1 能够根据操作规范，独立完成 Linux 系统的安装 1.1.2 能够根据操作规范，完成 Python 环境安装 1.1.3 能够根据操作规范，完成常见关系型数据库的安装 1.1.4 能够根据操作规范，完成数据库管理工具的安装 1.1.5 能够根据操作规范，完成数据可视化工具的安装	（1） 了解 Linux 操作系统 （2） 熟悉 Python 及其数据分析库 （3） 熟悉常见关系型数据库 （4） 熟悉常见数据可视化工具
	1.2 软件管理	1.2.1 能够根据操作规范，进行关系型数据库用户管理 1.2.2 能够根据操作规范，进行关系型数据库权限管理 1.2.3 能够根据操作规范，在线扩展 Python 第三方库 1.2.4 能够根据操作规范，离线扩展 Python 第三方库	（1） 掌握关系型数据库用户管理的方法 （2） 掌握关系型数据库权限管理的方法 （3） 掌握 Python 库的安装方法
	1.3 系统管理	1.3.1 能够根据操作规范，进行 Linux 系统用户管理 1.3.2 能够根据操作规范，进行 Linux 系统权限管理 1.3.3 能够根据操作规范，进行 Linux 系统的内存管理 1.3.4 能够根据操作规范，进行 Linux 系统的状态监控	（1） 掌握 Linux 用户管理的方法 （2） 掌握 Linux 权限管理的方法 （3） 掌握系统运行状态监控的方法与工具
2.数据采集与存储	2.1 数据质量评估	2.1.1 能够根据业务需求及数据质量标准，进行数据规范性评估 2.1.2 能够根据业务需求及数据质量标准，进行数据完整性评估 2.1.3 能够根据业务需求及数据质量标准，进行数据准确性评估 2.1.4 能够根据业务需求及数据质量标准，进行数据一致性评估 2.1.5 能够根据业务需求及数据质量标准，进行数据时效性评估 2.1.6 能够根据数据质量评估结果，独立完成数据质量评估报告	（1） 熟悉数据质量评估的指标 （2） 熟悉数据质量评估报告的构成
	2.2 数据采集	2.2.1 能够根据业务需求，制定网页数据采集方案 2.2.2 能够根据业务需求，进行网址分析、网页分析 2.2.3 能够根据业务需求，使用 Python 采集网页数据 2.2.4 能够根据业务需求，存储采集的结构化数据	（1） 熟悉爬虫的流程 （2） 掌握谷歌开发者工具的基础用法 （3） 掌握网址分析的基础方法 （4） 掌握网页构成与分类 （5） 掌握 Xpath 解析网页的方法

	2.3 数据存储	2.3.1 能够根据业务需求，选择数据库管理工具 2.3.2 能够根据业务需求，将结构化文件数据导入关系型数据库 2.3.3 能够根据业务需求，导出关系型数据库数据为结构化文件 2.3.4 能够根据业务需求，运用 SELECT 语句实现数据查询	（1） 掌握常见数据库管理工具的用法 （2） 掌握基础 SQL 语句的用法
3.数据分析与可视化	3.1 数据处理	3.1.1 能够根据业务需求与数据现状，进行数据中缺失值的识别与处理 3.1.2 能够根据业务需求与数据现状，进行数据中异常值的识别与处理 3.1.3 能够根据业务需求与数据现状，进行数据的其他清洗操作 3.1.4 能够根据业务需求与数据现状，进行数据合并	（1） 掌握缺失值的识别与处理方法 （2） 掌握异常值的识别与处理方法 （3） 掌握数据合并的方法
	3.2 数据分析	3.2.1 能够根据业务需求与数据现状，进行描述性统计分析 3.2.2 能够根据业务需求与数据现状，进行相关性分析 3.2.3 能够根据业务需求与数据现状，进行对比分析 3.2.4 能够根据业务需求与数据现状，进行交叉分析	（1） 掌握常见的描述性分析方法 （2） 掌握常见的相关性分析方法 （3） 掌握对比分析、交叉分析等其他数据分析方法
	3.3 数据可视化	3.3.1 能够根据业务需求，选择数据可视化工具 3.3.2 能够根据业务需求，使用数据可视化工具对数据进行基本的操作与配置 3.3.3 能够根据业务需求，绘制基础的可视化图形 3.3.4 能够根据业务需求，辅助业务人员完成数据可视化大屏	（1） 掌握散点图及其相关图形的绘制 （2） 掌握折线图及其相关图形的绘制 （3） 掌握面积图及其相关图形的绘制 （4） 掌握柱形图及其相关图形的绘制 （5） 掌握饼图及其相关图形的绘制 （6） 掌握仪表盘、雷达图等其他图形的绘制

3.3.1.2.实验教学

序号	实验名称	实验要求
1	新冠疫情数据采集	能够掌握网页网址分析方法 能够掌握 Xpath 解析网页方法 能够掌握数据导入关系型数据库的方法
2	学生校园消费行为分析	能够掌握数据清洗的方法 能够掌握数据合并的方法 能够掌握对比分析
3	百货商场用户画像描	能够掌握用户画像绘制的方法

	述	能够掌握数据预处理的方法 能够掌握描述性分析
4	自动售货机数据可视化大屏	能够掌握可视化图形的绘制方法 能够掌握数据可视化工具的使用方法

3.3.2. 中级

3.3.2.1.理论教学

工作领域	工作任务	职业技能要求	教学目标
1. 平台管理	1.1 软件安装	1.1.1 能够根据操作规范，完成 Linux 系统集群的安装与配置 1.1.2 能够根据操作规范，完成 Hadoop、Storm、Spark 大数据系统的安装与配置 1.1.3 能够根据操作规范，完成 IDE 集成开发环境的安装与基础配置 1.1.4 能够根据操作规范，完成分布式数据库、分布式文件系统的安装与基础配置 1.1.5 能够根据操作规范，完成 ETL 工具的配置与安装	（1）掌握 Linux 集群系统的配置过程 （2）掌握 Hadoop、Spark 大数据组件的配置过程 （3）掌握 Eclipse、IntelliJ、PyCharm 开发工具的安装配置、基本功能
	1.2 软件管理	1.2.1 能够根据操作规范，对大数据平台进行状态监控、异常分析 1.2.2 能够根据监控与分析结果，对常见问题进行处理 1.2.3 能够根据操作规范，完成大数据平台的升级操作 1.2.4 能够根据操作规范，完成大数据组件性能优化	（1）了解集群系统和大数据平台的管理 （2）掌握 Linux 系统文件管理与编辑方法 （3）掌握 Linux 系统中压缩与解压方法 （4）掌握 Linux 网络的设置与维护方法
	1.3 系统管理	1.3.1 能够根据操作规范，完成 Linux 系统文件管理与编辑 1.3.2 能够根据操作规范，完成 Linux 系统压缩与解压 1.3.3 能够根据操作规范，完成 Linux 系统的磁盘管理与维护 1.3.4 能够根据操作规范，完成 Linux 系统的网络设置与维护	
2. 数据采集与存储	2.1 数据采集	2.1.1 能够根据业务需求，进行终端协议分析 2.1.2 能够根据业务需求，进行客户端数据采集 2.1.3 能够根据业务需求，进行手机 APP 数据采集 2.1.4 能够根据业务需求，完成采集的非结构化数据的存储	（1）掌握采集客户端数据的方法 （2）掌握采集 APP 数据的方法
	2.2 数据存储	2.2.1 能够根据业务需求，运用 SQL 语句实现常规数据查询 2.2.2 能够根据业务需求，进行关系型数据库性能优化 2.2.3 能够根据业务需求，进行关系型数据库的备份与恢复 2.2.4 能够使用 Python 访问关系型数据库，进行数据操作	（1）掌握常见 SQL 高级语句的使用方法 （2）掌握优化数据库性能的方法 （3）掌握备份与恢复数据库的方法

	2.3 数据整合	2.2.1 能够根据业务需求，选择不同的 ETL 工具 2.2.2 能够根据业务需求，实现关系型数据库数据的抽取 2.2.3 能够根据业务需求，实现本地文件数据的抽取 2.2.4 能够根据业务需求，将数据装载至关系型数据库	（1）掌握 ETL 的各个组成的作用 （2）掌握 ETL 常见工具的功能和特点 （3）掌握 ETL 方案设计准则
3. 数据分析与可视化	3.1 数据处理	3.1.1 能够根据业务需求与数据现状，进行数据标准化处理 3.1.2 能够根据业务需求与数据现状，进行离散化处理 3.1.3 能够根据业务需求与数据现状，进行独热编码处理 3.1.4 能够根据业务需求与数据现状，进行业务指标构建	（1）掌握常见数据变换方法 （2）掌握常见的分类算法 （3）掌握常见的聚类算法 （4）掌握常见的回归算法 （5）掌握常见的智能推荐算法 （6）掌握常见的关联规则算法
	3.2 数据挖掘	3.2.1 能够根据业务需求，构建分类模型 3.2.2 能够根据业务需求，构建聚类模型 3.2.3 能够根据业务需求，构建回归模型 3.2.4 能够根据业务需求，构建智能推荐模型 3.2.5 能够根据业务需求，构建关联规则模型	
	3.2 数据可视化	3.3.1 能够根据业务需求使用数据可视化工具将数据以图表的形式进行展示 3.3.2 能够根据业务需求，在业务主管的指导下根据数据分析可视化结果，形成有效的数据分析报告 3.3.3 能够通过数据分析可视化结果，得出有效的分析结论 3.3.4 能够根据业务需求，实现数据可视化大屏设计	（1）掌握常见可视化工具的功能和基础应用 （2）掌握可视化报告的书写规范和准则
4. 项目管理	4.1 需求管理	4.1.1 能够根据项目现状，制定需求管理计划 4.1.2 能够收集业务需求并进行整理归档 4.1.3 能够根据业务，使用常用的需求收集工具与技术 4.1.4 能够根据业务，筛选需求	（1）熟悉项目管理内容 （2）熟悉需求管理过程和收集需求流程 （3）熟悉进度管理过程和常用工具 （4）熟悉变更管理的规则和内容
	4.2 进度管理	4.2.1 能够根据业务需求，对项目活动进行排序 4.2.2 能够根据业务需求，对项目活动所需资源进行规划 4.2.3 能够根据业务需求，制定项目进度计划 4.2.4 能够根据计划执行情况，调整项目计划	
	4.3 变更管理	4.3.1 能够根据项目变更原则，制定研发计划 4.3.2 能够根据项目变更工作流程，推动项目研发 4.3.3 能够根据项目需求，控制变更频次 4.3.4 能够根据项目需求，制定版本发布和回退计划	

3.3.2.2.实验教学

序号	实验名称	实验要求
1	大数据平台配置	1. Linux 集群系统安装 2. Linux 集群系统配置 3. Hadoop 分布式安装和配置 4. Spark 分布式安装和配置
2	实时商务日志采集系统	1. 能够根据采集需求设计数据采集方案 2. 能够通过数据采集工具配置数据采集流程

		3. 安装数据采集工具，实现数据采集到 mysql
3	基于 Hive 的数据 ETL	1. 能够掌握 Hive 的基本操作 2. 能够根据业务需求分析 ETL 流程
4	广电大数据用户智能推荐	1. 能够掌握数据探索方法 2. 能够掌握智能推荐算法
5	人口数据特征关系分析	1. 能够掌握 Python 可视化分析方法 2. 能够掌握编写可视化分析报告
6	电子商务实时推荐系统	1. Flume、Kafka 数据采集工具应用 2. 能够掌握 Spark 数据统计应用 3. 能够掌握实时推荐流程设计和实现

3.3.3. 高级

3.3.3.1.理论教学

工作领域	工作任务	职业技能要求	教学目标
1.数据采集与存储	1.1 数据采集	1.1.1 能够根据业务需求，进行大数据采集系统的配置 1.1.2 能够根据业务需求，进行大数据采集操作 1.1.3 能够使用 Python 调用大数据采集工具，获取采集数据 1.1.4 能够根据业务需求，设计大数据采集方案	（1）熟悉 Kafka 的功能和工作流程 （2）熟悉 Flume 的功能和工作流程 熟悉常见的大数据采集工具
	1.2 数据存储	1.2.1 能够运用非关系型数据库工具，进行非结构化数据查询 1.2.2 能够根据业务需求，进行非关系型数据库的备份与恢复 1.2.3 能够根据业务需求，进行非关系型数据库的性能优化 1.2.4 能够根据业务需求，使用 Python 访问非关系型数据库，实现非结构化数据的操作 1.2.5 能够根据业务需求，使用 Python 访问分布式文件系统，进行文件操作	（1）了解非关系型数据库的概念、与关系型数据库的区别 （2）熟悉 MongoDB 数据库的特性、对象和数据类型 （3）掌握常见的 MongoDB 数据库操作
	1.3 数据整合	1.3.1 能够根据业务需求，实现数据转换操作 1.3.2 能够根据业务需求，实现 ETL 全流程编排 1.3.3 能够根据业务需求，实现定时 ETL 1.3.4 能够根据业务需求，完成数据仓库方案设计	（1）掌握创建 Kettle 转换工程的方法 （2）掌握 Kettle 脚本的使用方法 （3）掌握 Kettle 作业的创建方法
2.数据分析与可视化	2.1 数据挖掘	2.1.1 能够根据业务需求实现算法选型 2.1.2 能够运用算法优化工具，实现算法参数调优，提升算法性能 2.1.3 能够根据业务需求使用分布式技术实现算法的并行计算，提升计算效率 2.1.4 能够根据业务需求，使用自动机器学习框架，进行数据挖掘	（1）掌握模型选择的常用方法 （2）掌握常用指标挖掘的方法 （3）熟悉常用算法优化的方法 （4）熟悉 Spark MLlib 算法
	2.2 文本挖掘	2.2.1 能够根据业务需求，实现文本分词与去停用词 2.2.2 能够根据业务需求，实现文本向量化	（1）掌握分词与去停用词 （2）掌握正则表达式清洗文本数据 （3）掌握文本向量化的常用算法

		2.2.3 能够根据业务需求，实现文本分类 2.2.4 能够根据业务需求，实现文本聚类 2.2.5 能够根据业务需求，实现关键词提取 2.2.6 能够根据业务需求，实现情感分析	（4）掌握文本分类的常用算法 （5）掌握文本聚类的常用算法 （6）掌握关键词提取的常用算法 （7）掌握情感分析的常用算法
	2.3 深度学习建模	2.3.1 能够根据业务需求，选择合适的深度学习框架 2.3.2 能够根据业务需求，实现全连接神经网络 2.3.3 能够根据业务需求，实现卷积神经网络 2.3.4 能够根据业务需求，实现循环神经网络 2.3.5 能够算法结果，进行深度学习算法评价	（1）掌握常见的全连接神经网络算法 （2）掌握常见的卷积神经网络算法 （3）掌握常见的循环网神经网络算法 （4）掌握常见的生成对抗网络算法
3.项目管理	3.1 立项管理	3.1.1 能够根据业务状况，完成项目建议书 3.1.2 能够根据可行性研究步骤，完成项目可行性研究报告 3.1.3 能够根据可行性研究报告，进行项目效益的预测与评估 3.1.4 能够根据项目招投标流程，跟踪招投标进度	（1）熟悉立项管理相关的概念与方法 （2）熟悉质量管理的基础、过程和工具 （3）熟悉人力资源管理的过程、工具与文件 （4）熟悉风险的识别、分析、应对与控制
	3.2 质量管理	3.2.1 能够根据质量管理流程，规划质量管理 3.2.2 能够根据现有的质量管理标准体系，实施质量保证 3.2.3 能够正确使用项目质量管理规划阶段技术与工具 3.2.4 能够正确使用项目质量管理执行阶段技术与工具	
	3.3 人力资源管理	3.3.1 能够根据人力资源管理的流程，绘制项目组织图 3.3.2 能够根据项目需求，组建项目团队 3.3.3 能够根据项目需求，制定人力资源管理计划 3.3.4 能够根据项目需求，制定团队绩效评价	
	3.4 风险管理	3.3.1 能够结合现有情况，识别项目风险 3.3.2 能够运用定性分析，分析项目风险 3.3.3 能够运用定量分析，分析项目风险 3.3.4 能够针对可能出现的风险制定风险应对方案	

3.3.3.2.实验教学

序号	实验名称	实验要求
1	数据仓库配置	1. 能够掌握数据仓库的基本组成 2. 能够章数据仓库的基本架构 3. 能够掌握数据仓库的配置方法
2	P2P 信用贷款数据指标挖掘	1. 能够掌握实现指标变换的方法 2. 能够掌握实现指标筛选的方法 3. 能够掌握实现特征构造的方法
3	消费者意见投诉挖掘	1. 能够掌握分词方法 2. 能够掌握去停用词方法

		3. 能够掌握 LDA 主题模型进行情感分析 4. 能够掌握词云图的绘制方法
4	广电大数据用户画像	1. 能够掌握数据探索的方法 2. 能够掌握 Spark MLlib 的用法 3. 能够掌握用户画像的基本概念与实现方法
5	姓名判别男女文本分类	能够掌握词袋模型 能够掌握文本向量化方法 能够掌握文本分类深度学习神经网络 能够掌握分类模型评测方法
6	基于混合模型的股市情感分析	1. 能够掌握模型评价的方法 2. 能够掌握模型选择的方法 3. 能够掌握深度学习模型的构建方法

4. 考核方案

4.1. 考核标准

1+X 大数据应用开发（Python）职业技能等级考核评价由考核评价组织实行统一大纲、统一试题、统一时间、统一标准、统一证书、统一组织的考核评价制度，原则上每年举行两次考核，即上半年和下半年各组织考核一次。考核标准分为合格、不合格，达到合格标准就代表通过，没有规定合格人数限制。

4.2. 考核方式

1+X 大数据应用开发（Python）职业技能等级考核分为理论知识与实践技能操作两部分。初、中、高三个等级的考核方式均设为 2 个科目：理论知识考核和实践技能操作考核。理论知识考核为闭卷考试，采用上机考试形式，考试时长为 120 分钟。实践技能操作考核为上机动手操作，考试时长为 240 分钟。

4.3. 评分标准

1+X 大数据应用开发（Python）职业技能等级考核评分实行百分制计分。理论考核试卷满分为 100 分，合格标准为 60 分；实践技能操作考核试卷满分为 100 分，合格标准为 60 分。

考核成绩占比：

（1）初级：理论成绩占比 35%，实操成绩占比 65%；

（2）中级：理论成绩占比 40%，实操成绩占比 60%；

（3）高级：理论成绩占比 45%，实操成绩占比 55%。

总成绩＝技能知识得分×技能知识考试占比＋技能实操得分×技能实操考评占比

理论知识和实践技能操作两部分考试成绩均合格，且总成绩合格的考生才可以获得相应级别的职业技能等级证书。

4.4. 考试题型

（1）理论知识题型

【初级】题型包括单选题、多选题、判断题、填空题、简答题、论述题 6 种题型，每种题型的题数与分数如下表。

题型	单选题	多选题	判断题	填空题	简答题	论述题
题数	10	10	10	10	4	2
分数	10	20	10	10	24	26

【中级】题型包括单选题、多选题、判断题、填空题、简答题、论述题 6 种题型，每种题型的题数与分数如下表。

题型	单选题	多选题	判断题	填空题	简答题	论述题
题数	10	10	10	10	4	2
分数	10	20	10	10	24	26

【高级】题型包括单选题、多选题、判断题、填空题、简答题、论述题 6 种题型，每种题型的题数与分数如下表。

题型	单选题	多选题	判断题	填空题	简答题	论述题
题数	10	10	10	10	4	2
分数	10	20	10	10	24	26

（2）实践操作技能题型

【初级】试卷包含 4 道实践操作试题，试题形式包括软件安装、软件操作、应用编程、案例分析等，每种题型的分数如下表。

题型	软件安装	软件操作	应用编程	案例分析
分数	20	20	20	40

【中级】试卷包含 3 道实践操作试题，试题形式包括软件安装、应用编程、案例分析等，每种题型的分数如下表。

题型	软件安装	应用编程	案例分析
分数	25	35	40

【高级】试卷包含 2 道实践操作试题，试题形式包括案例分析、项目实战等，每种题型的分数如下表。

题型	案例分析	项目实战
分数	50	50

4.5. 考核权重

（1）大数据应用开发（Python）【初级】考核权重

工作任务	技能要求	知识要求	总权重
平台管理	20%	10%	30%
数据采集与存储	20%	10%	30%
数据分析与可视化	25%	15%	40%

（2）大数据应用开发（Python）【中级】考核权重

工作任务	技能要求	知识要求	总权重
平台管理	15%	5%	20%
数据采集与存储	15%	10%	25%
数据分析与可视化	25%	15%	40%
项目管理	5%	10%	15%

（3）大数据应用开发（Python）【高级】考核权重

工作任务	技能要求	知识要求	总权重
数据采集与存储	25%	15%	40%
数据分析与可视化	25%	15%	40%
项目管理	5%	15%	20%

4.6. 考核平台

采用专用开发的考试考核平台。

5. 试点院校建设

5.1. 试点院校申请条件

5.1.1. 开办相关专业

具备办学许可的法人单位，且培训学校已开设泰迪 1+X 认证所对应的相关专业，该专业近 3 年连续招生且每年招收全日制在校生不低于 30 人。

5.1.2. 领导高度重视

试点学校领导高度重视 1+X 职业技能等级证书考试的组织及实施工作，注重学生的技能训练与提高。

5.1.3. 具备师资队伍

相关专业具备培训能力的专兼职师资队伍，团队成员不少于 4 人。其中，“双师型”教师不少于 50%，行业企业专家比例不低于 20%，具有满足模块化教学需要的教学团队，专业带头人应具有高级职称。

5.1.4. 实训场地设备

教学场地需配备必要的多媒体和专业实训设备，可以同时满足不低于 40 人进行理论学习和实践实训操作。

5.1.5. 教学管理团队

专职教学管理团队成员不少于 2 名，技术服务团队成员不少于 2 名。有固定的办公场所。团队负责人能够充分调动资源，提供培训所需的保障条件。

5.2. 试点院校申请流程

5.3. 试点院校工作内容

5.3.1. 人才培养方案

依据职业技能等级证书认证标准，调整专业人才培养方案、课程体系及教学内容，推进教育教学改革。

5.3.2. 教学资源开发

依据职业技能等级证书标准，开展课程标准制定、教学资源开发（泰迪合作支持）。

5.3.3. 教学团队建设

依据职业技能等级证书认证培训要求，融合行业企业资源，实施教师分工协作的模块化教学。（教师个体知识能力—教师团队知识能力一分工协作教学一达成教学目标）。

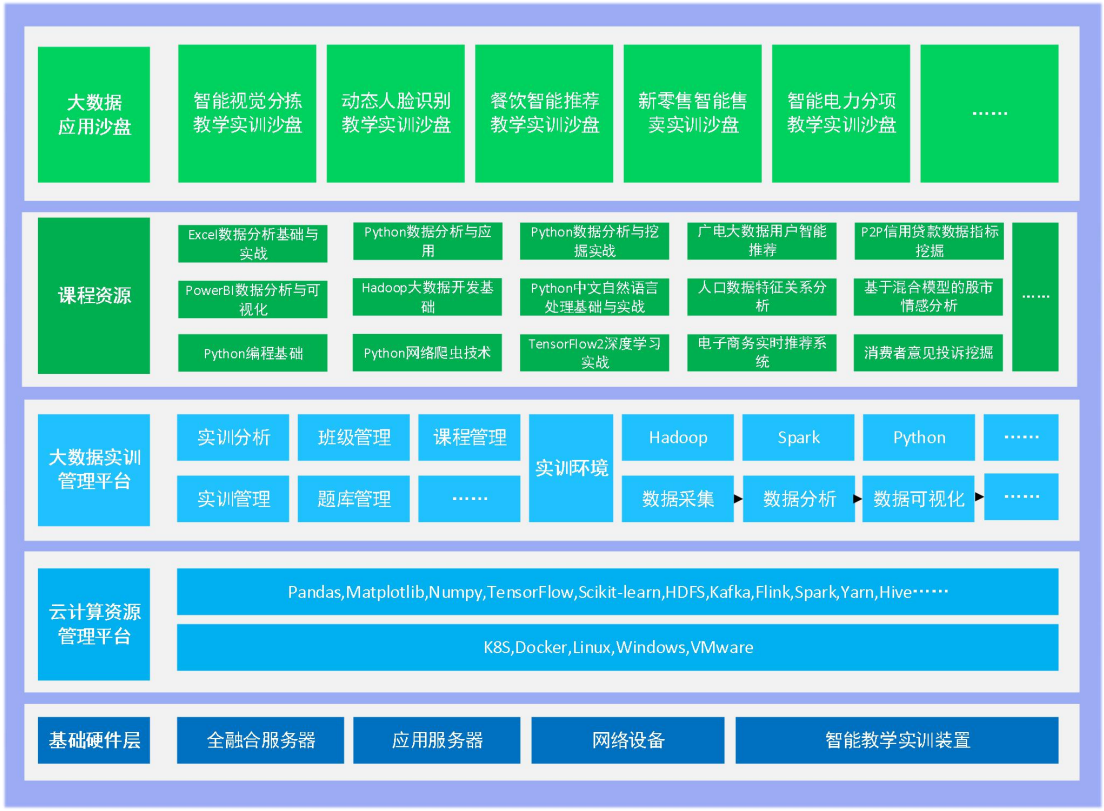
5.3.4. 保障完善条件

建立建全相关组织机构、制度流程、资金条件、激励机制。

5.4. 试点院校实训基地建设

为满足院校职业技能岗位人才培养需求，泰迪科技针对 1+X《大数据应用开发（Python）职业技能等级标准》要求为院校提供优质的实训环境。硬件上基地配备 3 台全融合服务器、1 台应用平台服务器、1 台数据采集服务器、1 套服务器机柜、1 套机架式 KVM 切换器、1 套管理交换机、2 套硬件交换机、61 套大数据工作站、1 系列配套软硬件实训平台，供学生进行 Python 大数据应用开发职业技能初、中、高证书考核实训。

5.4.1. 实训基地整体描述



整体的实训基地主要由 5 部分组成，包括基础硬件层、云计算资源管理平台、大数据实训管理平台、课程资源、数据应用沙盘，各部分具体作用如下：

（1）基础硬件层：主要为实训基地提供实训的硬件基础支撑，包括全融合服务器、应用服务器、数据采集服务器、管理交换机、硬件交换机等必需硬件设备。

（2）云计算资源管理平台：主要指的是支撑大数据应用开发的系统与软件服务，包括 Linux 操作系统、Docker 容器、Hadoop 大数据集群、Python 脚本编程环境、各类机器学习、深度学习框架等内容。

（3）大数据应用开发教学实训平台：主要为师生提供高可用的大数据应用开发实训环境。实训环境统一由大数据实训管理平台进行管理，同时还提供了课程管理、人员管理、考试管理、作业管理等功能。

（4）课程资源：主要是为实训基地提供实训课程。实训课程涵盖整个 1+X《大数据应用开发（Python）职业技能等级标准》，包括但不限于《Excel 数据分析基础与实战》、《Power BI 数据分析与可视化》、《Python 编程基础》、《Python 网络爬虫技术》、《Python 数据分析与应用》、《Hadoop 大数据开发基础》、《Python 数据分析与挖掘实战》、《Python 中文自然语言处理基础与实战》、《TensorFlow 2 深度学习实战》。此外还会提供对应课程配套的项目案例，让实训更加贴合企业场景。

（5）大数据应用沙盘：主要作用是在课程资源的基础上，提供更加真实、可见的场景。通过大数据应用沙盘，不仅可以学习到沙盘对应场景的业务知识，更加能够直观感受大数据技术在业务场景中的应用以及效果。

5.4.2. 实训系统介绍

5.4.2.1. 基础硬件建设

（1） 全融合服务器

名称	参数配置	单位	数量	参考产品型号
全融合计算服务器	规格：2U机架式 CPU：Intel Xeon 系列；双路CPU；主频2.1GHz；每颗14核心； 内存:总容量≥256GB DDR4； 硬盘：3*480GB（SSD 企业级）、2*240GB（SSD 企业级）、3*4TB(SATA HDD 7.2K); 网卡：双口万兆网卡10Gbps 1块，板载双口千兆网卡1Gbps 1块；具备独立的千兆管理网口，附2个光模块； RAID卡： 标准外插RAID控制器, 1GB 缓存，支持RAID 0/1/5/6/10/50/60级别； I/O插槽：PCI-E 3.0插槽总数≥8个； 电源：提供1+1高效冗余热插拔电源 主机安全认证：支持服务器厂商自研的主机安全系统加固软件，从操作系统内核实现对服务器的安全加固，与服务器同一品牌。该系统可实现内核级安全加固，增强型身份认证、服务完整性检测、注册表防篡改机制等功能； 售后服务：由原厂商直接提供三年质保、首年免费维护服务； 提供国家环保标志认证、3C国家强制性产品认证和中国节能认证证书复印件并加盖投标人公章。	台	3	R730/R740/浪潮 Inspur_NF5280M4

（2） 应用平台服务器

名称	参数配置	单位	数量	参考产品型号
应用平台服务器	规格：2U机架式 CPU：Intel Xeon系列CPU；数量≥2颗；主频≥2.0GHz； 每颗CPU≥12核心； 内存：总容量≥160GB，DDR4，最大扩充768GB； 硬盘：企业级存储硬盘；SSD 240GB，≥2块；SATA 7.2K 3.5英寸 2TB，≥4块，支持热插拔； 网卡：集成千兆网卡；接口数量≥2 RAID卡： RAID控制器，缓存≥1GB，支持RAID 0/1/5/6/10/50； 电源：提供1+1高效冗余热插拔电源； 售后服务：由原厂商直接提供三年质保； 提供国家环保标志认证、3C国家强制性产品认证和中国节能认证证书复印件并加盖投标人公章。	台	1	DELL R730 浪潮 Inspur_NF5270M4

（3） 数据采集服务器

名称	参数配置	单位	数量	参考产品型号

数据采集服务器	规格：2U 机架式 CPU：Intel Xeon 系列 CPU；数量≥2 颗；主频≥2.0GHz； 每颗 CPU≥12 核心； 内存：总容量≥160GB，DDR4，最大扩充 768GB； 硬盘：企业级存储硬盘；SSD 240GB，≥2 块；SATA 7.2K 3.5 英寸 2TB，≥4 块，支持热插拔； 网卡：集成千兆网卡；接口数量≥2 RAID 卡： RAID 控制器，缓存≥1GB，支持 RAID 0/1/5/6/10/50； 电源：提供 1+1 高效冗余热插拔电源； 售后服务：由原厂商直接提供三年质保； 提供国家环保标志认证、3C 国家强制性产品认证和中国节能认证证书复印件并加盖投标人公章。	台	1	DELL R730 浪潮 Inspur_NF5270M4

（4） 服务器机柜

使用服务器机柜，是采用专业的设备存放计算机和相关控制设备，可以提供对存放设备的保护，屏蔽电磁干扰，有序、整齐地排列设备，方便以后维护设备。机柜可以存放服务器、网络交换机、控制台等。服务器机柜配置如下。

名称	参数配置	单位	数量	参考产品型号
机柜	规格：42U 机柜尺寸为：宽≥600mm×长≥1000mm×高≥2000mm，立柱间距为 485mm（19 英寸标准），立柱厚度为≥2.0mm 材质：机柜采用优质冷轧钢材质，脱脂喷塑工艺面板 功能：可安装主机、服务器、KVM 切换器等服务器设备 安全：机柜底部设有接地保护，保证工作时的安全 拆卸：机柜顶部和底部设有布线口，可根据需求进行拆装 标准：符合《网络机柜技术条件》要求	台	1	大唐 卫士

机柜系统性地解决了计算机应用中的高密度散热、大量线缆附设和管理、大容量配电及全面兼容不同厂商机架式设备的难题，从而使数据中心能够在高稳定性的环境下运行。

（5） 管理交换机

名称	参数配置	单位	数量	参考产品型号
管理交换机	网络标准：802.3ah 和 802.1ag 端口：24 个万兆 10G SFP+端口，2 个 4 万兆 40G QSFP+ 速率：100/1000/10000Mbps，40000Mbps 交换容量：1.28Tbps/12.8Tbps 包转发率：480Mpps Mac 地址表：32k 支持全双工工作模式 输入电源：额定电压范围：100-240V AC；50/60Hz；最大电压范围：90-264V AC；47/63Hz 网络模块：SFP/SFP+万兆多模 6 个、光电转换模块 2 个	台	1	华为 S6720S-26Q-LI-24S-AC

万兆光交换机交换机用于服务器和服务器、服务器与应用交换机的通讯连接，由于大数据需要大容量、高速度的数据和信息传输，特别是服务器集群之间，计算节点之间，需要超高速、毫秒级的计算速度，因此需要较高配置的万兆交换机，才能保证大数据的计算实时性、快速性和响应的稳定性，并且科研实验时，服务器计算和响应准时性，不卡顿、不影响实训。

（6） 应用交换机

把网络节点上话务承载装置、交换级、控制和信令设备以及其他功能单元的集合的集合体，即把用户线路、电信电路和(或)其他要互连的功能单元根据单个用户的请求连接起来。具体配置如下。

名称	参数配置	单位	数量	参考产品型号
应用交换机	1. 网络标准：IEEE 802.3 、IEEE 802.3u、IEEE 802.3ab、IEEE 802.3x 2. 端口：48 个 10/100/1000Base-T 以太网端口，4 个 100/1000 Base-X SFP 光口 3. 速度：10/100/1000/1000Mbps 自适应 4. 交换容量：336Gbps 5. 包转发率：78Mpps 6. Mac 地址表：16k 7. 支持半双工、全双工、自协商工作模式 8. 电源模式： 100~240V AC，50/60HZ	台	2	华为 S1720-52GWR-4P

（7） 大数据工作站

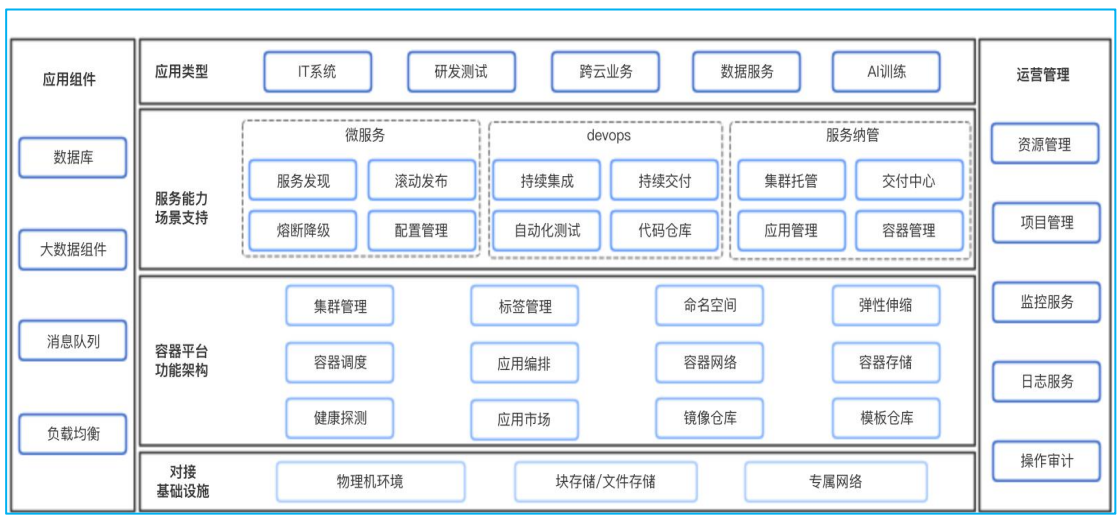
名称	参数配置	单位	数量	参考产品型号
大数据工作站	操作系统：Win10 家庭中文版 CPU：Intel i5 系列，频率 2.8GHz，核心数：6 核 显卡：独立显卡，显存≥2GB 内存：容量≥8GB，DDR4，插槽数量：2 个，最大支持容量 32GB 硬盘：固态硬盘 容量≥128GB；机械硬盘 容量≥1TB ， 转速 7200 转/分钟 显示器：尺寸≥21.5 英寸，高清屏 鼠标：有线鼠标 键盘：有线键盘 USB 接口：4 个 视频接口：VGA\HDMI 接口 网口：RJ45 接口 电源：≥180W ES 电源 品牌：国际知名名牌 品质管理：制造商通过 ISO9000 系列体系认证	台	61	戴尔

5.4.2.2.云资源管理平台

云计算资源管理平台，简称容器云管理平台，是对实验室所有服务器资源，结合 Docker 容器技术，进行云化

处理，使得服务器资源（服务器中的软件、系统、CPU、内存、存储、网络等）成为可管理及维护的云计算服务中心。提供计算、存储、网络、安全等方面的功能和应用，一方面实现更加精细化的资源管理，控制成本，提供资源利用率，另一方面基于 Docker 容器技术，极大加快实训平台搭建、启动速度，提高平台稳定性及弹性伸缩能力。以安全、简单、智能为设计理念，带来稳定、高效、高性能、快速的实训平台基础设施建设和云解决方案，帮助高校建立一套全新的云计算服务中心。

容器云资源管理，提供软件定义资源方式：软件定义计算、软件定义网络、软件定义存储；提供多种网络组网方式。



容器云资源管理平台功能如下：

5.4.2.3.集群管理

（1） 用户可以一键创建灵活的管理集群，支持集群弹性伸缩，节点支持升降配。

（2） 用户独占容器集群，可自定义专有网络 VPC 等环境，保证集群安全隔离。

（3） 整合命名空间，提供一个集群内不同环境的逻辑隔离能力。

5.4.2.4.应用管理

（1） 应用管理：支持通过标准镜像发布应用，也支持通过模版发布应用，应用内服务一键部署/停止。

（2） 快速发布：容器组秒级发布、回滚，利用滚动升级不中断业务更新服务。

（3） 服务发现：可通过负载均衡域名或服务名称加端口访问服务，可避免服务后端变化时 IP 变更带来的影响。

（4） 存储支持：支持数据卷管理，对有状态服务数据进行多形式的持久化存储。

（5） 动态扩缩：服务灵活水平扩展，应对业务快速变化。

（6） 配置项：配置项以数据卷或环境变量的方式挂载到容器组中，支持可视化和 YAML 两种编辑形式。

（7） 安全灾备：容器异常自动恢复，服务内容器可跨集群部署，可快速迁移。

5.4.2.5. 交付中心

（1） 本地仓库：提供安全、高可用的私有镜像仓库以及私有 Chart 仓库；拥有丰富的权限控制，针对不同集群、项目进行读写权限分配。

（2） 应用市场：提供官方 Chart 包，结合 Helm 功能简化了 Kubernetes 部署应用的版本控制、打包、发布、删除、更新等操作

（3） 镜像市场：定期更新 DockerHub 官方主流镜像，提供 DockerHub 官方镜像加速拉取功能。

（4） 模板仓库：集成 Kubernetes 配置项目管理简化应用模版管理。

5.4.2.6. 运维管理

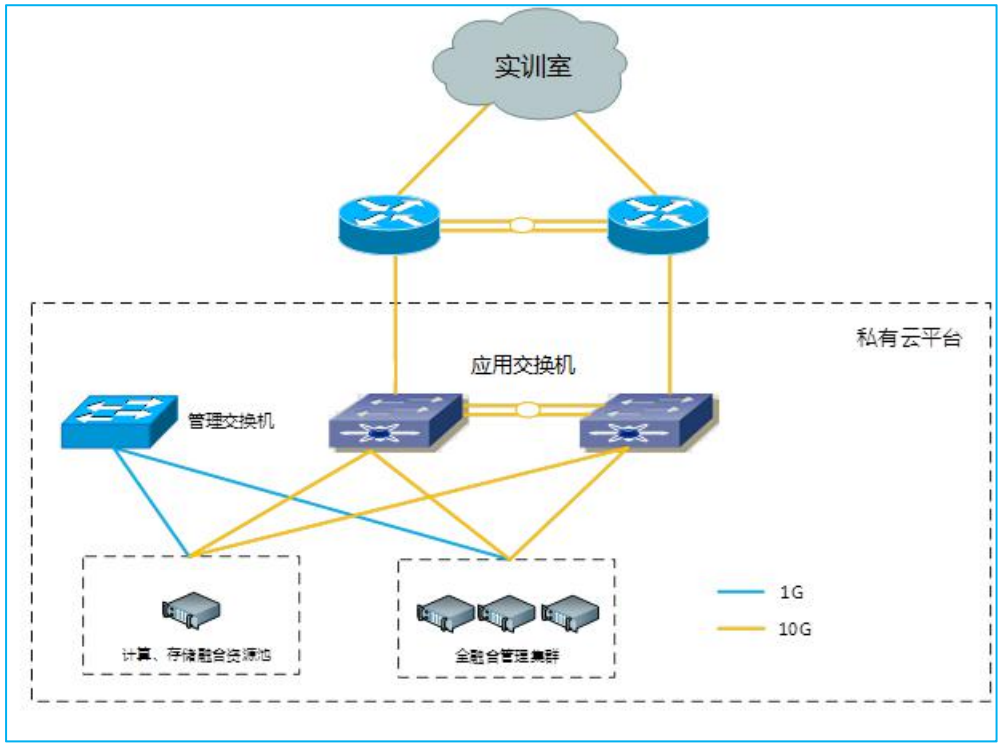
（1） 云监控的集成与使用

容器云云监控为容器服务的集群、应用服务、容器组、实例等提供即开即用的监控数据采集、聚合展示、报警功能。用户可以验证集群、应用是否正常运行，并创建相应的报警机制。使用容器云云监控可节省用户自建容器监控的各项成本，结合云监控的一站式服务保障业务的稳定运行。有关云监控的更多信息，请参阅云监控产品帮助文档。

（2） 日志服务的集成与使用

基于物理机或虚拟机部署的应用，日志采集相关技术都比较完善，有比较健全的 Logstash、Fluentd、FileBeats 等。但在容器架构中，尤其在 Kubernetes 中，日志采集并没有很好的解决方案，自建技术成本高昂。容器云容器服务支持多种方式进行应用日志的管理，通过容器云提供的原厂日志服务，用户可以享受一站式的日志管理，方便对

容器服务集群内的容器日志进行集中管理、实时查询、统计分析、归档备份，在海量容器中实现快速定位并解决问题。日志服务支持采集容器文本日志，支持采集容器标准输出流日志。



5.4.2.7.大数据应用开发教学实训平台

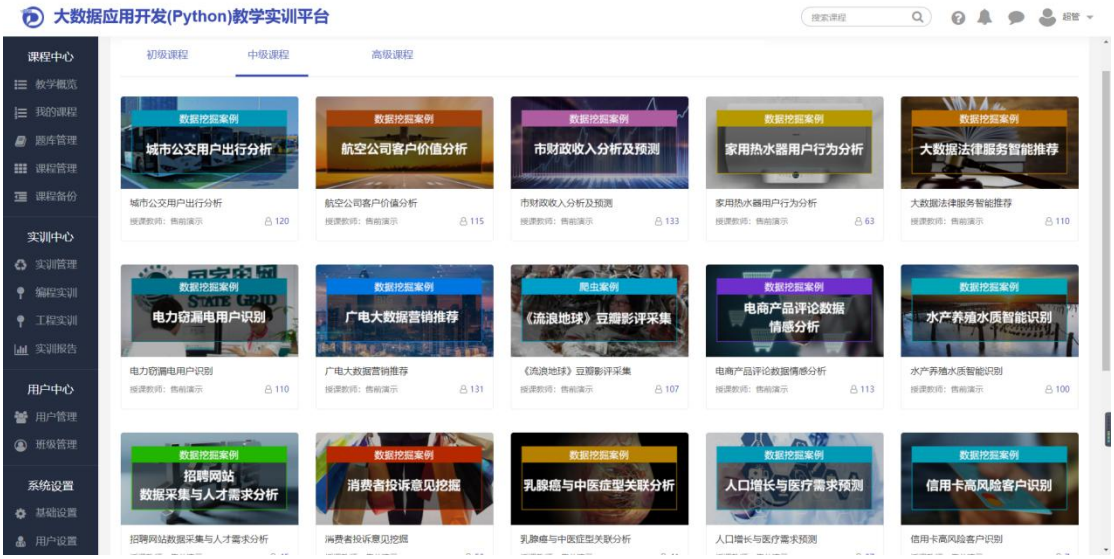
大数据应用开发（Python）教学实训平台提供的功能与环境与《大数据应用开发（Python）职业技能等级标准》及其课程一一对应，各个平台具体功能与对应实训课程如下表所示。

序号	平台名称	主要功能	工作任务	对应课程
1	大数据应用开发教学实训管平台	管理实训课程		
2	Python 实训平台	进行有关 Python 脚本编程有关实训	数据采集 数据处理	《Python 编程基础》 《Python 数据分析与应用》
3	大数据实训平台	进行有关 Hadoop 大数据集群有关实训	软件安装 软件管理 系统管理 数据存储	Hadoop 大数据技术基础
4	互联网大数据采集平台	进行数据采集有关实训	数据采集	数据采集

5	大数据整合平台	进行数据整合和迁移有关实训	数据整合 数据存储	数据整合与预处理
6	数据挖掘建模平台	进行有关数据分析与挖掘相关的实训	数据分析 数据挖掘 文本挖掘 深度学习 建模	《Python 数据分析与应用》 《Python 数据分析与挖掘实战》 《Python 中文自然语言处理基础与实战》 《TensorFlow 2 深度学习实战》
7	大数据可视化平台	进行有关数据可视化相关实训	数据可视化	数据可视化
8	大数据应用沙盘	通过场景进行数据挖掘相关学习	数据分析 数据挖掘	《Python 数据分析与应用》 《Python 数据分析与挖掘实战》

5.4.2.8.大数据实训管理平台

大数据实训管理平台主要为大数据应用技术技能实训和教学提供“一站式”的服务,采用 B/S 架构，基于 Web 管理，做为进入其他软件工具平台的统一入口。所有的实训课程及案例资源进行统一管理，包括课程视频、课程 PPT、实训指导书、作业管理、考试管理、成绩管理及用户管理等。



大数据实训管理平台主要功能说明如下：

（1） 采用 B/S 架构，即浏览器/服务器架构。

（2） 支持用户权限区分。分为系统管理员、教师、学生；支持权限管理，可进行用户权限配置、用户角色定义、用户角色分配等。

（3） 支持修改个人账户信息、姓名、头像等。

（4） 提供课程管理功能。可创建新的课程类别，并对具体课程类别进行编辑、删除等操作。支持在具体的课程类别下创建新的课程。

（5） 支持教学课程的创建，支持自定义配置课程信息及资源，配置信息及资源为：课程名称、课程起讫时间、课程封面、课程成员、教学 PPT、教学视频、实验环境、作业和考试等。

（6） 支持教学视频在线播放和教学 PPT 在线浏览。

（7） 支持批量上传作业或考试题目到平台题库中，且题库可直接应用于作业或考试。

（8） 支持客观题（如选择题、判断题和填空题等）自动评分，答题完毕马上可以看到成绩。

（9） 支持学生选课并学习相应的课程。查看教师分配的所有学习资源，提交实验报告或作业，进行在线考试，查看教师已批改作业的成绩详情等。

（10） 提供成绩管理功能，对成绩等级及分数段相关信息进行设置。

5.4.2.9. 实训环境

(1) Python 编程实训平台

Python 编程实训平台是一套建立在虚拟化层上基于 Python 的平台，它更注重于学生的实践环节，即编程应用能力。平台提供 Python 环境，让学生在学习“Python 编程基础”课程后，进行编程练习，还提供 Python 爬虫实验让

学生在学习“数据采集与网络爬虫技术”课程后的实践操作，让学生在掌握了理论的基础上，结合平台的实验学习与实操，有效地解决学生缺少实践经验、缺乏实践应用能力等问题。学生通过基于 Python 的实验，配合相应的上机实验指导书，动手实操，学生可在短时间内，掌握 Python 编程技术，更能满足大数据应用开发考核的要求，从而具备考取证书的能力。

（1） 边看边做：学生可直接在本地客户端上进行实验，界面分为左右两栏。平台界面左边可查看实验指导书，右边可操作本地虚拟桌面。

（2） 虚拟机和本地物理机间文件互传

（3） 支持实训报告在线提交，并支持提交本地文件报告和实训环境中的文件报告这 2 种方式。

(2) 大数据开发实训平台

大数据开发实训平台是一套基于 B/S 架构建立在虚拟层上的教学实训平台，是基于 Web 的虚拟云主机系统，用户可通过浏览器即可访问平台进行实验。可通过课程实验进入大数据开发实训平台。

每个实验对应着相应搭载好的满足此实验条件的系统镜像，点击进入 Web 虚拟云主机继续相应的实验操作。界面可分为左右两栏，平台界面左边可查看实验指导书和教学 PPT，右边可操作实验虚拟云主机桌面。虚拟云主机桌面支持全屏展示。

每个 Web 虚拟云主机就是一个 Hadoop 集群，标配 3 个服务器节点。基于此环境学生可不用关注环境的搭建而专注于大数据应用开发的知识学习和能力提升，结合课程学习要求，使用此环境可快速进行 Hadoop 技术相关的大数据应用开发。

5.4.2.10. 互联网大数据采集平台

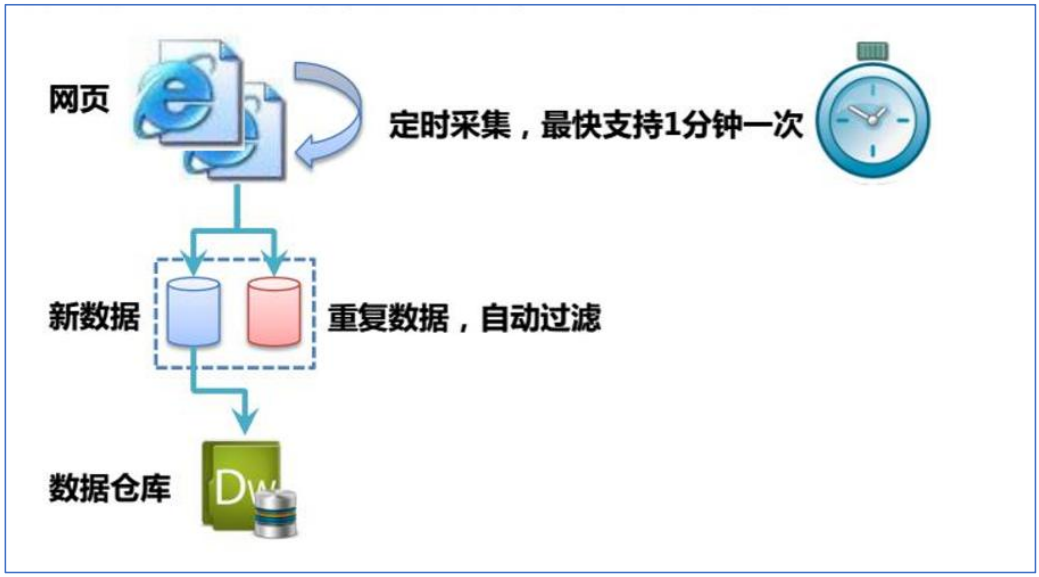
互联网大数据采集平台是强大且易用的互联网数据采集平台，可简单快速地将网页数据转化为结构化数据，存储 EXCEL、CSV 或数据库等多种形式，并且提供基于云计算的大数据云采集解决方案，实现精准、高效、大规模的数据采集，提供多种操作模式，满足不同用户的个性化需求。其智能模式可实现输入网址全自动化导出数据，是国内首个大数据一键采集平台。

互联网大数据采集平台以完全自主研发的分布式云计算平台为核心，可以在很短的时间内，轻松从各种不同的网站或者网页中获取大量的规范化数据，帮助任何需要从网页获取信息的客户实现数据自动化采集、编辑、规范化，从而降低获取信息的成本、提高效率。在政府、高校、企业、银行、电商、科研、汽车、房产、媒体等众 多行业及领域均有广泛应用。

作为一款通用的网页数据采集器，并不针对于某一网站某一行业的数据进行采集，而是网页上所能看到或网页源码中有的文本信息几乎都能采集，市面上 98% 的网页都可以用商业互联网采集平台进行采集。

使用本地采集（单机采集），除了可以实现绝大多数网页数据的爬取，还可以在采集过程中对数据进行初步的清洗。如使用程序自带的正则工具，利用正则表达式将数据格式化，在数据源头即可实现去除空格、筛选日期等多种操作。其次商业互联网采集平台还有提供分支判断功能，可对网页中信息进行是与否的逻辑判断，实现用户筛选需求。

云采集除具有本地采集（单机采集）的全部功能之外，还可以实现定时采集，实时监控，数据自动去重并入库，增量采集，自动识别验证码，API 接口多元化导出数据以及修改参数。同时利用云端多节点并发运行，采集速度将远超于本地采集（单机 采集），多 IP 在任务启动时自动切换还可避免网站的 IP 封锁，实现数据采集的最大化。



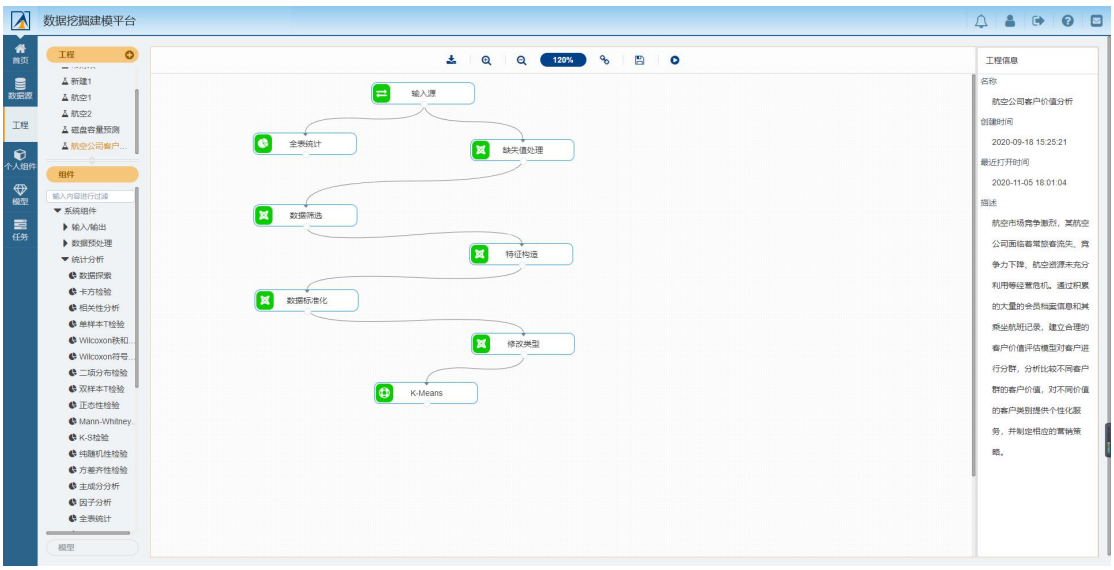
5.4.2.11. 大数据整合平台

大数据整合平台是一套 ETL（Extract-Transform-Load）工具，主要满足数据整合与预处理类课程的实训需求，主要提升大数据应用开发涉及的数据整合、迁移、储存等相关实践技能，学生可在应用中实现将数据从来源端经过抽取、转换、加载至目的端的过程。使用本工具，可更高效、简便地将数据从业务系统迁移到数据分析数据库，并实现对数据的清洗、修改、计算、集成等处理。

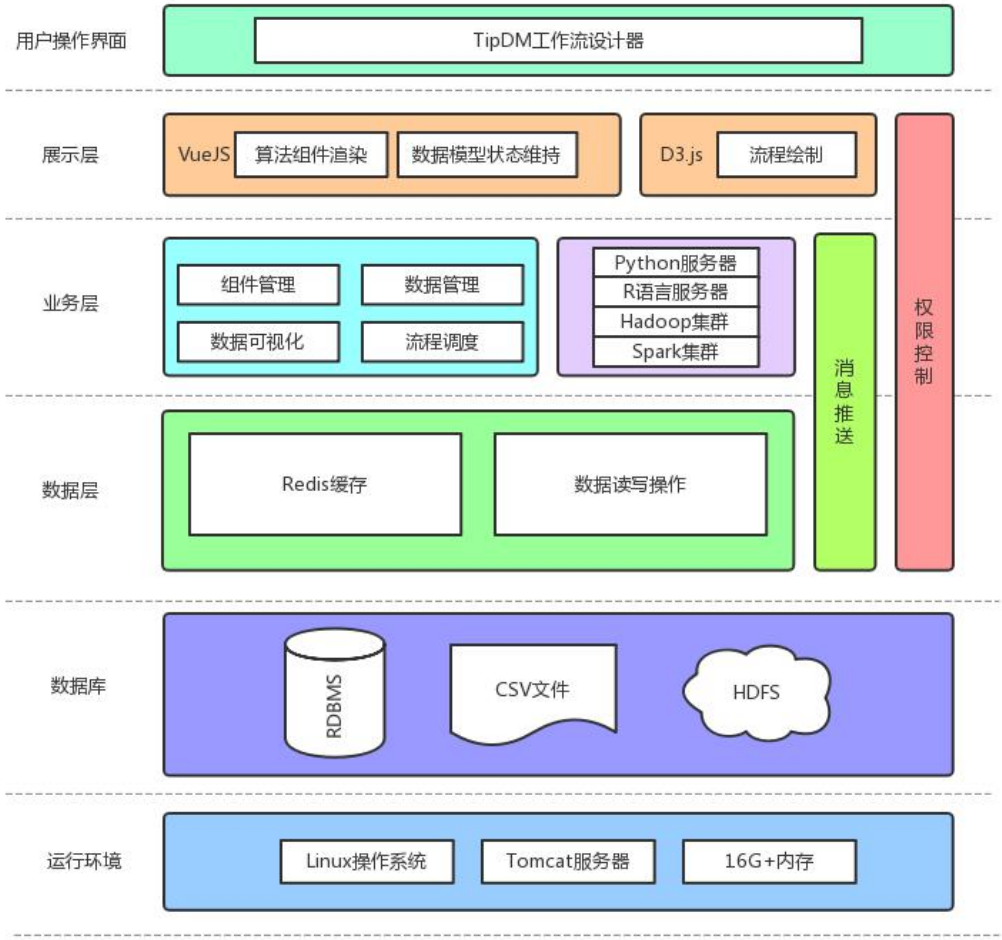
大数据整合平台主要由 Job Automation、Schedule Servert、Monitor Center、ManagerServer、Agent、Common Job Component 等组成。平台采用了先进的 J2EE 技术架构，具有很强的跨平台性。部署简便，维护简单，容易使用。支持分步式的多机集群，能承载大规模数据的高负荷运行，具有良好的稳定性。平台采用了三层架构，结构清晰，具有良好的扩展性、稳定性和容错性。平台的各个组件可以单独进行使用，从而提高工具平台的灵活性。

5.4.2.12. 数据挖掘建模平台

数据挖掘建模平台是由广东泰迪智能科技股份有限公司自主研发，面向大数据挖掘项目的工具。平台使用 JAVA 语言开发，采用 B/S 结构，可通过浏览器进行访问。平台提供了基于 Python 和 Hadoop/Spark 引擎的数据分析功能。平台支持工作流，用户可在没有 Python 等编程语言基础的情况下，通过拖拽的方式进行操作，以流程化的方式将数据输入输出、统计分析，数据预处理、分析与建模等环节进行连接，从而达成大数据分析的目的。



产品主要由首页模块，数据源模块、算法组件模块、工程模块、个人组件模块，模型模块，任务模块、接口拓展模块构成。整体产品的架构图所下图所示。



（1） 首页：成功登录平台后，进入的第一个页面就是首页，首页用于展示模板。模板主要用于标准大数据分析案例的快速创建和展示。通过模板，能够建立一个个无需导入数据，设置参数就能够快速运行的工程。同时，每一个模板的创建者都具有模板的所有权，能够对模板进行管理。

（2） 数据源：数据空间模块主要用于数据集的导入与管理，用户可根据数据大小选择来源于文件或者来源于数据库。来源于文件支持从本地导入任意类型的数据；来源于数据库支持从 DB2、SQL Server、MySQL、Oracle、PostgreSQL 等常用关系型数据库导入数据。与此同时，每一份导入的数据都能够进行数据预览，数据删除等操作。

（3） 工程：工程模块主要用于大数据分析流程化案例的创建与管理。通过工程模块，能够创建空白工程，进行大数据项目流程的配置，建模结果可通过可视化报告进行查看。对于完成度优秀的工程，可以将其保存为模板，让其他使用者学习和借鉴。

（4） 算法组件：算法组件分为系统组件和个人组件两大部分。其中系统组件为泰迪科技提供的通用组件，个人组件提供的是一个用户自定义组件的入口。

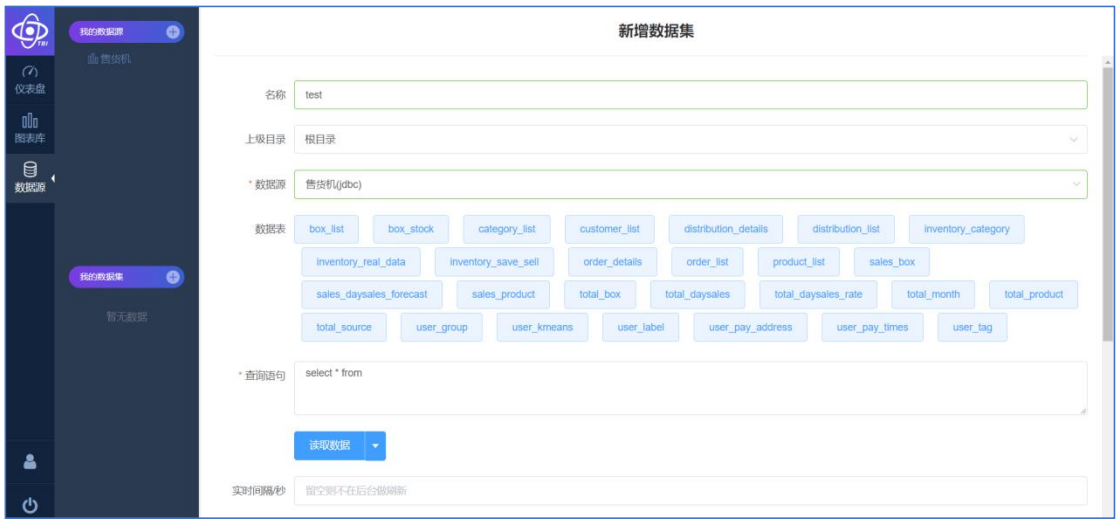
（5） 模型：模型管理模块主要针对分类算法，对训练集使用某分类算法得出的模型，里面将包括模型的输入、输出、算法参数的信息。可将该模型部署到一批新的数据上用于验证或者预测。模型除了支持导入、

导出 PMML 文件外，还提供分享功能。

（6） 任务：在日常的数据分析与挖掘的过程中，常常会遇见不需要在当前请求下立刻执行的任务，或者是需要定时去做的一些任务，这些需要一个用来存放任务的队列，一个任务执行定时器，以及一个用来执行任务的工具。任务模块就是基于这些需求开发的，支持常规任务定义，任务对列，任务调度。

5.4.2.13. 数据可视化平台

数据分析可视化平台是一套基于海量网络数据分析，实现大数据商业智能的文本分析和可视化平台。主要服务于数据分析与可视化类课程，帮助学生学习数据挖掘、机器学习和数据智能可视化等相关知识。



5.4.3. 课程资源建设

5.4.3.1.实训课程

(1) 《Excel 数据分析基础与实战》

一、 **课程简介**

大数据时代已经到来，在商业、经济及其他领域中基于数据和分析去发现问题并做出科学、客观的决策越来越重要。Excel 作为常用的数据分析工具之一，在数据分析技术的研究和应用中，扮演着至关重要的角色。通过本课程的学习，使学生学会使用 Excel 编辑数据，通过排序、筛选、分类汇总等方式探索数据，通过多种函数的使用处理数据，将理论与实践相结合，为将来从事以 Excel 为生产力工具的人员奠定基础。

二、 **课时数**

理论教学 5 学时，实践教学 27 学时，总计 32 学时。

三、 **课程资源**

包含 44 份实训指导书、21 个课程视频、14 份课程 PPT、76 份数据。

四、 **课程内容**

项目 1 认识 excel2016

项目 2 输入数据（1）

项目 2 输入数据（2）

项目 2 输入数据（3）

项目 2 输入数据（4）

项目 3 美化工作表（1）

项目 3 美化工作表（2）

项目 3 美化工作表（3）

项目 3 美化工作表（4）

项目 4 获取文本数据

项目 5 获取网站数据

项目 6 获取 MySQL 数据库中的数据

项目 7 对订单数据进行排序

项目 8 筛选订单数据的关键信息

项目 9 分类汇总每位会员的消费金额

项目 10 制作数据透视表（1）

五、 实训目录

实训 2　获取 PostgreSQL 数据库中的数据

实训 3　获取 sales 的数据

项目 7　对订单数据进行排序：

实训 1　对订单数据进行排序

实训 2　多种排序方式

实训 3　对销售业绩进行排序

项目 8　筛选订单数据的关键信息：

实训 1　筛选订单数据的关键信息

实训 2　高级筛选

实训 3　筛选便利店销售业绩

项目 9　分类汇总每位会员的消费金额：

实训 1　分类汇总每位会员的消费金额

实训 2　嵌套分类汇总

实训 3　自助便利店销售业绩分类汇总

项目 10　制作数据透视表：

实训 1　制作数据透视表

实训 2　数据透视表操作

实训 3　创建订单数据透视表

项目 11　使用日期或时间函数完善员工数据：

实训 1　使用日期和时间函数完善员工数据

实训 2　创建和提取日期和时间数据

实训 3　完善【自助便利店会员信息】

项目 12　使用数学函数处理企业的营业数据：

实训 1　使用数学函数处理某企业的营业数据

实训 2　取整函数

实训 3　完善【自助便利店销售数据】

项目 13　使用统计函数处理企业的营业数据：

实训 1　使用统计函数处理企业的营业数据

实训 2　计算方差与标准差

实训 3　使用数组公式

实训 4　完善【8 月商品销售数据】

项目 14　使用宏生成工资条：

实训 1　使用宏生成工资条

实训 2　使用 VBA 编程

实训 3　创建宏

(2)　《Power BI 数据分析与可视化》

一、**课程简介**

在大数据时代已经到来，在商业、经济及其他领域中基于数据和分析去发现问题并做出科学、客观的决策越来越重要。数据分析技术将帮助企业用户在合理时间内获取、管理、处理以及整理海量数据，为企业经营决策提供积极的帮助。数据分析作为一门前沿技术，广泛应用于物联网、云计算、移动互联网等战略新兴产业。有实践经验的数据分析人才已经成为了各企业争夺的热门。为了推动我国大数据，云计算，人工智能行业的发展，满足日益增长的数据分析人才需求，特开设 Power BI 数据分析与可视化课程。

二、**课程资源**

包含 18 份实训指导书、13 个课程视频、11 份课程 PPT、123 份数据、25 份代码。

三、**课程内容**

第 1 章　数据分析与可视化概述

1.1　安装

1.2　界面介绍

第 2 章　数据获取

2.1　获取 excel 数据

2.2　获取 web 数据

2.3　从数据库中获取数据

第 3 章　M 语言数据建模与处理

3.1　M 语言获取网络分页数据

第 4 章　DAX 语言数据处理

4.1　数据的集成

4.2　数据的清洗

第 5 章　数据分析可视化

5.1　变换数据

5.2　数据归约

第 6 章　数据分析报表

四、 实训目录

第 7 章　Power BI 移动版数据部署

实训 1　部署超市运营分析报表

第 8 章　自助售货机综合案例

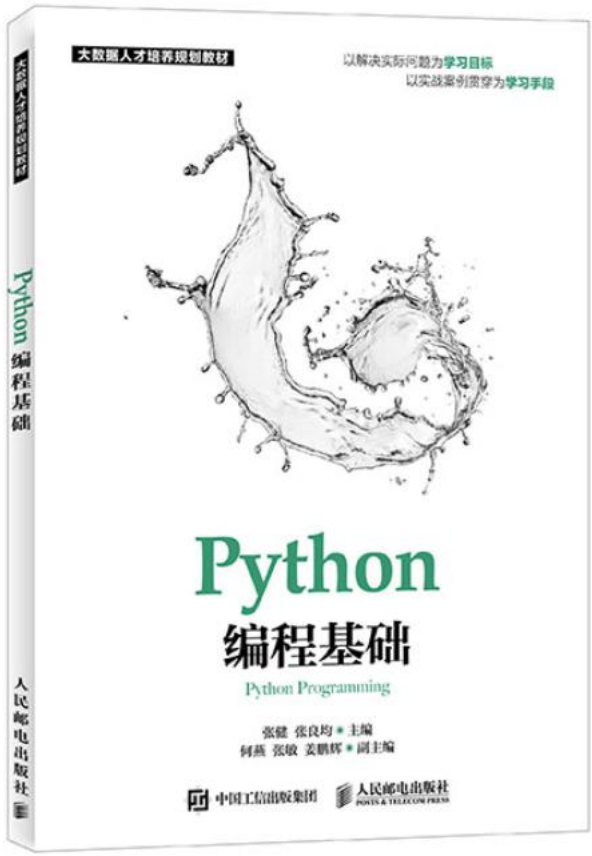
实训 1　数据预处理

实训 2　数据分析与可视化

实训 3　餐饮综合案例报表整理和发布

(3)　　Python 编程基础

一、 图书封面



二、 课程简介: Python 可以用于数据统计、分析、可视化等任务，以及机器学习、人工智能等领域。大量的第三方模块所支持的内容涵盖了从统计计算到机器学习，从金融分析到生物信息，从社会网络分析到自然语言处理，从各种数据库各种语言接口到高性能计算模型等领域。《Python 编程基础》是大数据与人工智能 Python 系列课程的入门课程。课程以任务为导向，能满足完全面向对象的 Python 的高校教学工作，可以作为高校中数学和统计学等专业的基础课程。

三、 课时数：理论教学 24 学时，实践教学 24 学时，总计 48 学时。

四、 课程资源：包含 35 份实验指导书、32 个课程视频、7 份课程 PPT、7 份数据、25 份代码。

五、 课程内容：

第 1 章 准备工作：

1.1 Python 认识

1.2 搭建 Python 环境

1.3 了解常用 Python IDE 并创建一个应声虫程序

第 2 章 Python 基础知识：

2.1 Python 的固定语法

2.2 了解 Python 变量与相互转化数值型变量

2.3 字符串类型基本操作

2.4 字符串内建函数

2.5 掌握常用操作运算符及优先级

第 3 章 Python 数据结构：

3.1 认识数据结构与列表

3.2 列表的增删改查

3.3 列表推导式

3.4 元组

3.5 字典

3.6 集合

第 4 章 程序流程控制语句：

4.1 条件分支语句

4.2 循环

4.3 嵌套循环与多变量迭代、列表解析

第 5 章 函数：

5.1 自定义函数

5.2 调用自定义函数

5.3 嵌套函数、全局变量与局部变量

5.4 匿名函数与高阶函数

5.5 存储并导入函数模块

第 6 章 面向对象编程：

6.1 认识面向对象编程

6.2 类与绑定 self

6.3 类的专有方法

6.4 创建对象

6.5 迭代器

6.6 继承与其他方法

第 7 章 文件基础：

7.1 认识文件、读取整个文件

7.2 with 语句读取文件与设置工作路径

7.3 读取 txt、csv 文件

7.4 os 模块与 shutil 模块

六、 实训目录

第 2 章 Python 基础知识:

实训 1　创建字符串变量并提取里面的数值

实训 2　计算圆形的各参数

实训 3　对用户星座进行分析

实训 4　通过表达式计算给定的三个数值均值、方差、标准差

第 3 章 Python 数据结构:

实训 1　创建一个列表（list）并进行增删改查操作

实训 2　转换一个列表为元组（tuple）并进行取值操作

实训 3　创建一个字典（dict）并进行增删改查操作

实训 4　将两个列表转换为集合（set）并进行集合运算

实训 5　计算出斐波那契数列前两项给定长度的数列，并删除重复项和追加数列各项之和为新项

实训 6　用户自定义查询菜单，输出查询结果

实训 7　简单的好友通讯录管理程序

实训 8　对两个给定的数进行最大公约数、最小公倍数的分析

第 4 章 程序流程控制语句:

实训 1　实现考试成绩划分

实训 2　实现一组数的连加与连乘

实训 3　使用冒泡排序法排序

实训 4　输出数字金字塔

实训 5　猜数字游戏

实训 6　统计字符串内元素类型的个数

第 5 章　函数:

实训 1　自定义函数实现方差输出

实训 2　使用匿名函数添加列表元素

实训 3　存储并导入函数模块

实训 4　构建一个计算列表中位数的函数

实训 5　使用 lambda 表达式实现对列表中的数求平方

第 6 章　面向对象编程:

实训 1　创建 Car 类

实训 2　创建 Car 对象

实训 3　迭代 Car 对象

实训 4　产生 Land_Rover 对象（子类）

实训 5　在精灵宝可梦游戏创建小火龙角色，对给出的各属性进行迭代和私有化

实训 6　对小火龙游戏角色采用继承的方式

第 7 章　文件基础:

实训 1　对 txt 文件进行读写

实训 2　对 csv 文件进行读写

实训 3　os 模块

实训 4　shutil 模块

实训 5　计算 iris 数据集的均值

实训 6　编程实现文件在当前工作路径的查找

(4)　　Python 数据分析与应用

一、　图书封面

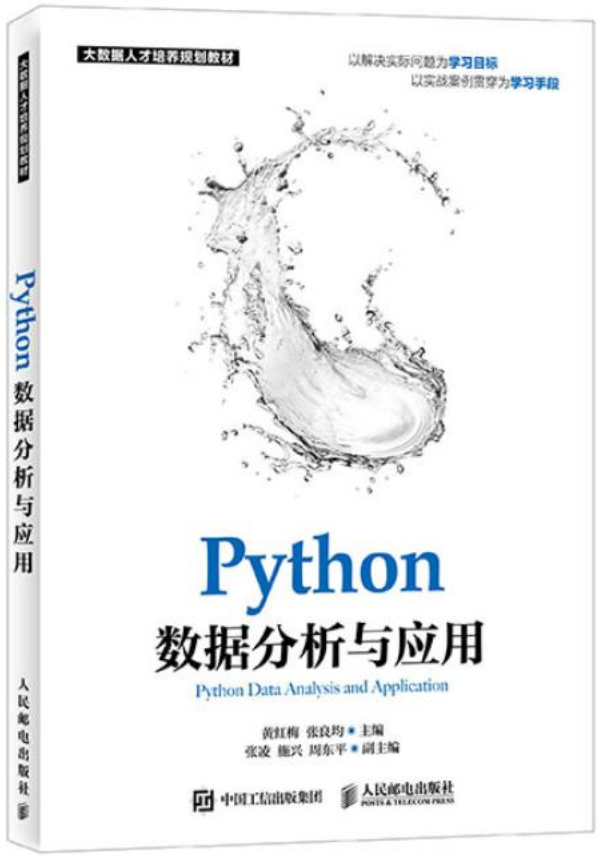


图 5-1 Python 数据分析与应用

二、 **课程简介：**数据分析技术将帮助企业用户在合理时间内获取、管理、处理以及整理海量数据，为企业经营决策提供积极的帮助。数据分析作为一门前沿技术，广泛应用于物联网、云计算、移动互联网等战略新兴产业。《Python 数据分析与应用》课程是大数据与人工智能 Python 系列课程的核心课程。课程以任务为导向，将 Python 数据分析知识点融入其中，能够让学生在练中学，学会即应用。

三、 **课时数：**理论教学 36 学时，实践教学 28 学时，总计 64 学时

四、 **课程资源：**包含 34 份实验指导书、23 个课程视频、7 份课程 PPT、25 份数据、7 份代码。

五、 **课程内容：**

第 1 章 Python 数据分析概述：

1.1 数据分析概述

1.2 熟悉 Python 数据分析的工具

1.3 安装 Python 的 Anaconda 发行版

1.4 掌握 Jupyter NoteBook 常用功能

第 2 章 NumPy 数值计算基础：

2.1 掌握 numpy 数组对象 ndarray_x264

2.2 掌握 Numpy 矩阵与通用函数

2.3 利用 Numpy 进行统计分析

第 3 章 Matplotlib 数据可视化基础：

3.1 掌握绘图基础语法与常用参数

3.2 分析特征间关系

3.3 分析特征内部数据分布与分散情况

第 4 章 pandas 统计分析基础：

4.1 读写不同数据源的数据

4.2 掌握 DataFrame 的常用操作

4.3 转换与处理时间序列数据 2

4.4 使用分组聚合进行组内计算

4.5 创建透视表与交叉表

第 5 章 使用 pandas 进行数据预处理：

5.1 合并数据

5.2 清洗数据

5.3 标准化数据

5.4 转换数据

第 6 章 使用 scikit-learn 构建模型：

6.1 使用 sklearn 转换器处理数据

6.2 构建并评估聚类模型

6.3 构建并评估分类模型

6.4 构建并评估回归模型

六、 实训目录：

第 2 章 NumPy 数值计算基础:

实训 1：掌握 NumPy 数组对象 ndarray

实训 2：掌握 NumPy 矩阵与通用函数

实训 3：利用 NumPy 进行统计分析

实训 4：创建数组并进行运算

实训 5：创建一个国际象棋的棋盘

第 3 章 Matplotlib 数据可视化基础:

实训 1：掌握绘图基础语法与常用参数

实训 2：分析特征间的关系

实训 3：分析特征内部数据分布与分散状况

实训 4：分析 1996~2015 年人口数据各个特征的分布与分散状况

实训 5：分析 1996~2015 年人口数据特征间的关系

第 4 章 pandas 统计分析基础:

实训 1：读写不同数据源的数据

实训 2：掌握 DataFrame 的常用操作

实训 3：转换与处理时间序列数据

实训 4：使用分组聚合进行组内计算

实训 5：创建透视表与交叉表

实训 6：读取并查看 P2P 网络贷款数据主表的基本信息

实训 7：提取用户信息更新表和登录信息表的时间信息

实训 8：使用分组聚合方法进一步分析用户信息更新表和登录信息表

实训 9：对用户信息更新表和登录信息表进行长宽表转换

第 5 章 使用 pandas 进行数据:

实训 1：合并数据

实训 2：清洗数据

实训 3：标准化数据

实训 4：转换数据

实训 5：插补用户用电量数据缺失值

实训 6：合并线损，用电量趋势与线路告警数据

实训 7：标准化建模专家样本数据

第 6 章 使用 scikit-learn 构建模型:

实训 1：使用 sklearn 转换器处理数据

实训 2：构建并评价聚类模型

实训 3：构建并评价分类模型

实训 4：构建并评价回归模型

实训 5：使用 sklearn 处理 wine 和 wine_quality 数据集

实训 6：构建基于 wine 数据集的 K-Means 聚类模型

实训 7：构建基于 wine 数据集的分类模型

实训 8：构建基于 wine_quality 数据集的回归模型

(5)　　Python 网络爬虫实战

一、 图书封面

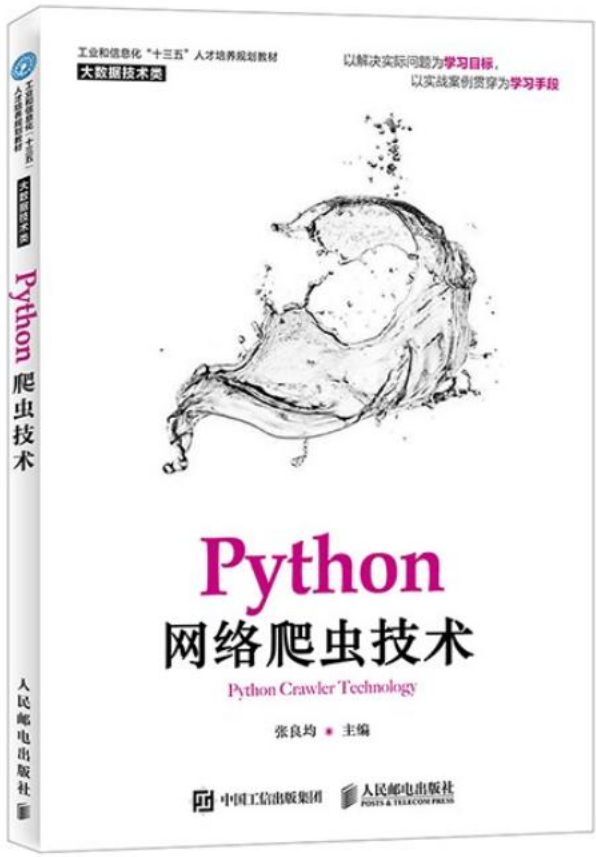


图 5-2　Python 网络爬虫技术

二、 课程简介：网络爬虫（又被称为网页蜘蛛，网络机器人，在 FOAF 社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。《Python 网络爬虫实战》是大数据与人工智能 Python 系列课程的进阶课程。课程以任务为导向，详细陈述了不同网页的爬取，以及最流行爬虫框架的使用。

三、 课时数：理论教学 14 学时，实践教学 18 学时，总计 32 学时

四、 课程资源：包含 27 份实验指导书、37 个课程视频、6 份课程 PPT、2 份数据、4 份代码。

五、 课程内容：

第 1 章 Python 爬虫环境与爬虫简介：

1.1 Python 网络爬虫实战介绍

1.2 认识爬虫

1.3 认识反爬虫

1.4 配置 python 爬虫环境

第 2 章 网页前端基础：

5.7 小结

第 6 章 终端协议分析：

6.1 终端协议分析概述

6.2 了解 HTTP Analyzer 工具

6.3 爬取千千音乐 PC 客户端数据

6.4 小结

六、 实训目录：

第 2 章 网页前端基础：

实训 1：使用 Socket 库进行 TCP 编程

实训 2：使用 Socket 库进行 UDP 编程

实训 3：使用 Socket 库连接百度首页

第 3 章 简单静态网页爬取：

实训 1：urllib3 库实现 HTTP 请求

实训 2：Requests 库实现 HTTP 请求

实训 3：正则表达式模块解析网页

实训 4：Xpath 解析网页

实训 5：Soup 库解析网页

实训 6：MySQL 数据存储

实训 7：生成 GET 请求并获取指定网页内容

实训 8：搜索目标节点并提取文本内容

实训 9：在数据库中建立新表并导入数据

第 4 章 常规动态网页爬取：

实训 1：逆向分析爬取动态网页

实训 2：使用 Selenium 库爬取动态网页

实训 3：存储数据至 MongoDB 数据库

实训 4：爬取网页“httpwww.ptpress.com.cn”推荐图书的信息

实训 5：爬取某网页的 Java 图书信息

实训 6：将数据储存到 MongoDB 数据库中

第 5 章 模拟登录：

实训 1：使用表单登录方法实现模拟登录

实训 2：使用 Cookie 登录方法实现模拟登录

实训 3：使用表单登录方法模拟登录数睿思论坛

实训 4：使用浏览器 Cookie 模拟登录数睿思论坛

实训 5：基于表单登录后的 Cookie 模拟登录数睿思论坛

第 6 章 终端协议分析：

实训 1：爬取千千音乐 PC 客户端数据

实训 2：分析人民日报 APP

实训 3：抓取千千音乐 PC 客户端的推荐歌曲信息

实训 4：爬取人民日报 APP 的旅游模块信息

(6) Hadoop 大数据技术基础

一、 图书封面

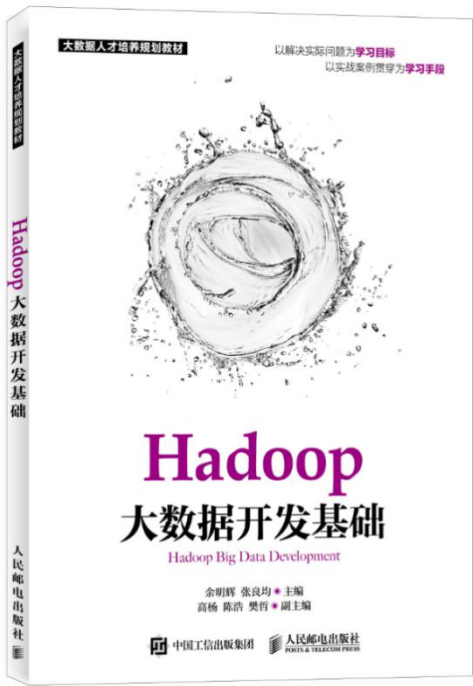


图 5-3 Hadoop 大数据技术基础

二、 课程简介: Hadoop 作为处理大数据的分布式存储和计算框架，得到了国内外大小型企业广泛的应用。Hadoop 是一个可以搭建在廉价 x86 服务器上的分布式集群系统架构，它具有可用性高、容错性高和可扩展性高等优点。《Hadoop 大数据技术基础》大数据技术系列课程的核心课程。课程详述了 Hadoop 提供的开放式平台，可以在完全不了解底层实现细节的情形下，开发适合自身应用的分布式程序。

三、 课时数：理论教学 20 学时，实践教学 28 学时，总计 48 学时

四、 课程资源：包含 23 份实验指导书、70 个课程视频、8 份课程 PPT、41 份数据、69 份代码。

五、 课程内容：

第 1 章 Hadoop 介绍：

1.1.1 Hadoop 简介与发展历史

1.1.2 Hadoop 的特点

1.2.1 分布式文件系统-HDFS

1.2.2 分布式计算框架-MapReduce

1.2.3 集群资源管理器-YARN

1.2.4 分布式公共设施—Common

1.3 Hadoop 生态系统

1.4 Hadoop 应用场景

第 2 章 Hadoop 集群的搭建及配置：

2.1.1 创建 Linux 虚拟机

2.1.2 设置固定 IP

2.1.3 远程连接虚拟机

2.1.4 虚拟机在线安装软件

2.2.1 在 windows 下安装 Java

2.2.2 在 Linux 下安装 Java

2.3.1 修改配置文件

2.3.2 克隆虚拟机

2.3.3 配置 SSH 免密码登录

2.3.4 配置 ntp 时间同步服务&格式化

2.3.5 启动关闭集群

2.3.6 监控集群

第 3 章 Hadoop 基础操作：

3.1.1 查询集群的存储系统信息

3.1.2 查询集群的计算资源信息

3.2.1 了解 HDFS 文件系统及其命令

3.2.2 掌握 HDFS 的基本操作

3.3.1 了解 Hadoop 官方的示例程序包

3.3.2 提交 MapReduce 任务集群运行

3.4.1 查询多个 MapReduce 任务

第 6 章 电影网站用户性别预测：

6.1.1 KNN 算法简介

6.1.2 KNN 算法原理及流程

6.2.1 获取数据

6.2.2 数据变换-数据连接

6.2.3 数据变换-数据统计

6.2.4 数据清洗逻辑

6.2.5 数据清洗的实现过程

6.2.6 划分数据集

6.3.1 实现思路

6.3.2 代码实现

6.4.1 评价思路

6.4.2 实现分类评价

6.4.3 寻找最优 K 值

6.4.4 KNN 算法优缺点

六、 实训目录：

第 2 章 Hadoop 集群搭建：

实训 1：Java 安装及 Hadoop 完全分布式集群搭建

第 3 章 Hadoop 基础操作：

实训 1：查看 HDFS 上的文件内容

实训 2：查看 Hadoop 集群的基本信息

实训 3：上传文件到 HDFS 目录

实训 4：运行首个 MapReduce 任务

实训 5：统计文件中所有单词的平均长度

实训 6：查询与中断 MapReduce 任务

第 4 章 MapReduce 编程入门：

实训 1：使用 Eclipse 创建 MapReduce 工程

实训 2：编程实现按日期统计访问次数

实训 3：编程实现按访问次数排序

实训 4：获取成绩表的最高分记录

实训 5：实现对两个文件中数据的合并与去重

实训 6：统计 Hadoop 出现的次数

第 5 章 MapReduce 编程进阶：

实训 1：筛选日志文件生成序列化文件

实训 2：API 读取序列化日志文件

实训 3：优化日志文件统计程序

实训 4：Eclipse 提交日志文件统计程序

实训 5：统计全球每年的最高气温和最低气温

实训 6：筛选气温在 15 到 25 度之间的数据

实训 7：计算学生平均成绩

实训 8：QQ 好友推荐

第 6 章 基于大数据构建观影用户画像：

实训 1：电影网站用户性别预测之数据预处理

实训 2：电影网站用户性别预测之模型构建与寻优

(7) Python 中文自然语言处理实战

一、 **课程简介：**早在上个世纪，已存在人工文本分析挖掘，并广泛应用在密码学等领域，由于技术的受限，这项技术得不到很好的传承与推广，直到近十几年科技的进步使这一领域迅速发展。文本挖掘已广泛应用于信息检索、自动问答、数据挖掘、语言翻译等领域。《Python 文本挖掘实战》是大数据与人工智能 Python 系列课程的实战课程。课程着重讲解了文本信息转化为数据，进行建模分析，提炼出核心内容、分析文本数据之间的关系等内容，是学习文本挖掘的首选课程。

二、 **课时数：**理论教学 16 学时，实践教学 16 学时，总计 32 学时

三、 **课程资源：**包含 9 份实验指导书、18 个课程视频、2 份课程 PPT、9 份数据、14 份代码。

四、 **课程内容：**

第 1 章 文本预处理技术：

1.1 文本挖掘概述

第 2 章 常见文本分类器及评估：

2.1 文本预处理_正则表达式

2.2 中文分词概述

2.2.1 机械分词法

2.2.2 马尔科夫链分词法

2.2.3 隐马尔可夫模型（HMM）

2.2.4 viterbi 算法

2.2.5 隐马尔可夫与 viterbi 算法应用

2.2.6 jieba 库_jieba 分词

2.3 绘制词云

第 3 章 文本向量化表示：

3.1 文本向量化表示

第 4 章 垃圾短信分类模型构建：

4.1 案例：垃圾短信识别_数据抽取

4.2 案例：垃圾短信识别_文本清洗

4.3 案例：垃圾短信识别_分词与去除停用词

4.4 案例：垃圾短息识别_绘制词云

4.5 案例：垃圾短信识别_文本向量化表示

4.6 案例：垃圾短信识别_文本分类器

4.7 案例：垃圾短信识别_分类模型评估

五、 实训目录：

第 1 章 文本预处理技术：

实训 1：正则表达式

实训 2：中文分词：匹配法

实训 3：中文分词：HMM

实训 4：中文分词：HMM 的维特比算法实现分词

实训 5：绘制词云

第 2 章 垃圾短信分类模型构建：

实训 1：文本分类：数据探索

实训 2：文本分类：数据预处理

实训 3：文本分类：绘制词云图

实训 4：文本分类：识别垃圾短信

(8) Python 数据分析与挖掘实战

一、 课程简介:算法的相关任务往往会受到数据变化、计算能力和经验性判断等的限制。《Python 数据分析与挖掘实战》是大数据与人工智能 Python 系列课程的核心课程。课程深入讲解了机器学习中的常用算法，详细陈述

了每种算法解决问题时的思路。让学员掌握各个算法的应用场景，算法理论基础，编程实现、模型评价体系等，为后续课程的学习及从事数据挖掘的开发和项目业务奠定基础。

二、 **课时数：**理论教学 36 学时，实践教学 28 学时，总计 64 学时

三、 **课程资源：**包含 15 份实验指导书、49 个课程视频、9 份课程 PPT、13 份数据、8 份代码。

四、 **课程内容：**

第 1 章 机器学习绪论：

1.1 引言

1.2 基本术语

1.3 假设空间&归纳偏好

第 2 章 模型评估与选择

2.1 经验误差与过拟合

2.2 评估方法

2.3 性能度量

2.4 性能度量 Python 实现

第 3 章 回归分析

3.1 线性回归基本形式

3.2 线性回归模型的 Python 实现

3.3 波士顿房价预测的 Python 实现

3.4 逻辑回归介绍

3.5 研究生入学录取预测的 Python 实现

第 4 章 决策树

4.1 从女生相亲到决策树

4.2 明天适合打球吗

4.3 决策树拆分属性选择

4.4 决策树算法家族

4.5 泰坦尼克号生还者预测—数据预处理

4.6 泰坦尼克号生还者预测—模型构建与预测

4.7 决策树可视化

第 5 章 神经网络

5.1 单个神经元介绍

5.2 经典网络结构介绍

第 10 章 小结

10.1 小结

五、 实训目录：

第 1 模块 回归分析

实训 1：完成波士顿房价预测模型

实训 2：对研究生是否被录取进行预测

第 2 模块 决策树

实训 1：决策树算法自编

实训 2：用决策树算法构建鸢尾花分类模型

第 3 模块 神经网络

实训 1：自定义 sigmoid 激活函数

实训 2：网络输入到输出

实训 3：网络权值和阈值更新

实训 4：网络模型训练

实训 5：网络模型预测

第 4 模块 KNN 与朴素贝叶斯

实训 1：求距离矩阵

实训 2：找邻居

实训 3：归类

实训 4：自编 KNN 算法实现鸢尾花分类

第 5 模块 聚类分析

实训 1：对鸢尾花数据进行 K-Means 聚类

第 6 模块 支持向量机

实训 1：用支持向量机解决鸢尾花分类

(9) TensorFlow2 深度学习实战

一、 课程简介

2016 年，Alpha Go 引爆全球，这一事件极大引发了人们对人工智能（AI）的关注。如今，AI 正深刻改变我们的社会与经济形态，作为人工智能的重要组成部分，深度学习愈发受到学术界和产业界的关注。2006 年，是深度学习元年，Hinton 提出了深层网络训练中梯度消失问题的解决方案：无监督预训练对权值进行初始化+有监督训练微调。2012 年，Hinton 课题组首次参加 ImageNet 图像识别比赛，其通过构建的 CNN 网络 AlexNet

一举夺得冠军，且碾压第二名（SVM 方法）的分类性能。也正是由于该比赛，CNN 吸引到了众多研究者的注意。本课程将带你进入深度学习的世界，主要展示如何利用深度神经网络 CNN 及 RNN 来完成经典任务。

二、 课时数

理论教学 16 学时，实践教学 16 学时，总计 32 学时

三、 课程资源

包含 8 份实训指导书、38 个视频、7 份 PPT、8 份代码、6 份数据。

四、 课程内容

第 1 章 引言

1.1 引言

第 2 章 卷积神经网络 CNN

2.1 卷积神经网络主体结构介绍

2.2 卷积操作：局部连接和权值共享

2.3 卷积过程示例

2.4 卷积网络与全连接网络计算量对比

2.5 非线性映射函数 ReLU

2.6 池化操作

2.7 全连接操作

2.8 多 filter 卷积

2.9 高维输入处理

2.10 卷积操作函数的参数说明

2.11 卷积和池化的代码实现

2.12 结果可视化

第 3 章 循环神经网络 RNN

3.1 循环神经网络引入

3.2 RNN 传输过程介绍

3.3 为什么是隐藏层神经元

3.4 权值训练：随时间反向传播

3.5 RNN 的常见变体

3.6 RNN 识别 MNIST 手写数字介绍

3.7 RNN 核心步骤的代码演示

3.8 RNN 网络搭建前准备：结构参数设置

五、 实训目录

5.4.3.2.项目实训

(1) 家用热水器用户行为分析【BP 神经网络】

一、 资源：

包含视频、PPT、实训指导书、数据、代码。

二、 背景：

居民在使用家用电器过程中，会因地区气候、区域不同、用户年龄性别差异，形成不同的使用习惯。家电企业若能深入了解其产品在不同用户群的使用习惯，开发新功能，就能开拓新市场。

三、 目标：

根据热水器采集到的数据，识别出洗浴事件。

四、 流程：

1) **数据抽取：**从国内某热水器生产厂商处抽取用户的用水数据。
2) **数据预处理：**删除冗余特征；划分用水事件；确定单次用水事件时长阈值；构建用水时长与频率特征、用水量与波动特征；筛选候选洗浴事件。
3) **模型构建：**将数据划分为训练集和测试集，构建神经网络模型，评价神经网络模型。
4) **模型解读：**在洗浴事件的识别上精确率（precision）非常高，达到了 96%，同时召回率（recall）也达到了 70%以上，可以确定此模型是有效并且效果良好的能够用于实际的洗浴事件的识别中。

五、 技术点：冗余特征处理；划分事件；确定阈值；特征构建；神经网络模型。

六、 案例内容：

1) 案例背景
2) 删除冗余特征
3) 划分用水事件
4) 确定单次用水事件时长阈值
5) 构建用水时长与频率特征
6) 构建停顿特征
7) 构建用水量与波动特征
8) 筛选候选洗浴事件
9) 模型构建

七、 实训目录：

实训 1：预处理热水器用户用水数据

实训 2：构建用水行为特征并筛选用水事件

实训 3：构建 BP 神经网络模型

(2) 市财政收入分析及预测【SVR】

一、 资源：

包含视频、PPT、实训指导书、数据、代码

二、 背景：

在我国现行的分税制财政管理体制下，地方财政收入不但是国家财政收入的重要组成部分，而且具有其相对独立的构成内容。如何制定地方财政支出计划，合理分配地方财政收入，促进地方的发展，提高市民的收入和生活质量是每个地方政府需要考虑的首要问题。因此，地方财政收入预测是非常必要的。

三、 目标：

根据历史数据预测 2014 年和 2015 年的财政收入。

四、 流程：

1) 数据抽取：从《统计年鉴》中抽取相关财政的数据。

2) 数据探索：分析数据特征的相关性。

3) 数据预处理：使用 Lasso 回归选取财政收入预测的关键特征。

4) 模型构建：结合使用灰色预测和 SVR 算法构建财政收入预测模型；评价模型。

5) 模型解读：根据模型评价指标可以看出，建立的支持向量回归模型拟合效果优良，可以用于预测财政收入。

五、 技术点：特征的相关性；Lasso 回归；灰色预测算法；SVR 算法，预测模型评价。

六、 案例内容：

1) 财政收入预测背景介绍
2) 数据基本情况介绍
3) 分析目标解读
4) 项目流程介绍
5) 求解 person 相关系数
6) person 相关系数解读
7) 了解 Lasso 回归方法
8) Lasso 回归选取关键特征的实现
9) Lasso 回归数据写出及相应解读
10) 关键特征数据读取及准备
11) GM11 特征值预测
12) GM11 特征数据整理及写出

13) 数据标准化

14) 模型训练及预测

15) 结果可视化

16) 教学目标确认

17) 案例任务点拆解

18) 技能梳理与串联

19) 重难点解析及分享

七、 实训目录：

实训 1：分析财政收入数据特征的相关性

实训 2：Lasso 模型选取财政收入预测的关键特征

实训 3：灰色预测法 GM(1,1)预测各自变量值

实训 4：支持向量回归 SVR 预测财政收入

(3) 城市公交用户出行分析【OD 矩阵模型】

一、 资源：

包含视频、PPT、实训指导书、数据、代码

二、 背景：

城市交通情况对于城市规划，居民城市归属感，城市品牌有着至关重要的影响。大城市的可持续发展，应该立足当前、着眼长远，倡导绿色环保出行，大力优先发展城市公共交通，构建性能优良的交通系统工程，是解决城市交通拥堵的有效手段。

三、 目标：

利用公交车载 GPS 数据与公交刷卡数据，构建模型，分析居民出行规律，并提出城市公交站点设置的优化建议。

四、 流程：

1) 数据抽取：选取某城市的地面公交车 GPS 监控数据和地面公交车刷卡数据。

2) 数据探索：绘制折线图分析 2014 年 6 月 09 日至 2014 年 6 月 13 日每个时间段刷卡的人数。

3) 数据预处理：数据归约；缺失值处理；数据合并。

4) 模型构建：构建 DBSCAN 聚类模型，得到每个站点，并计算每个站点的上下车人数，得到 OD 矩阵。

5) 模型应用：根据每个站的上下车人数，提供站点的优化方案。

五、 技术点：数据归约；缺失值处理；数据合并；DBSCAN 聚类模型；OD 矩阵。

六、 案例内容：

1) 案例背景

2) 数据情况与挖掘目标

3) 分析方法与过程、数据抽取

4) 数据探索

5) 数据预处理

6) 数据读取（Python 实现）

7) 数据预处理（Python 实现）

8) 数据探索（Python 实现）

9) 案例思路与密度聚类分析

10) 构建 OD 矩阵模型

11) 密度聚类（Python 实现）

12) 分时段（Python 实现）

13) 构建 OD 矩阵模型（Python 实现）

七、 实训目录：

实训 1：数据探索分析与预处理

实训 2：模型构建

(4) 电力窃漏电用户识别【随机森林】

一、 资源：

包含视频、PPT、实训指导书、数据、代码

二、 背景：

电力是以电能作为动力的能源。发明于 19 世纪 70 年代，电力的发明和应用掀起了第二次工业化高潮。成为人类历史 18 世纪以来，世界发生的三次科技革命之一，从此科技改变了人们的生活。20 世纪出现的大规模电力系统是人类工程科学史上最重要的成就之一，是由发电、输电、变电、配电和用电等环节组成的电力生产与消费系统。它将自然界的一次能源通过机械能装置转化成电力，再经输电、变电和配电将电力供应到各用户。据统计，全国每年因窃电造成的损失都在 200 亿元左右；被查获的窃电案件不足总窃电案件的 30%。

三、 目标：

根据电力营销系统与计量自动化系统数据，构建窃漏电用户识别模型，自动检测判断是否存在窃漏电行为。

四、 流程：

1) 数据抽取：从营销、计量自动化系统收集目标数据。

2) 数据探索：统计出各个用电类别的窃漏电用户分布情况；随机抽取一个正常用电用户和一个窃漏电用户，采用周期性分析对用电量进行探索；异常用电电量探索；

3) 数据预处理：将非居民用电类别的用电数据过滤掉；过滤节假日的用电数据；插补缺失值。

4) 指标构造：构造电量趋势下降指标；构造线损指标；构造告警指标。

5) 模型构建：将数据划分为训练集和测试集，占比分别为 80%、20%；构建 CART 决策树模型；评价模型。

6) 结果诊断：用构建好的窃漏电用户识别模型计算用户的窃漏电诊断结果，实现窃漏电用户实时诊断，并与实际稽查结果作对比。

五、 技术点：pyplot 图形绘制；缺失值处理；CART 决策树模型。

六、 案例内容：

1) 案例背景
2) 项目案例整体流程
3) 数据抽取
4) 数据探索分析
5) 数据预处理
6) 特征构建
7) 模型构建与评价
8) 代码实现流程梳理
9) 数据探索代码实现
10) 告警指标构建代码实现
11) 随机森林模型构建与评估

七、 实训目录：

实训 1：数据抽取与探索分析

实训 2：数据预处理

实训 3：模型构建

(5) 航空公司客户价值分析【K-Means 聚类】

一、 资源：

包含视频、PPT、实训指导书、数据、代码

二、 背景：

民航的竞争除了三大航空公司之间的竞争之外，还将加入新崛起的各类小型航空公司、民营航空公司，甚至国外航空巨头。航空产品生产过剩，产品同质化特征愈加明显，于是航空公司从价格、服务间的竞争逐渐转向对客户的竞争。随着高铁、动车等铁路运输的兴建，航空公司受到巨大冲击。目前航空公司已积累了大量的会员档案信息和其乘坐航班记录，利用这些记录今夕特征分析可以对不同价值的客户制定相应的营销策略。

三、 目标：

对不同价值的客户类别提供个性化服务，制定相应的营销策略。

四、 流程：

1) 数据抽取：从航空公司处抽取会员档案信息和其乘坐航班记录。

2) 数据预处理：处理数据缺失值与异常值；构建航空客户价值分析关键特征 L、R、F、M、C；标准化 L、R、F、M、C 特征。

3) 模型构建：构建 K-Means 聚类模型，对客户进行分群。

4) 模型应用：根据每个群的特点，可定义五个等级的客户类别：重要保持客户、重要发展客户、重要挽留客户、一般客户、低价值客户，并根据这五个等级的客户制定 3 种营销策略：会员的升级与保级、首次兑换积分、交叉销售。

五、 技术点

缺失值处理；异常值处理；构建特征；K-Means 聚类模型。

六、 案例内容

1) 案例背景
2) 案例目标
3) 数据读取
4) 剔除票价为空的记录
5) 剔除异常记录
6) RFM 模型介绍
7) LRFMC 模型
8) 构造入会时长特征
9) 剩余特征构造
10) 使用 K-means 算法进行客户分群
11) 获取 K-Means 聚类结果
12) 聚类结果可视化
13) 小结

七、 实训目录

实训 1：预处理航空客户数据

实训 2：使用 K-Means 算法进行客户分群

(6) 广电大数据营销推荐【协同过滤】

一、 资源：

包含视频、PPT、实训指导书、数据、代码

二、 背景：

随着经济的不断发展，人民的生活水平显著提高，对生活质量的要求也在提高。互联网技术的快速发展适应了时代的需求。为人们提供了许多娱乐的渠道。其中“三网融合”为人们在信息化时代利用网络等高科技手段获取所需的信息提供了极大的便利性。

在三网融合的大背景下，广播电视运营商与众多的家庭用户实现信息实时交互。广电为了提升自身的竞争力，需要不断提高用户体验，基于已有数据挖掘其价值。

三、 目标：

利用产品信息数据，对用户提供个性化精准推荐服务，有效提升用户的转化和生命周期价值。

四、 流程：

1) **数据抽取：**从某集团的大数据平台抽取收视行为信息数据、账单数据、订单数据、收费数据及用户状态数据。
2) **数据预处理：**在收视行为信息数据中，去重，处理异常值数据；在账单数据与收费数据中，删除特殊线路的用户和政企用户；在订单数据中，去重，删除与分析无关的数据，选择符合时间规则的数据；在用户状态数据中，删除与分析无关的数据。
3) **数据探索：**绘制条形图查看用户观看总时长，绘制折线图查看付费频道与点播回看的周观看时长；对比分析工作日与周末观看时长；对所有收视频道名称的观看时长与观看次数进行贡献度分析；建立标签库；构建用户画像（客户特征、业务特征、兴趣爱好）；基于每个数据，构建相关特征；通过爬虫来获取一些新的产品标签数据。
4) **模型构建：**构建基于物品的协同过滤算法的推荐模型；构建基于 Simple TagBased TF-IDF 算法的标签推荐模型；构建 Popular 流行度推荐模型。
5) **模型解读：**计算分类准确度、召回率指标，对比基于物品的协同过滤算法的推荐模型与 Popular 流行度推荐模型的性能，可以发现协同过滤算法推荐效果优于流行度算法。

五、 **技术点：**重复值处理；异常值处理；pyplot 图形绘制；用户画像；构建特征；爬虫；基于物品的协同过滤算法的推荐模型；基于 Simple TagBased TF-IDF 算法的标签推荐模型；Popular 流行度推荐模型。

六、 案例内容：

1) 背景与目标
2) 目标分析与解读
3) 数据介绍
4) 收视数据探索
5) 异常数据探索
6) 收视数据处理介绍

7) 去除特殊线路和政企用户记录

8) 去除直播记录中不关机顶盒的数据记录

9) 去除累计超过 3 小时或小于 4 秒的直播记录

10) 订单数据预处理介绍

11) 订单数据处理-业务品牌和用户状态筛选

12) 订单数据预处理-产品失效时间和去重处理

13) 保存预处理后的数据

14) 用户观看电视时长可视化

15) 热门频道的可视化

16) 构建标签库介绍

17) 客户标签的计算方式

18) 产品标签体系

19) 客户标签体系介绍

20) 客户标签计算方法

21) 电视依赖度标签计算-低

22) 电视依赖度标签计算-中高

23) 用户画像构建

24) 协同过滤推荐

25) 基于流行度的推荐

26) 案例小结及平台呈现

七、 实训目录：

实训 1：数据准备

实训 2：构建用户画像

实训 3：客户价值分析

实训 4：节目信息的获取

实训 5：构建基于物品的协同过滤推荐模型

实训 6：构建基于 Simple TagBased TF-IDF 的标签推荐模型

实训 7：构建 Popular 流行度推荐模型

实训 8：模型评价与结果分析

(7) 大数据法律服务智能推荐【协同过滤】

一、 资源：

包含视频、PPT、实训指导书、数据、代码

二、 背景：

随着电子商务规模的不断扩大，商品个数和种类快速增长，顾客需要花费大量的时间才能找到自己想买的商品。这种浏览大量无关的信息和产品过程无疑会使淹没在信息过载问题中的消费者不断流失。为了解决这些问题，个性化推荐系统应运而生。以帮助电子商务网站为其顾客购物提供完全个性化的决策支持和信息服务。

三、 目标：

深入了解用户访问网站行为，对不同需求的用户进行相关的服务页面的推荐。

四、 流程：

1) **数据抽取：**从某法律网站出抽取用户访问网页的数据。
2) **数据探索：**分析网页类型；网页点击情况绘图。
3) **数据预处理：**删除不符合规则的网页；还原翻页网址；划分网页类别。
4) **模型构建：**将婚姻知识类的数据转换成 0-1 二元型数据，使用基于物品的协同过滤算法对数据进行建模，并对预针对每个用户进行推荐。

五、 技术点：类型统计；正则匹配；字符串数据处理；基于物品的协同过滤算法的推荐模型。

六、 案例内容：

1) 智能推荐介绍
2) 背景与挖掘目标
3) 读取用户访问数据
4) 统计不同类型的网页访问次数
5) 探索 101 和 1999 类型网页的访问次数
6) 统计用户点击次数
7) 网页点击分析
8) 翻页网址探索
9) 脏数据探索
10) 脏数据处理规则汇总
11) 脏数据处理操作实现
12) 翻页网址处理
13) 协同过滤推荐算法介绍
14) 提取婚姻类型数据

15) 模型构建准备工作介绍

16) 将用户划分为训练用户和测试用户

17) 将数据集划分为训练集和测试集

18) 构建用户物品矩阵

19) 自定义函数求杰卡德相似系数

20) 模型构建-求物品相似度矩阵

21) 模型推荐及性能评价方法介绍

22) 构建测试用户网址浏览字典

23) 模型推荐

24) 模型性能评价

25) 教学目标确认

26) 案例任务点拆解

27) 技能梳理与串联

28) 重难点解析

29) 教学技巧分享

七、 实训目录：

实训 1：读取数据

实训 2：数据探索分析

实训 3：数据预处理

实训 4：模型构建与评价

(8) 《流浪地球》豆瓣影评采集【Selenium】

一、 资源：

包含视频、PPT、实训指导书、数据、代码

二、 概要：

2019 年 2 月 5 日电影《流浪地球》正式在中国内地上映，业界明星都对该电影给予极高的评价，可是公映后，豆瓣评分却一度下降，观众对该电影的评价呈现 2 个极端，甚至已经演变成一场失控的舆论混战。

三、 目标：

根据豆瓣对《流浪地球》的短评数据进行文本挖掘及可视化的操作。

四、 流程：

1) 数据抽取：通过爬虫获取评论数据。

2) 数据处理：删除不符合分析的字符串符号。

3) 统计分析：绘制词云图展示总体评论；绘制词云图展示好评与差评；统计评分；绘制时序图查看评论数量随日期、时刻的变化；分析豆瓣评分的时间趋势。

五、 技术点：Selenium 爬虫；XPath 网页解析；数据保存；pyplot 图形绘制。

六、 案例内容：

1) 案例背景与挖掘目标
2) 短评数据爬取介绍
3) 安装 selenium 及配置 chromedriver
4) 获取用户名
5) 获取短评正文
6) 设置 cookies
7) 获取用户居住地和入会时间信息
8) 单页数据整理
9) 自定义获取单页数据的函数
10) 判定网页是否已被加载
11) 翻页爬取
12) 代码整理及小结
13) 短评正文数据预处理
14) 词频统计
15) 绘制整体评论数据的词云图
16) 好评差评词云图绘制及小结
17) 评分分数分布统计
18) 短评数量与日期的关系
19) 短评数量与时刻的关系
20) 不同评分数量与时间的关系
21) 评论最多的前十个城市
22) 评分数量与城市的关系
23) 总结
24) 教学目标确认
25) 案例任务点拆解
26) 技能梳理与串联
27) 重难点解析

28) 教学技巧分享

七、 实现过程：

实训 1：获取豆瓣短评数据

实训 2：分析好评与差评的关键信息

实训 3：分析评论数量及评分与时间的关系

实训 4：分析评论者的城市分布情况

(9) 电商产品评论数据情感分析【LDA 模型】

一、 资源：

包含视频、PPT、实训指导书、数据、代码

二、 背景：

网购盛行，许多人都能够上网网购，电商平台之间的竞争十分激烈。如今消费者的反馈通畅，并且在消费评论蕴含丰富信息。分析信息能够知道消费者的意见和评价。

三、 目标：

对京东平台上的热水器评论做文本挖掘分析，分析某一热水器的用户情感倾向，从评论文本中挖掘出该热水器的优点与不足。

四、 流程：

1) 数据抽取：获取‘美的’的评论数据。

2) 数据预处理：对评论数据进行文本去重、停用词去除、分词操作。

3) 模型构建：通过 LDA 模型对评论数据进行主题分析，形成 3 个主题。

4) 模型解读：主题 1 反映了美的热水器安装收费和售后服务问题；主题 2 反映的是美的热水器不满足用户需求等；主题 3 反映了美的热水器自己安装的问题。从热水器的质量和服务人员的素质上提升竞争力。

五、 技术点：文本去重、文本分词、LDA

六、 案例内容：

1) 背景与目标

2) 数据介绍

3) 数据读取及简单查看

4) 剔除换行符

5) 去除评论数据中的产品型号信息

6) 去除 html 语言中的表情符号

7) 文本去重

8) 分词及去停用词

9) 词云绘制

10) 文本情感分析介绍

11) 读取所需词表

12) 计算情感词分数

13) 程度副词计算

14) 否定词计算

15) 程度副词和否定词融合

16) 自定义分值计算函数

17) 所有评论数据的情感得分

18) 保存处理后的评论数据

19) LDA 主题模型介绍

20) 读取好评数据

21) LDA 主题模型构建

22) 小结

七、 实训目录：

实训 1：数据预处理

实训 2：分词并去除停用词

实训 3：根据情感评分划分正面评论与负面评论

实训 4：构建主题模型

(10) 水产养殖水质智能识别【决策树】

一、 资源：

包含视频、PPT、实训指导书、数据、代码

二、 背景：

水产养殖的关键因素之一是水质，养殖水体生态系统的平衡状况可通过水质颜色体现而传统水质监控的关键是行家。在这种过程中，行家判断存在着局限性：对个人经验要求高，存在主观性引起的观察性偏差观察结果的可比性、可重复性不高，不易推广应用。

三、 目标：

根据水质图片，利用图像处理技术和相应模型，实现水质的自动评价。

四、 流程：

1) 数据抽取：抽取某地区多个罗非鱼池水样图片数据。

2) 数据预处理：使用图像切割提取水样图像中央部分具有代表意义的图像；对切割后的图像提取其颜色矩，作为图像的颜色特征。

3) 构建分类模型：对建模数据进行数据标准化；划分训练集与测试集；构建支持向量机（SVM）模型。

4) 模型评价：将测试集带入构建的模型，得到预测结果；使用混淆矩阵评价水质。

五、 技术点：图像切割、颜色矩提取、决策树、混淆矩阵。

六、 案例内容：

1) 案例背景与目标
2) 读取一张图片数据
3) 获取图片数据的像素值矩阵
4) 截取图像的有效区域
5) 水质图像特征-颜色矩
6) 三个颜色矩的 Python 实现
7) 如何进行批量化数据转换
8) 自定义函数获取指定路径中的所有图片名称_x264
9) 处理所有图片数据
10) 数据处理代码整理
11) 模型构建与性能评估
12) 教学目标确认
13) 案例任务点拆解
14) 技能点梳理及串联
15) 重难点解析
16) 教学技巧分享

七、 实训目录：

实训 1：数据清洗

实训 2：特征提取

实训 3：建模前数据整理

实训 4：模型构建与评估

(11) 招聘网站数据采集与人才需求分析【Request】

一、 资源

包含视频、PPT、实训指导书、数据、代码

二、 概要

在这个信息高速发展的时代，人才市场网络化的产生，使得网络招聘越来越成为如今社会的主流趋势，它以招聘范围广、方便迅速、不受时空限制等区别于传统招聘的优势成为越来越多求职者和企业单位青睐的招聘渠道，在人力资源招募与配置方面中起着至关重要的作用。同时，随着互联网、云计算和大数据产业的兴起，面对海量的网络数据，数据分析、数据挖掘等相应行业也正快速发展。网络招聘信息反映着各行各业的发展现状，各地区发展水平，不同职业类型对人才基本条件、能力和素质的要求，以及对新兴行业的发展动向都有着最及时有效的传达。因此，对网络招聘信息进行分析研究，了解不同职业领域的需求特点，挖掘兴起的数据类行业相应的人才需求现状及发展趋势，为广大求职者提供正确的就业指导有着重要意义。

三、 目标

1) 爬取招聘网站全国范围内大数据、数据分析、数据挖掘、机器学习、人工智能等相关岗位的招聘信息。
2) 分析比较不同岗位的薪资、学历要求等情况，并进行可视化呈现。
3) 分析比较不同区域、行业对相关人才的需求情况，并进行可视化呈现。
4) 分析比较不同岗位的知识、技能要求。
5) 对大数据人才培养给出相关建议。

四、 流程

1) **数据采集**：从某主流招聘网站中采集大数据相关岗位人才信息
2) **数据预处理**：对采集的各个字段进行预处理
3) **数据分析**：分析比较不同岗位的薪资、学历要求等情况，并进行可视化呈现。
4) **数据分析**：分析比较不同区域、行业对相关人才的需求情况，并进行可视化呈现。
5) **数据分析**：分析比较不同岗位的知识、技能要求。
6) **模型解读**： 经过对不同职业类型对人才的技术要求分析，得知在互联网的发展下，对于人才的沟通协调能力及自主学习能力都较为注重，且倾向具备一定编程基础、数据库基础的人才，也需要有创造力和管理能力的人才。

五、 技术点

数据采集；XPath 查询语言；字符串处理；数据可视化；数据合并；词云。

六、 案例内容

1) 背景与目标
2) 信息爬取介绍
3) 获取岗位名称数据
4) 获取目录页的所有字段信息
5) 获取二级网址的网页链接

6) 获取二级网址的所有字段信息

7) 对单一目录页中的所有二级网页信息进行抓取

8) 将第一个目录页的数据进行保存

9) 批量爬取及数据保存

10) 已爬取数据介绍

11) 根据岗位名筛选招聘信息

12) 统一岗位名称

13) 根据工资列筛选数据

14) 完成工资数据处理

15) 工作地点字段处理

16) 公司类型字段处理

17) 行业字段数据处理

18) 工作描述字段处理

19) 公司规模字段处理

20) 数据预处理小结

21) 热门招聘岗位可视化

22) 热门行业及公司招聘分析

23) 热门岗位的工资水平

24) 可视化综合分析

25) 岗位技能分析

26) 总结

七、 实训目录

实训 1 爬取招聘网站岗位信息

实训 2 岗位名称探索和岗位名称标准化

实训 3 工资数据处理

实训 4 工作地点数据处理

实训 5 公司类型数据处理和行业类型数据处理

实训 6 工作描述数据处理和公司人数数据处理

实训 7 热门岗位和热门行业可视化分析

实训 8 热门招聘公司和热门岗位的薪资待遇可视化分析

实训 9 热门行业的薪资待遇和热门城市的工资水平可视化分析

实训 10 热门城市的招聘分布和不同体量企业的薪资待遇可视化分析

实训 11 不同体量公司的用人需求和岗位技能可视化分析

(12) 消费者投诉意见挖掘【LDA】

一、 资源

包含视频、PPT、实训指导书、代码、数据。

二、 概要

12315 消费者投诉举报专线，为了切实解决广大消费者"投诉难"的问题，及时、方便、快捷受理消费者诉求，1999 年 3 月 15 日，国家工商行政管理总局在原信息产业部的大力支持下，在全国统一开通了 12315 消费者申诉举报专用电话。12315 专用号码的启用，进一步畅通了消费者的诉求渠道，更加方便工商部门及时受理和处理消费者申诉举报，更好地保护消费者权益，严厉打击制售假冒伪劣商品的行为，及时有效地查处各类经济违法违章案件，为维护市场经济秩序公平、公正，促进经济健康发展，起到了积极，有效的作用。在 2015 年，全国工商和市场监管部门依托 12315 网络，共处理消费者诉求 777.76 万件，同比增长 2.6%，为消费者挽回经济损失 18.6 亿元，其中，互联网接收量增幅较大；涉及消费维权的举报占两成。

三、 目标：

1） 掌握 pandas 库的常见操作,apply 函数和 lambda 函数的搭配使用。

2） 掌握 matplotlib 的绘图方法。

3） 掌握 jieba 分词法，并绘制出词云图。

4） 掌握 LDA 主题模型。

四、 流程

1) 数据探索：探索投诉举报咨询信息中的品牌分布。

2) 数据预处理：对文本信息进行数据预处理，文本去重、中文分词、停用词过滤。

3) 数据可视化：绘制词云图。

4) 模型构建：文档主题生成模型(LDA)

五、 技术点

jieba 分词，文本挖掘，数据可视化，LDA 文档主题生成模型

六、 案例内容

1) 消费者投诉举报信息意见挖掘

七、 实训目录

实训 1 投诉举报咨询信息中的品牌分布

实训 2 投诉类文本数据的词云图绘制

实训 3 LDA 主题模型构建

(13) P2P 信贷结果预测【逻辑回归】

一、 资源

包含视频、PPT、实训指导书、数据、代码

二、 概要

随着互联网金融飞速发展与居民收入水平的提高，居民的理财观念和消费习惯也有所转变，互联网金融理财观念日益深入人心。大数据时代的到来试得信息化技术逐步完善，P2P 网络贷款平台逐渐成为大众金融消费理财的主要途径。但目前多数 P2P 平台的贷款申请审核工作，依赖风控人员的经验，以线下纯人工审核的方式进行。这种方法效率低下、成本高昂、过于主观。根据客户相关信息，用量化手段提高风险管控效率，是各互联网金融企业的迫切需求。

三、 目标

1) 目标 1：了解 P2P 网络信贷的现状和数据。
2) 目标 2：掌握数据预处理的基本方法或相关函数。
3) 目标 3：掌握逻辑回归的基本原理与方法。
4) 目标 4：掌握 SMOTE 法和 WOE 法。

四、 流程

1) 数据抽取：从某互联网企业获取贷款申请数据。
2) 数据预处理：数据结构化、缺失值处理、属性规约、异常值处理、指标构建。
3) 模型构建：使用逻辑回归算法构建模型，并对模型结果进行分析·。
4) 模型优化：使用 SMOTE 法和 WOE 法对模型进行优化。

五、 技术点

自定义函数，字符串处理，正则表达式，逻辑回归，ROC 曲线，过采样，WOE 法

六、 案例内容

1) 了解 P2P 网络信贷数据与建模流程
2) 读取贷款申请数据并合并数据
3) 对贷款申请数据进行预处理
4) 运用逻辑回归算法构建个人贷款模型和企业贷款模型
5) 运用 SMOTE 法和 WOE 法优化模型

七、 实训目录

实训 1　数据预处理

实训 2　构建个人贷款模型

实训 3　构建企业贷款模型

实训 4　优化个人贷款模型

实训 5　优化企业贷款模型

(14)　乳腺癌与中医症型关联分析【Apriori】

一、 资源

包含视频、PPT、实训指导书、数据、代码

二、 概要

恶性肿瘤俗称癌症，当前已成为危害我国居民生命健康的主要杀手。应用中医药治疗恶性肿瘤已成为公认的综合治疗的方法之一，且中医药治疗乳腺癌有着广泛的适应证和独特的优势。从整体出发，调整机体气血、阴阳、脏腑功能的平衡，根据不同的临床证候进行辨证论治。确定“先证而治”的方向：即后续证侯尚未出现之前，需要截断恶化病情的哪些后续证侯。发现中医症状间的关联关系和诸多症状间的规律性，并且依据规则分析病因、预测病情发展以及为未来临床诊治提供有效借鉴。这样患者在治疗的过程中，医生可以有效的减少西医以及化疗治疗的毒副作用，为后续治疗打下基础。并且还能够帮助乳腺癌患者手术后体质的恢复、生存质量的改善，有利于提高患者的生存机率。

主要实现的挖掘目标为：（1）借助三阴乳腺癌患者的病理信息，挖掘患者的症状与中医证型之间的关联关系。（2）对截断治疗提供依据，挖掘潜性证素。

三、 目标

主要实现的目标为：

（1）借助三阴乳腺癌患者的病理信息，挖掘患者的症状与中医证型之间的关联关系。

（2）对截断治疗提供依据，挖掘潜性证素。

四、 流程

1) 数据获取：本案例采用调查问卷的形式对数据进行搜集，生成原始数据。

2) 数据探索：探索原始数据存在需要进行筛选和转换的数据。

3) 数据预处理：对数据进行清洗、属性规约和属性构造、数据离散化。

4) 模型构建：构建关联规则模型。

5) 模型解读：对关联规则的结果结合业务解读。

五、 技术点

属性构造；数据离散化处理；apriori 模型；模型解读应用。

六、 案例内容

1) 案例背景与挖掘目标

2) 案例流程及数据采集

3) 数据预处理

4) 模型构建

5) 关联规则介绍

6) 事务和项集

7) 支持度与置信度

8) 频繁项集

9) 频繁项集的挖掘过程

10) 代码实现&案例小结

七、 实训目录

实训 1：数据预处理和清洗

实训 2：属性规约和属性构造

实训 3：数据离散化

实训 4：模型构建

(15) 基于医学影像的血管三维重构【最近邻约束】

一、 资源

包含视频、PPT、实训指导书、数据、代码

二、 概要

这个案例的来源于序列图像的计算机三维重建。序列图像的计算机三维重建是应用数学和计算机技术在医学与生物学领域的重要应用之一；是医学和生物学的重要研究方法，它帮助人本由表及里、由浅入深地认识生物体的内部性质与变化，理解其空间结构和形态。

血管是血液流通的通路，其在生命活动中的重要性是众所周知，诊断师在临床中经常需要了解血管的分布、走向等重要信息。理想的血管可以看成是粗细均匀的管道，如何建立其数学模型是图像三维重构的重要一环。

三、 目标

1) 目标 1：计算管道的中轴线与半径，给出具体的算法，

2) 目标 2：绘制中轴线在 XY、YZ、ZX 平面的投影图。

四、 流程

1) 问题描述：了解问题背景，简化问题及合理假设。

2) 解题思路：必要假设及假设验证。

3) 数字图像的读取和显示

4) 血管半径计算：直接搜索法、内切圆算法、切线法。

5) 确定切片的圆心

6) 多圆心处理

7) 拟合平滑处理

8) 模型检验

五、 技术点

图像读取；半径计算；确定圆心；拟合平滑处理。

六、 案例内容

1) 背景与挖掘目标

2) 解题思路

七、 实训目录

实训 1 问题描述

实训 2 解题思路

实训 3 数字图像的读取与显示

实训 4 直觉思维法

实训 5 内切圆算法

实训 6 叠加算法

实训 7 确定切片的圆心

实训 8 多圆心处理

实训 9 拟合平滑处理

实训 10 模型检验

(16) 人口增长与医疗需求预测【ARIMA】

一、 资源

包含视频、PPT、实训指导书、代码、数据

二、 概要

未来的医疗需求与人口年龄结构、数量和经济发展等因素相关，合理预测能使医疗设施建设正确匹配未来人口健康保障需求，是保证深圳社会经济可持续发展的重要条件。然而，现有人口社会发展模型在面对深圳情况时，却难以满足人口和医疗预测的要求。为了解决此问题，本案例根据深圳人口发展变化态势以及全社会医疗卫生需求情

况收集数据、建立针对深圳具体情况的数学模型，预测深圳未来的人口增长和医疗需求。

主要挖掘建模目标：（1）根据 1979 年至 2013 年深圳市人口数据，建立模型预测未来五年深圳市户籍、非户籍、常住人口数量。（2）根据 2000 年、2005 年、2010 年三年的深圳市常住人口年龄结构，预测 2015 年深圳市常住人口年龄结构。（3）以病床需求量为参考，预测 2014 年深圳市医疗需求状况。

三、 目标

1） 数据平稳性判定

2） ARIMA 模型定阶

3） ARIMA 模型预测也评估

4） Leslie 矩阵的构建

5） Leslie 矩阵的改进优化预测

四、 流程

1) 数据获取：本案例采用从政府统计网站收集生成原始数据。

2) 数据探索：探索原始时间序列数据和不同时间节点的人口年龄结构。

3) 数据预处理：对数据进行序列化、归并数据量少的数据。

4) 模型构建：构建 ARIMA 模型和 Leslie 矩阵。

5) 模型解读：对 ARIMA 的预测结果和 Leslie 矩阵预测结果偏差的解读。

五、 技术点

平稳性检验；ARIMA 模型系数确定；残差检验；Leslie 矩阵构建和改进。

六、 案例内容

1) 任务背景和挖掘目标

2) 时间序列的平稳性检验

3) ARIMA 模型定阶

4) ARIMA 模型预测和评估

5) Leslie 矩阵的构建和改进

6) 医疗床位需求预测

七、 实训目录

实训 6　Leslie 矩阵构造和预测

实训 7　Leslie 矩阵改进

实训 8　基于改进的 Leslie 矩阵预测

(17)　金融服务机构资金流量预测【ARIMA】

一、　资源

包含视频、PPT、实训指导书、数据、代码

二、　概要

金融全球化的浪潮正以一种不可抗拒的趋势席卷全球，而伴随着我国加入 WTO（世界贸易组织）以及社会主义市场经济的快速发展，我国的金融市场也随之迅猛发展。一方面，金融活动在将储蓄转化为投资，在疏导社会资金流通，发挥实物资金流量的作用等方面都扮演着重要的角色。另一方面，金融资金流量规模的过快增长，大大地超过实物资金流量的增长，又会给市场带来供不应求，物价上涨等经济不稳定现象。

为了更有效地发挥金融活动对于实体经济的意义，资金流量预测成为金融服务机构一重要任务。

三、　目标

基于用户资金流入与资金流出的记录数据，构建资金流入精准预测模型

四、　流程

1) 数据抽取：从某金融服务机构处抽取用户资金流入与资金流出的记录。

2) 数据探索：绘制时序图、自相关图检验时间序列的平稳性。

3) 数据预处理：使用差分运算处理非平稳序列；检验序列的纯随机性。

4) 模型构建：对模型进行定阶；构建 ARIMA 模型；模型检验。

5) 模型应用：根据模型的平均误差可以看出，ARIMA(0,0,3)(1,1,1)为最优的预测模型。

五、　技术点

平稳性检验；非平稳序列处理；纯随机性检验；模型定阶；ARIMA 模型

六、　案例内容

1) 案例背景

2) 效验数据的平稳性与纯随机性

3) 模型构建

七、　实训目录

实训 1：数据探索与预处理

实训 2：模型构建与评估

5.4.4. 大数据应用沙盘

5.4.4.1.机智过人教学实训沙盘

本沙盘主要是通过深度学习和机器视觉技术实现无序物料的抓取。在教学过程中，重点在于讲授如何抽取及预处理沙盘产生的数据，如何设计机器视觉，如何构建神经网络等相关知识。在实训过程中，学生使用相关知识来实现数据采样，抓取预测，智能学习等多种功能及其组合，通过机器人行为做出直观的反馈验证学习成果。

(1) 机智过人教学实训装置（硬件）

机智过人教学实训装置（简称实训装置）的主要功能是通过采用深度学习技术实现对无序物料抓取，使得学生通过这一工程案例掌握深度学习的基本原理与应用过程。实训装置的结构布局如下图所示，包含机器人、相机、电脑、显示屏等设备。

图 5-4 实训装置的结构布局

实训装置的工作原理如下：由电脑根据深度学习算法提供抓取位置的坐标，机器人接收到信号后，移动到指定位置，将工件由取料盒抓到放料盒。

(2) 机智过人教学实训平台（软件）

机智过人教学实训沙盘的软件部分称为机智过人教学实训平台。该平台采用 C/S 架构，能够充分利用客户机的

算力资源，且界面美观，响应速度快。总共包括了数据源、预处理、构建网络、模型训练、模型发布 5 部分。

（1）数据源在机器学习项目流程中表示从业务系统抽取数据的过程。在机智过人教学实训平台中用于导入和展示数据源，提供了加载数据、数据源预览、翻页等功能，如下图所示。

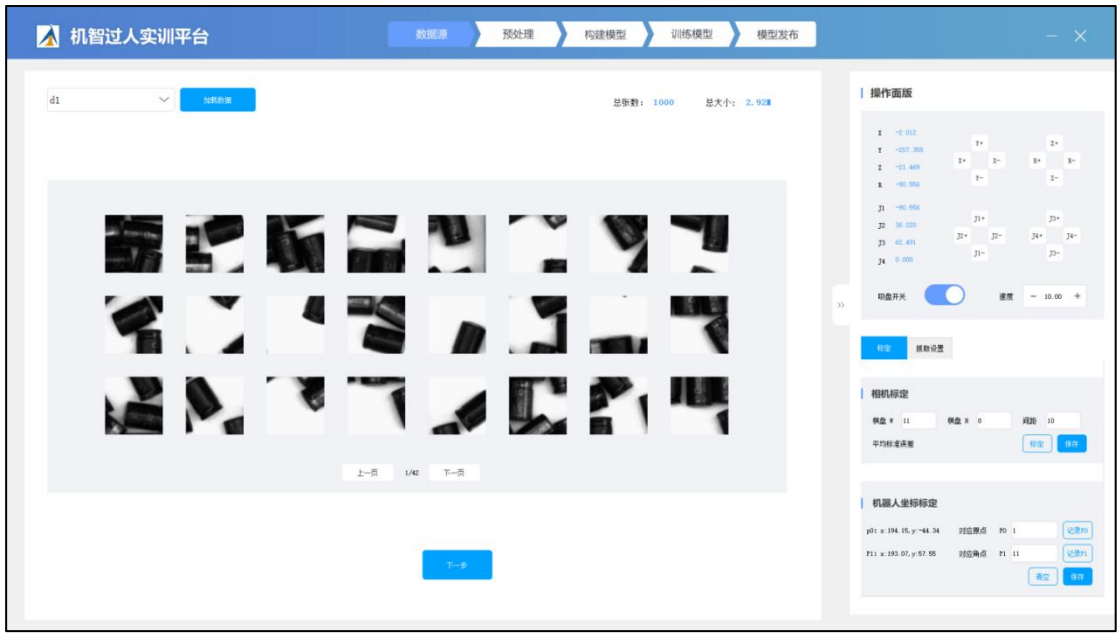
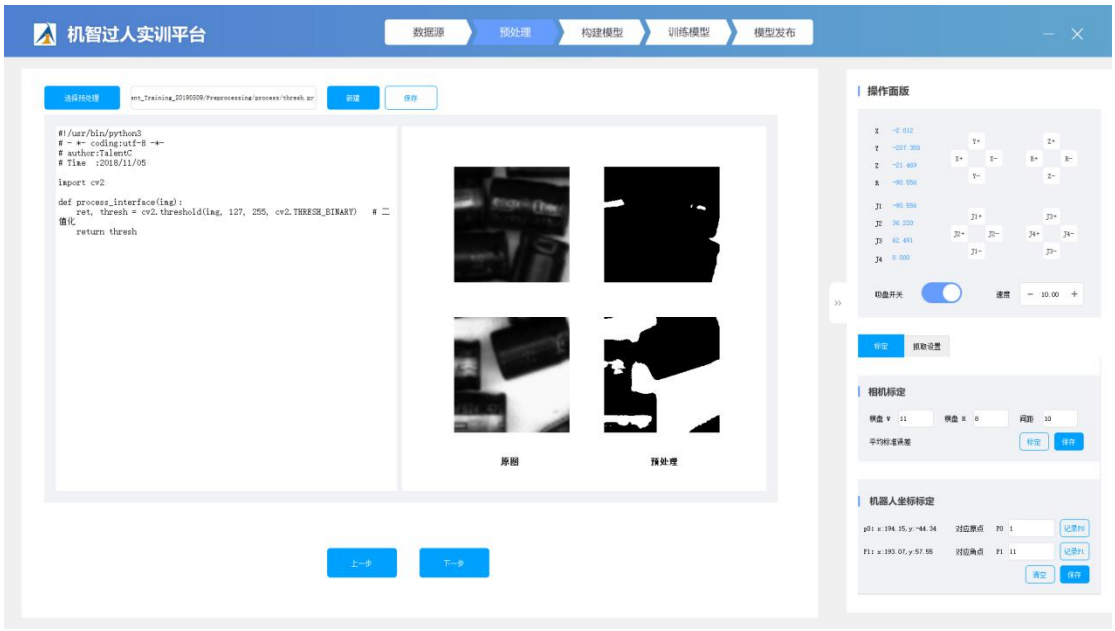
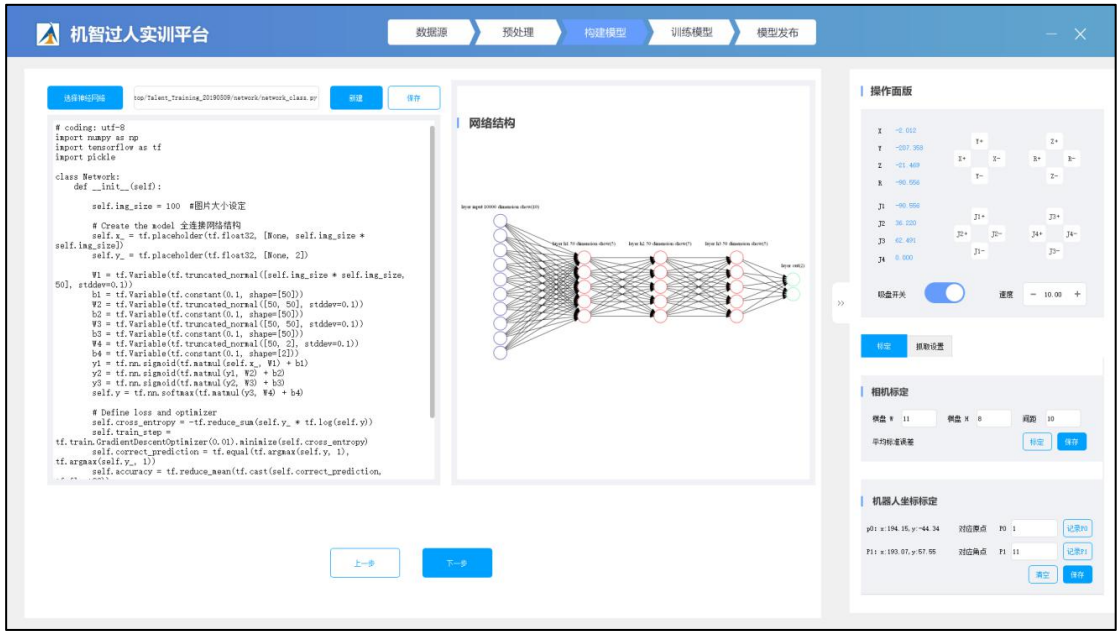


图 5-5　数据源界面

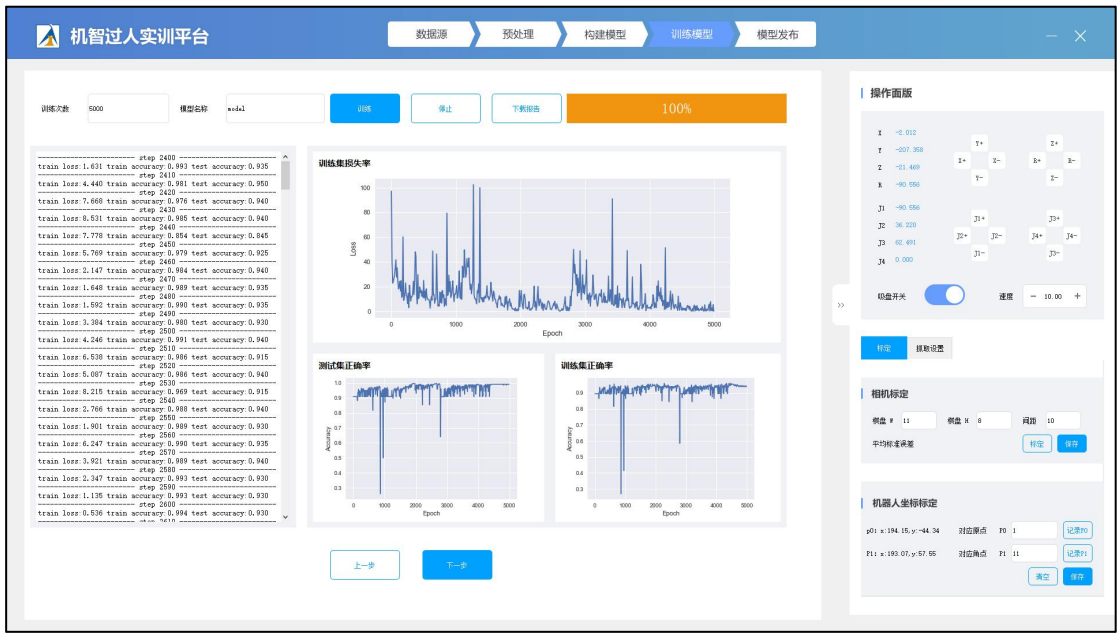
（2）预处理在机器学习流程中表示将数据进行清洗、合并、转换的过程，在机智过人教学实训平台中则表示对图像进行处理的过程，提供了处理组合选择、处理源码查看、处理结果实时预览等功能，如下图所示。

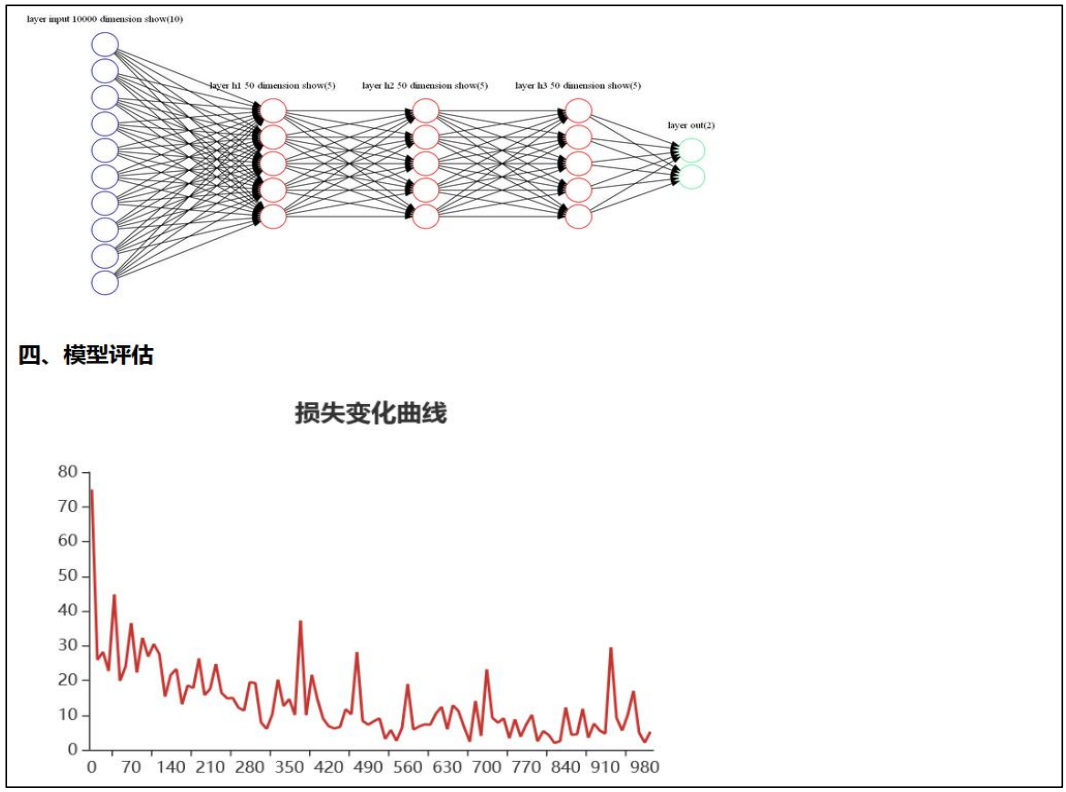


（3）构建模型在机器学习流程中表示模型构建的过程，在机智过人教学实训平台中则表示使用 TensorFlow 构建神经网络的过程，提供了神经网络选择、新建神经网络、保存神经网络、源码查看、网络结构图预览等功能，如下所示。



（4）模型训练部分在机器学习流程中表示训练模型，并做出评价的过程，在机智过人教学实训平台中则表示对上个步骤构建的神经网络模型进行训练，提供了训练次数设置、模型名称设置、训练过程准确率查看、训练集 loss 查看和模型报告下载等功能，如下图所示；模型报告如下图所示。





（5）模型在机器学习流程中表示模型部署于应用的过程，在机智过人教学实训平台中则表示将训练模型应用至机智过人教学实训装置的过程，提供了模型选择、机器人复位、抓取零件位置展示、成功率曲线展示、预测成功率展示和运行日志等功能，如下图所示。

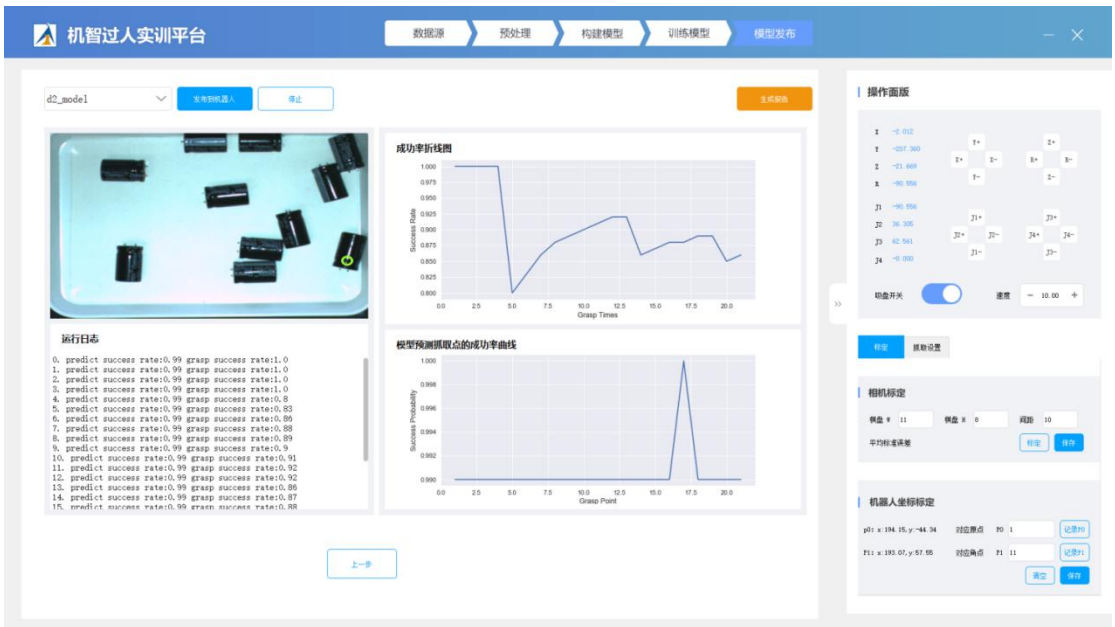


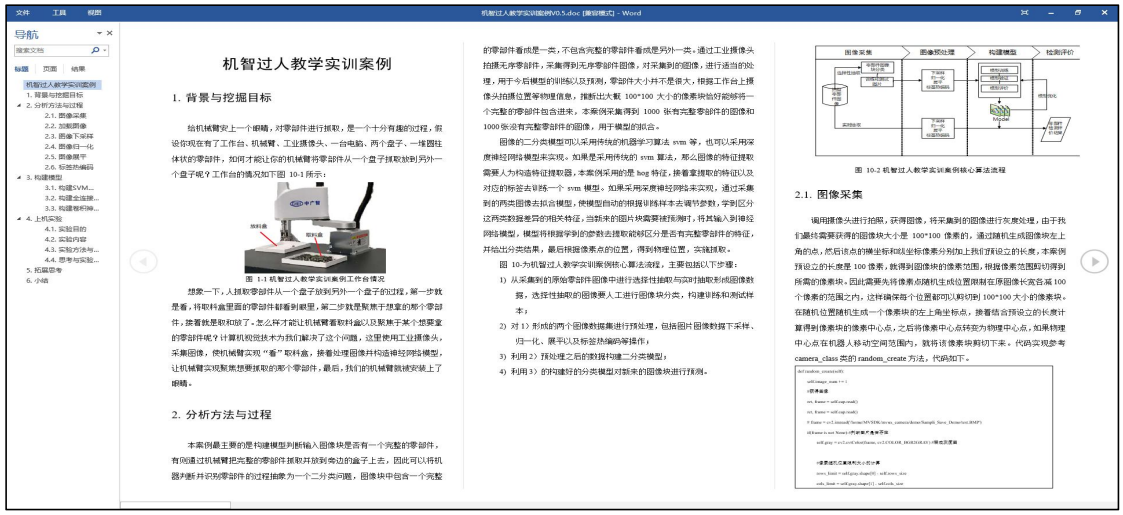
图 5-6　模型发布界面

(3)　配套教学资料（资源）

一、解决方案

解决方案详细地讲解了整个案例的分析过程，如下图所示，包括案例背景、分析方法与过程、模型构建、模型

评价等。同时，还设计了上机实验环节和拓展思考，可作为学员的课后作业。

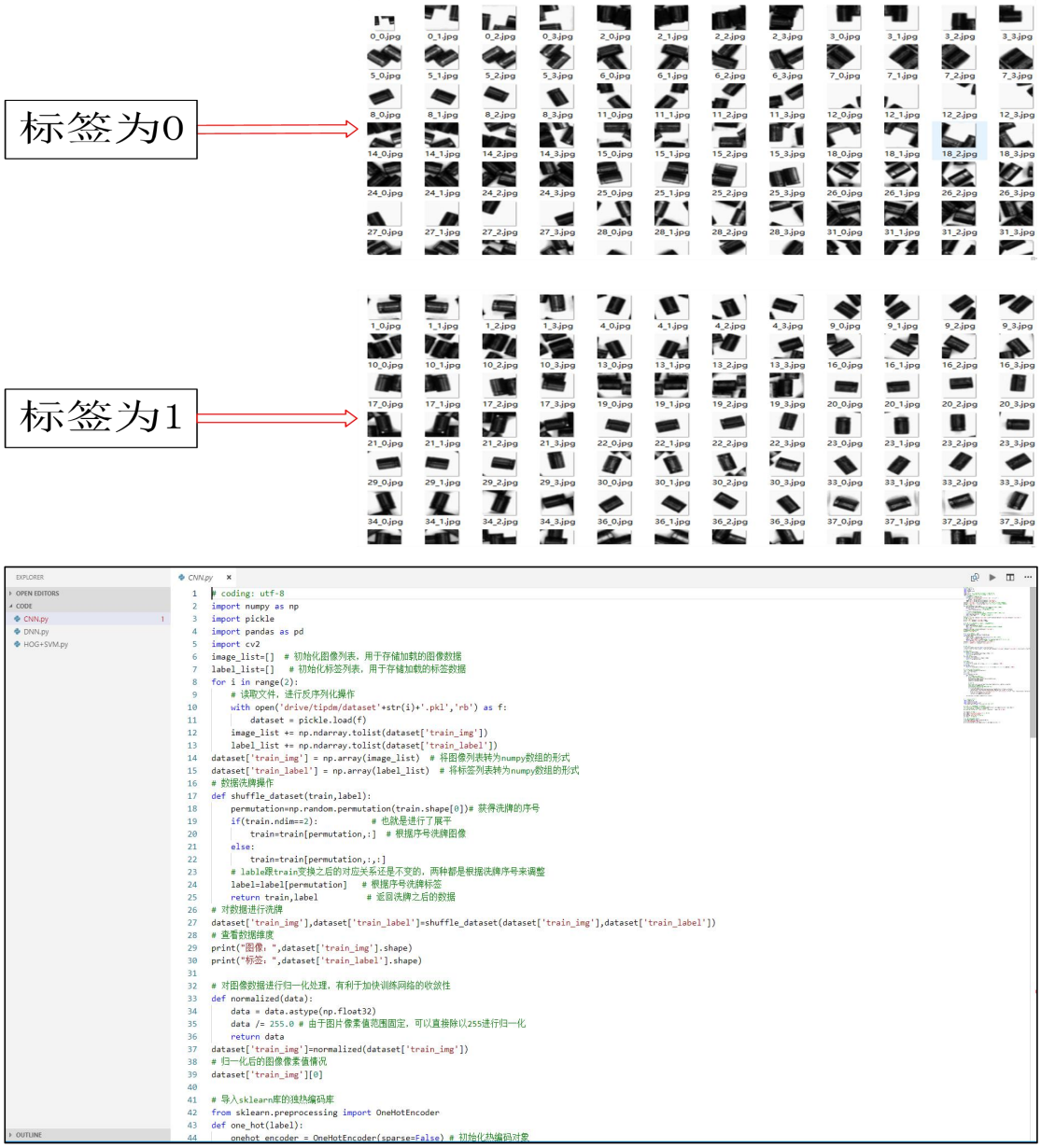


二、教学 PPT

PPT 案例课件是以案例教程为依托，以 PPT 形式展现案例的思路，可作为老师讲课的课件。教学 PPT 的部分截图（包括背景、分析过程等）如下图所示。

三、数据及代码

案例配备的数据和代码，其中数据是真实的企业数据，有利于积累图像处理与机器学习项目经验。数据和代码的部分截图如下图所示。



5.4.4.2.动态人脸识别教学实训沙盘

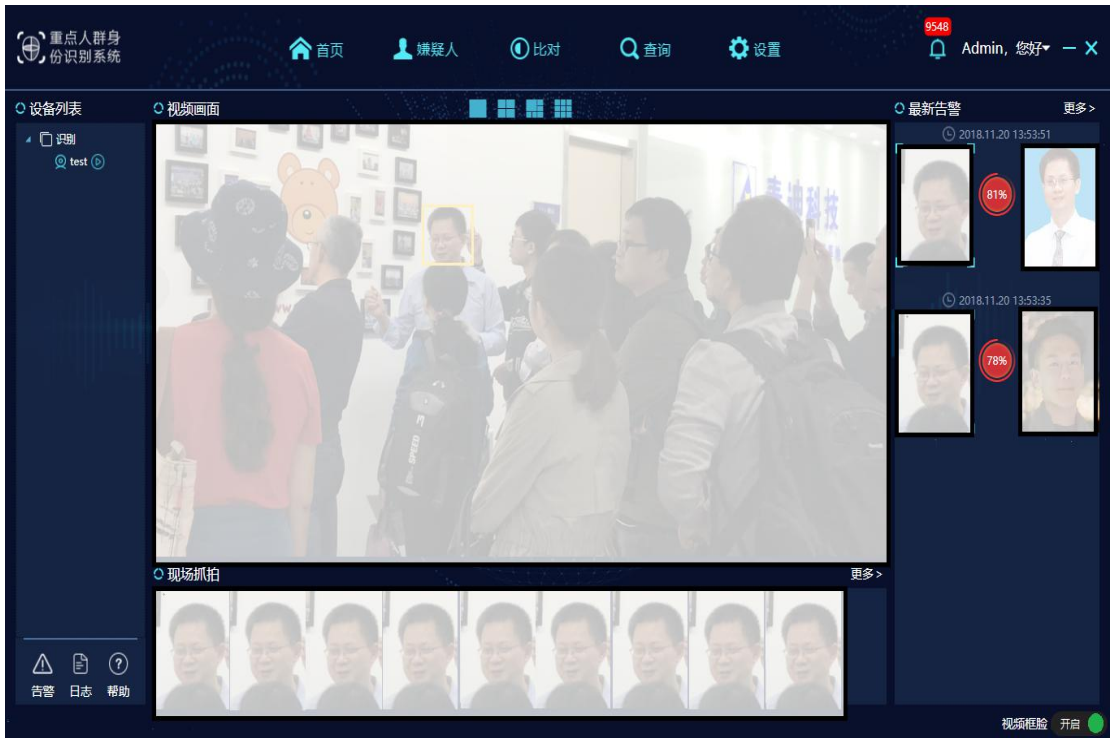
本沙盘主要是通过人脸识别技术实现动态人的身份识别。在教学过程中，重点在于讲授如何抽取及预处理沙盘采集的数据，如何自动监测及跟踪人脸，如何实现人脸对比识别等相关知识。在实训过程中，学生使用相关知识来实现人脸监测，动态跟踪，智能识别等多种功能及其组合，通过系统对比做出直观的反馈验证学习成果。

(1)　　动态人脸识别系统（软件）

动态人脸识别教学实训沙盘的主题就是动态人脸识别系统。该系统采用 C/S 架构，能够充分利用客户机的算力资源，且界面美观，响应速度快。

动态人脸识别系统总共包括了系统主界面，嫌疑人管理，静态比对，查询、告警 5 大板块。

系统主界面主要用于动态人脸识别效果展示。其中，最上部为导航栏，主要用于跳转至其他功能界面；左侧区域为摄像设备列表，用于选定在中央区域展示的设备；中央区域为摄像头实时画面，并可选是否框选出当前画面中识别出的人脸；下部区域主要用于当前摄像画面中识别出的人脸集中展示；右侧区域主要用于人脸比对结果展示，包括当前人脸，数据库中的相似人脸及其相似度。整体界面，如下图所示。



嫌疑人管理板块主要包括了嫌疑人分类管理、嫌疑人信息管理两部分。嫌疑人分类管理提供了嫌疑人分类的添加、修改、删除，嫌疑人管理则提供了嫌疑人的添加、修改、删除，如下图所示。

静态比对板块提供了离线本地图片视频文件的比对，包括了图片对比，视频对比，人脸比对三个功能。图片比对可以单张人脸图像或者多张人脸图像与嫌疑人库中的图像做比对。视频比对可以从视频中提取人脸照片，人脸比

对可以进行人脸 1:1 比对，如下图所示。

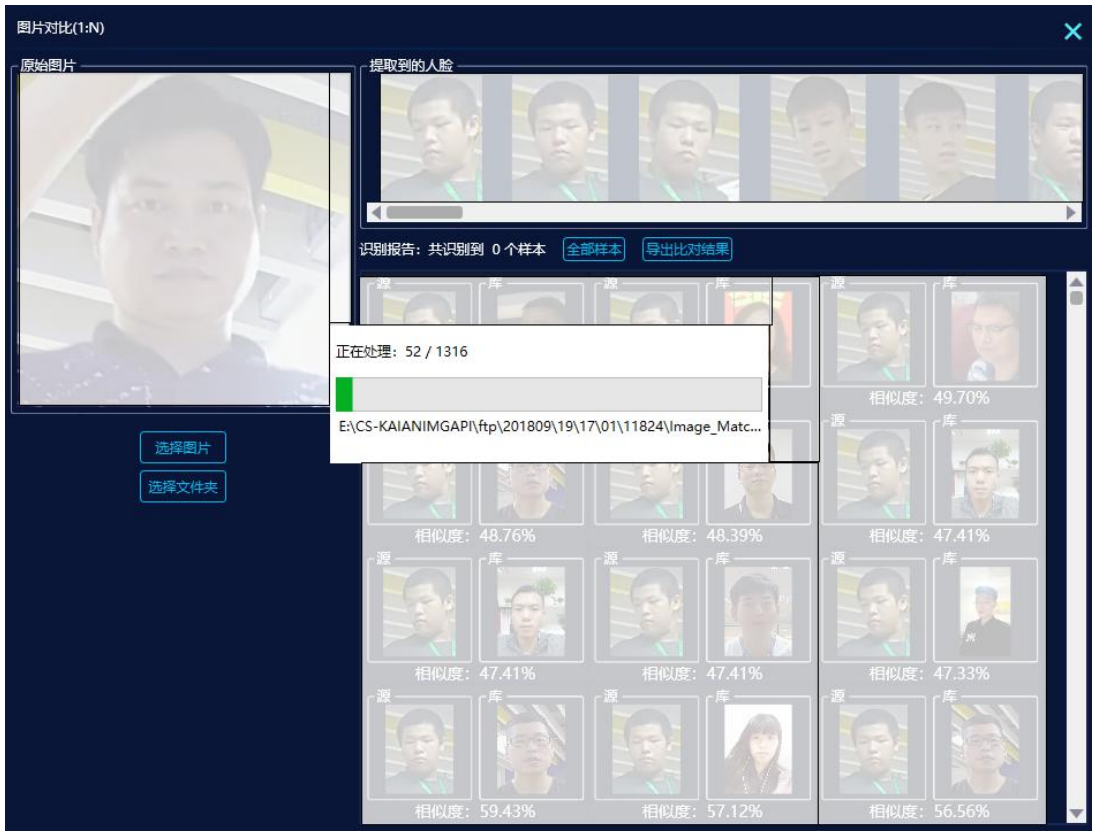


图 5-7-图片对比界面

图 5-8-人脸比对界面

查询板块主要提供了信息查询的功能，包括了历史识别，系统日志和提取图片。历史识别用于查询历史告警信息，系统日志用于查询系统日志信息,包含监控客户端日志、识别服务器日志、网络服务日志，提取图片用于查询系统所有设备提取到的人脸图片，如下图所示。

(2) 配套教学资料（资源）

一、解决方案

解决方案详细地讲解了整个案例的分析过程，如图 3-124 所示，包括案例背景、数据采集、数据预处理和人脸识别等。同时，还设计了上机实验环节和拓展思考，可作为学员的课后作业。

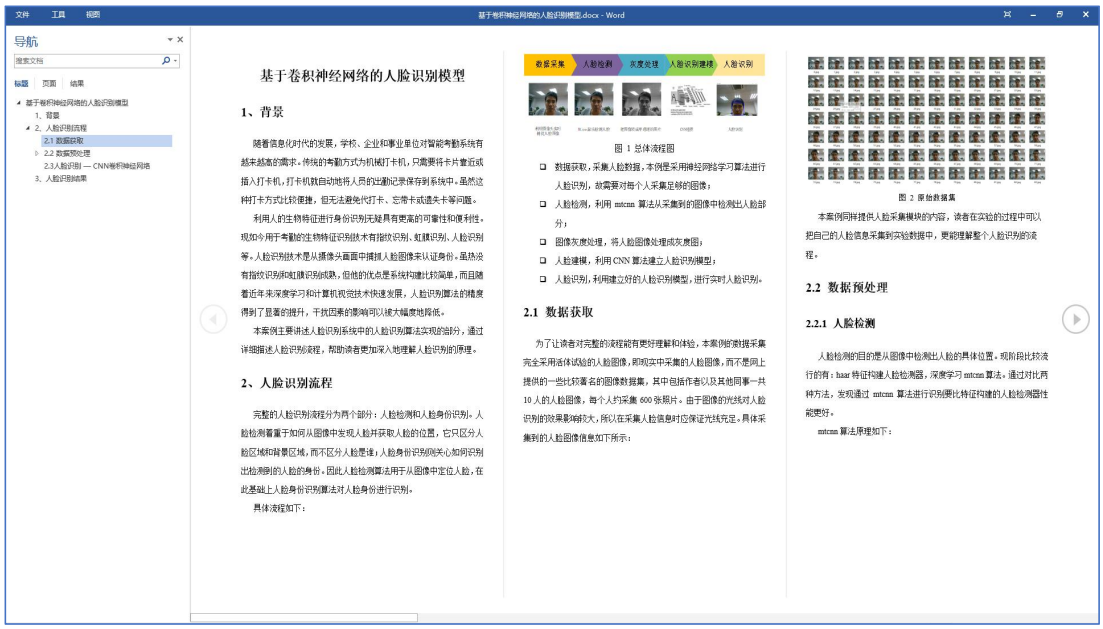


图 5-9-解决方案的总体流程图

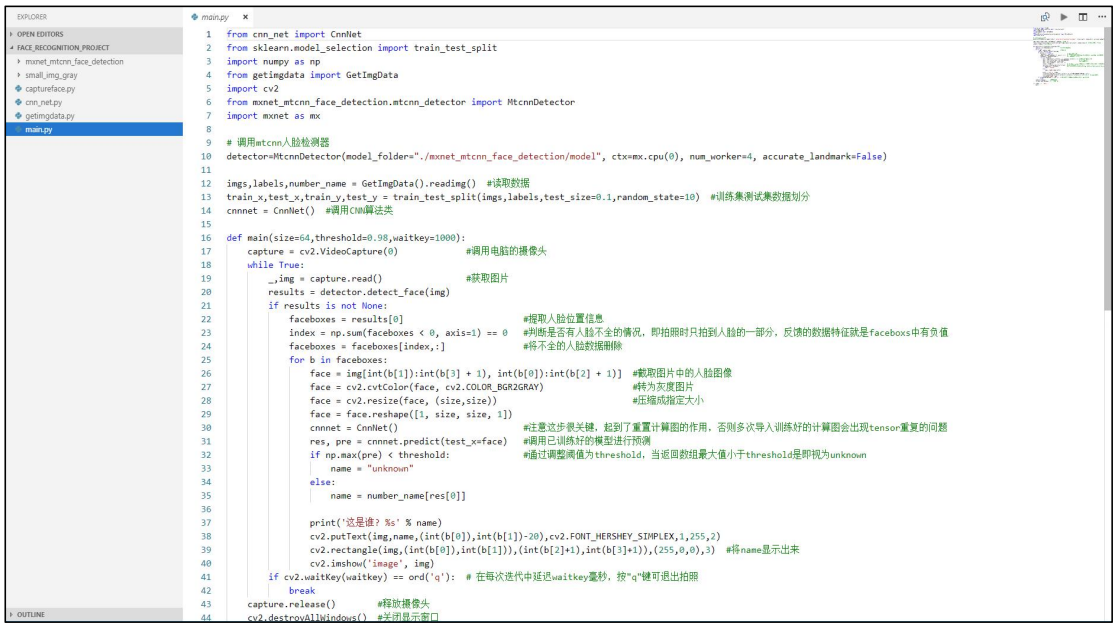
二、教学 PPT

PPT 案例课件是以案例教程为依托，以 PPT 形式展现案例的思路，可作为老师讲课的课件。教学 PPT 的部分截图（包括背景、分析过程等），如下图所示。

三、数据和代码

案例配备的数据和代码，其中数据是真实的企业数据，有利于积累人工智能相关项目经验。数据和代码的部分截图如下图所示。

图 5-10-配套数据



5.4.4.3.新零售智能售卖教学实训沙盘

(1) 产品概述

自动售货机沙盘是广州泰迪科技自主研发的教学实训沙盘，解决高校在培养商业大数据人才过程中，学生能力与企业需求不符的问题。在学习过程中，学生极少接触企业项目，所学理论知识无法准确向解决问题的技能转化，导致学生实战能力无法达到企业要求。沙盘是一个项目的载体，学生结合沙盘进行学习，可以了解企业项目的解决流程，磨练学生的技能，锻炼学生解决问题的能力。

本沙盘体现了自动售货机综合项目的全流程，通过自动售货机采集数据，定时将数据导出至分析数据库，对数据进行处理及分析，并以大屏的形式对数据分析形成的图表进行呈现。在教学中，师生可利用自动售货机数据，进行数据处理及指标构建，绘制分析图，形成分析报表，实现对销售分析、库存分析、用户分析等不同主题的分析。

(2) 产品优势

一、基于真实的零售业务场景：沙盘硬件为真实的自动售货机，学生可与售货机产生交互（购买行为）。处于真实的生产环境中，学生更易于理解业务场景。

二、数据采集维度丰富：盘通过自动售货机采集数据，基于自动售货机的交易、补货等行为，生成订单列表、售货机列表、商品列表、补货单列表、用户列表等数据表。

三、分析结果大屏展示：沙盘配置可视化大屏，对原始数据进行处理及指标构建后，形成的分析图表，可通过大屏分主题呈现。

四、数据持续获取：随着售货机的使用，可不断获取新的售货机数据，扩大数据量。

五、分析结果可验证：基于数据的持续生成，可通过本期实际数据验证上期分析结果，不断优化分析。

六、原始数据定时抽取：沙盘设置定时任务，自动将售货机数据导出至分析数据库，无需人工操作。

(3) 产品硬件构成

一、自动售货机

沙盘硬件为自动售货机，可采集数据。售货机实物参考图如下图所示。

图 5-11-自动售货机

二、可视化大屏

沙盘软件为可视化大屏，可展示分析结果。大屏部分界面如下图所示。

图 5-12-售货机大数据分析平台-总数据页面

(4) 配套资源

沙盘配套资源为一套项目解决方案，包括售货机历史数据，程序，方案文档，教学 PPT。满足老师教学及学生实训需求。

一、数据源

实时提取售货机数据，包含以下数据表。

表名	字段数（个）	记录数（条）
用户列表（uesr_list）	6	13874
售货机列表（box_list）	6	55
类目信息列表（category）	3	60
商品列表（product_list）	7	432
商品详情（product_details）	13	27
订单列表（order_list）	18	30019
订单详情（order_details）	27	39258
补货单列表（supply_list）	9	3351

补货单详情（supply_details）	20	31502

二、程序

提供解决本项目各阶段问题的相关程序。

A. 数据获取：获取接口数据、解析 json 串、嵌套循环解析 JSON 串等。

B. 数据探索：缺失值分析、异常值分析、分布分析、趋势分析、总量分析等。

C. 数据处理：字段选择、过滤记录、记录关联等。

D. 指标构建：销售量、订单量、客单价、商品价格区间、消费时段、存货周转天数等指标的构建

E. 分析数据表生成：销售金额变化趋势分析数据表、商品价格区间分析数据表等。

F. 建模：预测、分群、画像等模型。

三、方案文档

方案文档详细描述项目的解决过程及结果分析，方案目录如下图所示。

自动售货机综合项目

目 录

图 5-13-自动售货机综合项目解决方案

主要解决以下问题。

A. 根据售货机历史数据，对数据质量、数据趋势、数据总量等进行分析，构建销量、库存、用户三个方面的各项指标，并按要求绘制对应图表。

B. 完成销售分析，分析销售情况变化趋势，提出销售策略建议，预测未来某段时间的销量。

C. 完成库存分析，分析库存商品现状，提出商品类型升级建议，优化商品采购。

D. 完成用户分析，分析用户的行为特征，并对用户进行分群，绘制用户画像。

四、教学 PPT

教学 PPT 详略得当的介绍自动售货机综合项目的解决思路，帮助教师讲解本项目。

5.4.4.4.电力智能分项计量教学实训沙盘

(1) 产品概述

随着云时代的来临，大数据更加主流化，为响应国家教育局新增大数据专业的决策，各大高校争相开设大数据相关专业。作为交叉型学科，大数据专业的相关课程涉及数学、统计和计算机等学科知识，数据的分析、处理和建模等实操。这类理论知识晦涩难懂，且实操需要结合具体任务进行练习，才能融会贯通，真正掌握知识。但高校大多仍只通过老师口授理论知识，这样的授课方式过于单一且抽象化，学生不易理解。并且大数据专业为新设专业，高校对教学所需的设施和案例资源储备少，学生缺乏实战经验。

教学实训沙盘，是广州泰迪智能科技有限公司（简称“泰迪智能科技”）基于高校以及其他教育机构开展大数据专业相关课程教学过程中的实际情况，针对传统教学中学生理解浅、知识消化慢、实践环节弱这一短板而设计开发的教学实训沙盘。它能够让老师在讲解理论知识时，通过教学实训沙盘全方位模拟实际应用案例，形象生动地传递信息，让学生深入浅出地掌握知识点，在实验室环境下就能体验实际项目。教学实训沙盘与配套案例的相辅相成，既使得学生受益，又为老师授课带来便利。

电力分项计量沙盘利用家庭、办公、教学等常规用电场所的电力回路，针对整个电力回路进行用电数据的高频采集和分析，结合数学建模方法和深度学习、人工智能算法，实现对电力回路上的用电设备识别和分项计量。在教学过程中，使用电力分项计量沙盘，不仅能让学生将理论知识化抽象为具体，实现理论与实际结合，更能让学生接触实际项目，积累实战经验，真正解决实际生活中的问题。

本产品主要部分包括教学实训装置、后台管理系统、配套教学资料。

主要部分	组件名称	组件说明
硬件	教学实训装置	形状为一个手提箱子，内含总漏电保护、1 个总用电开关、1 个嵌入式网络接口、1 个工控机供电插座、1 个带开关对外供电插座、1 个总电源插口、4 个 USB 接口、1 台交换机、1 台网关(导轨式电表)、1 台主机、1 台显示屏、若干线材。
软件	后台管理系统	管理、储存与展示装置采集得到的用电数据。
配套教学资料	解决方案	详细地讲解了整个案例的分析过程。
	教学 PPT	以 PPT 形式展现案例的思路。

	数据及代码	案例配套的数据和代码。

(2) 产品硬件

沙盘使用工业能源网关终端采集数据，综合了先进的移动通信技术、专业级实时 Linux 嵌入式操作系统、高精度电能计量技术等，具备强大的扩展功能和系统升级空间。并且可以通过设置上传电器设备的阈值、功率抖动、时间间隔、心跳等参数，满足个性化需求。可按多种需求进行电力数据的抄表和存储，如单一设备或混合组设备、用电设备多种模式切换等情况。采集到的数据可通过沙盘软件（后台管理系统）进行管理，利用装置显示屏可实时进行数据预览，也可显示装置对用电设备的识别结果。

图 5-14 沙盘装置外观

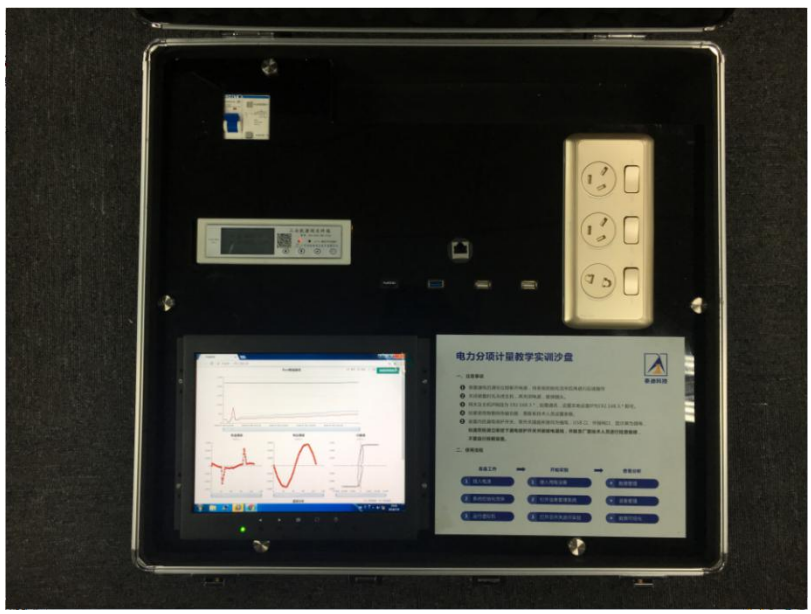


图 5-15 沙盘装置内部

(3) 产品软件

电力分项计量沙盘配备一套后台管理系统，系统采用 B/S 架构，无需安装，可直接通过浏览器登录进行访问。信息管理系统支持以下功能。

支持以视图形式显示实时产生的电力数据的变化趋势，如图所示：

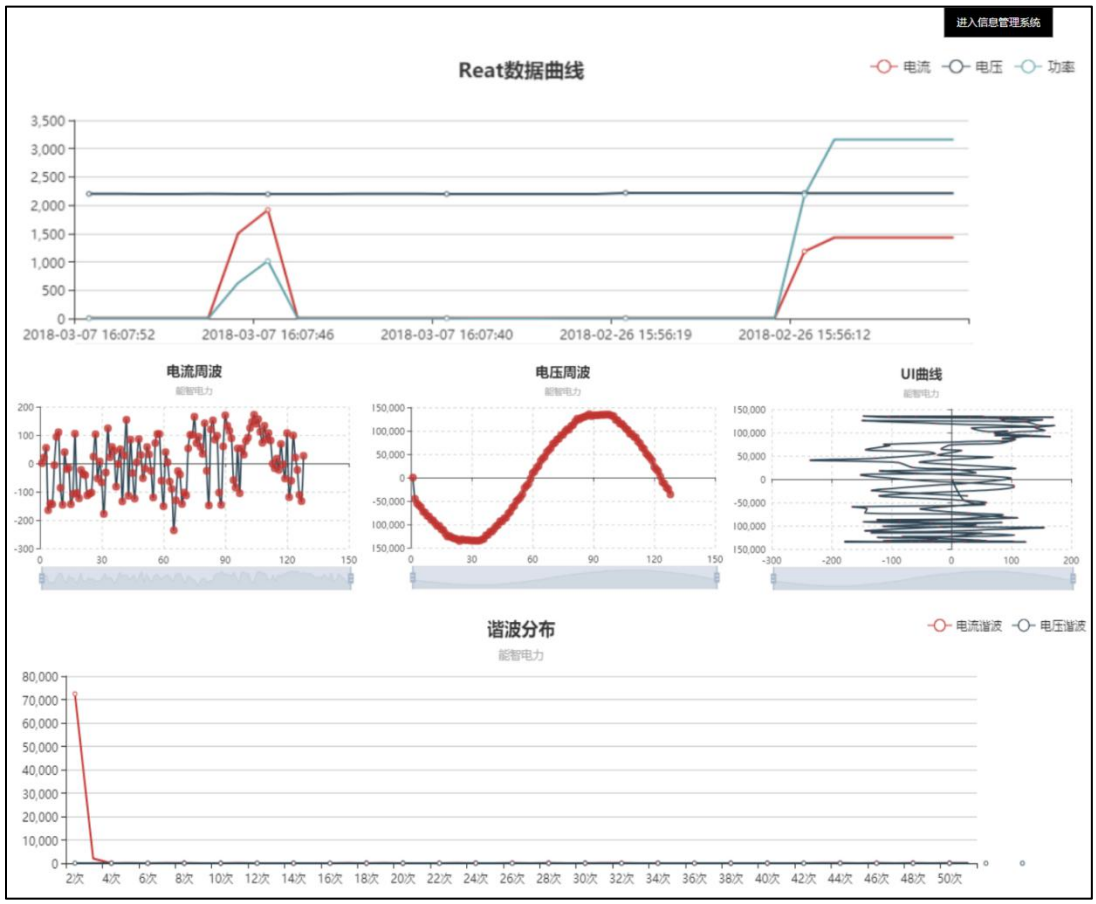
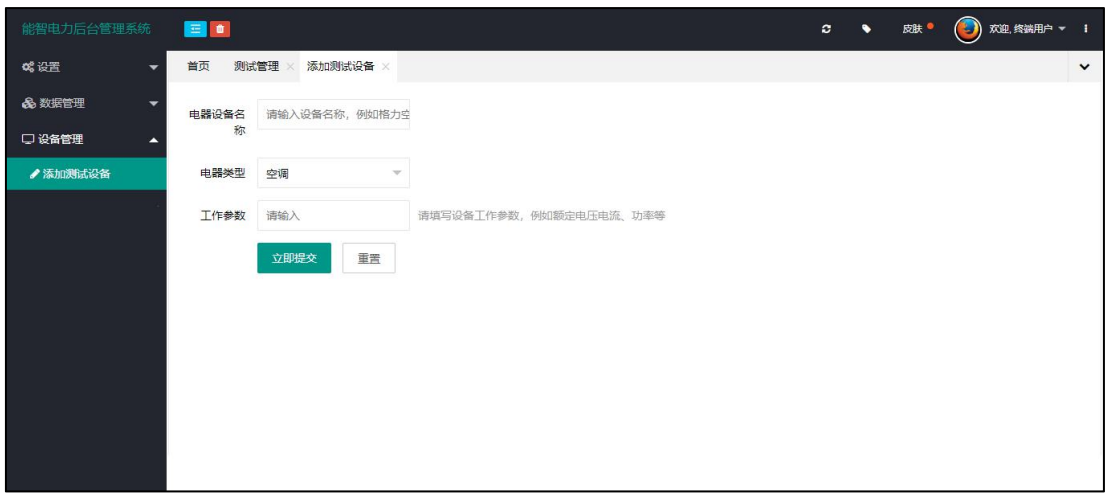
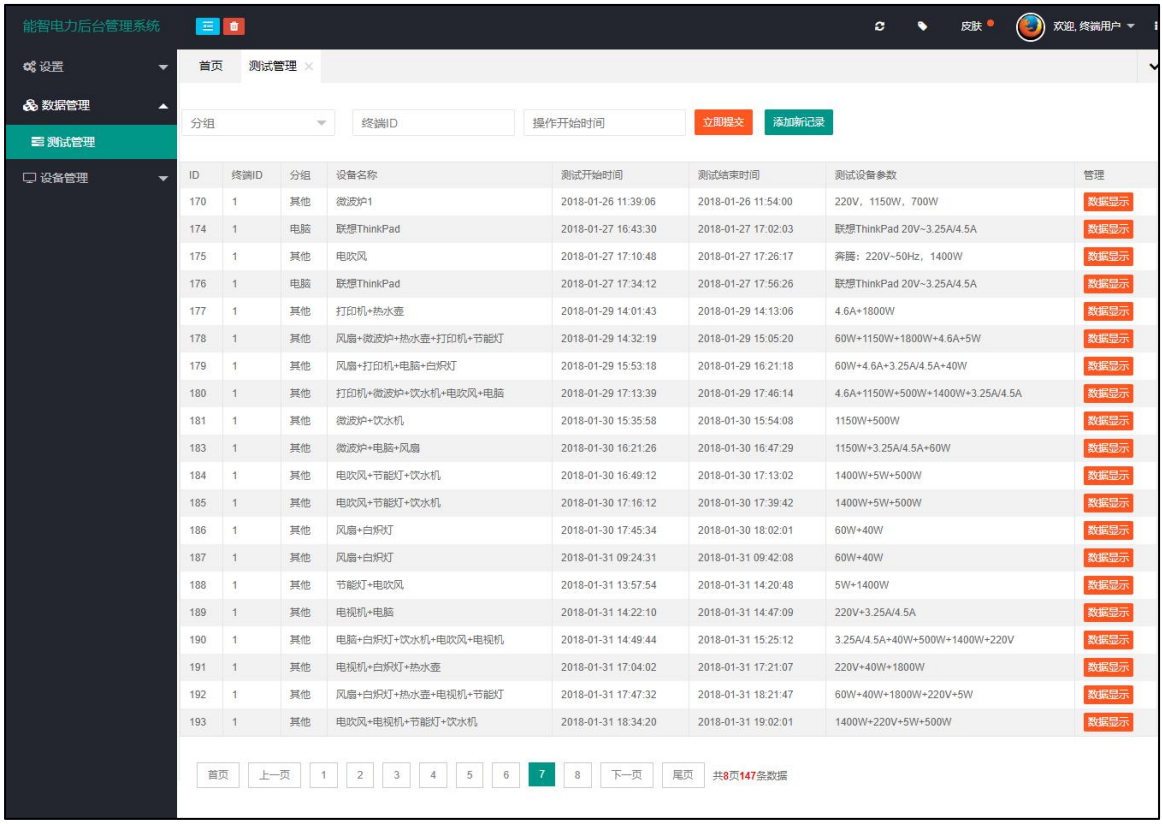


图 5-16 实时产生的电力数据的变化趋势图

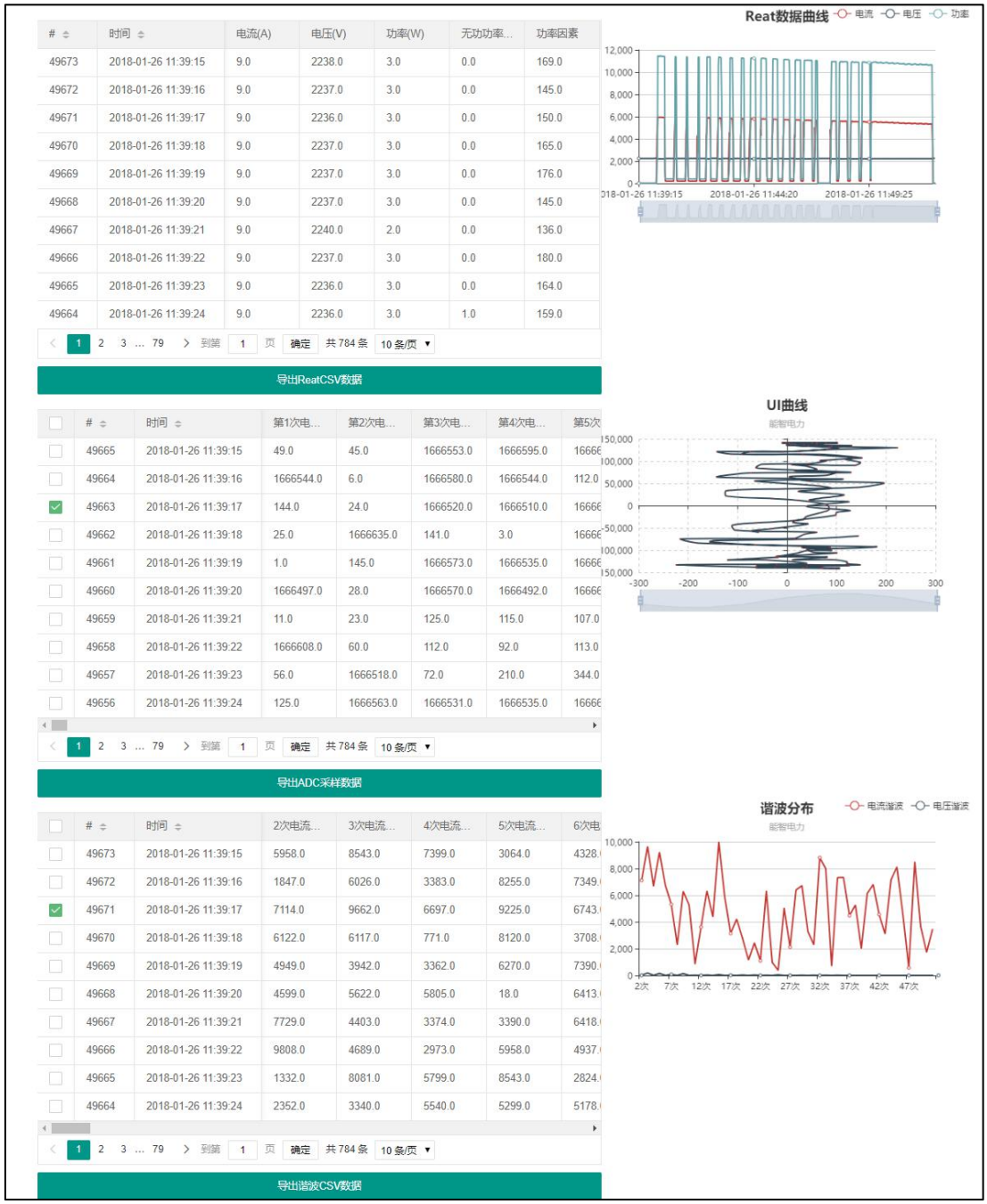
支持对测试设备的信息进行管理，包含添加设备、删除设备、修改设备信息、查找已添加设备。如图所示。



支持查看测试记录数据，以及提供数据可视化功能，可视化部分包括电压、电流、有功功率、无功功率、功率因素的 reat 数据曲线图，周波采样数据的 UI 曲线图，谐波数据的谐波分布图。如图所示：



支持数据导出，可将采集到的数据根据需求导出至本地文件或数据库，数据可使用 Python、R、Matlab 等专业工具进行教学和分析。如图所示：



(4) 配套教学资料（资源）

一、解决方案

解决方案详细地讲解了整个案例的分析过程，如图 3-124 所示，包括案例背景、数据预处理、模型构建等。同时，还设计了上机实验环节和拓展思考，可作为学员的课后作业。

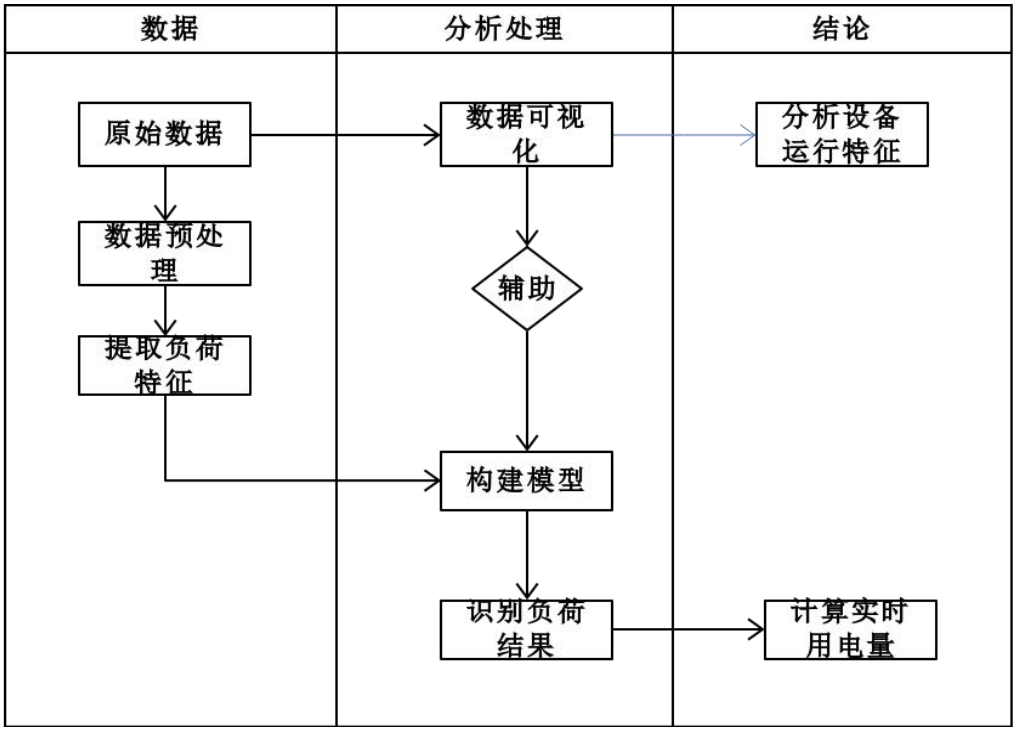


图 5-17 解决方案的总体流程图

二、教学 PPT

PPT 案例课件是以案例教程为依托，以 PPT 形式展现案例的思路，可作为老师讲课的课件。教学 PPT 的部分截图（包括背景、分析过程等）如图所示：

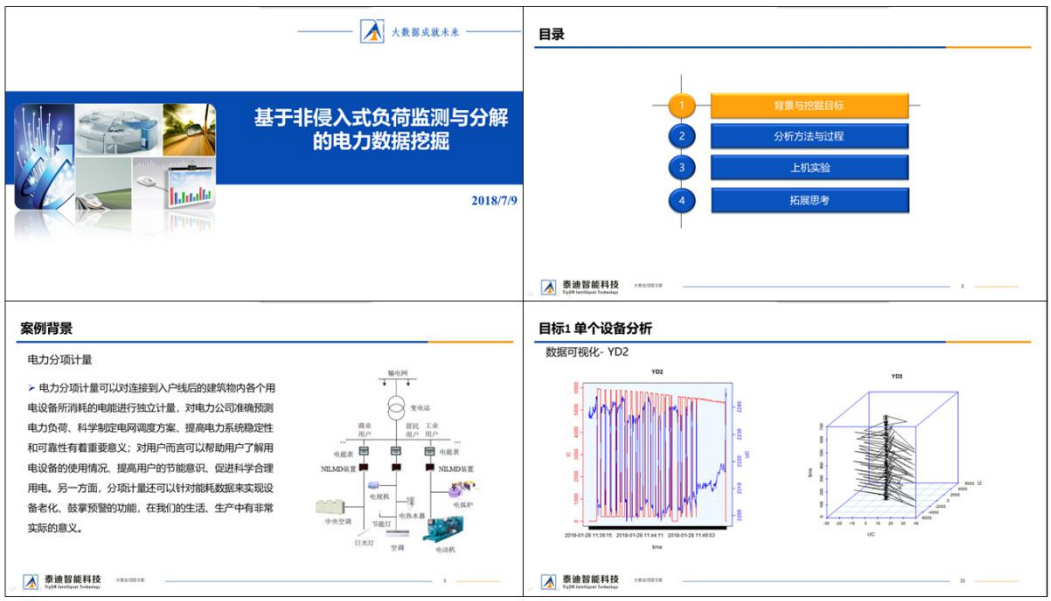


图 5-18 配套教学 PPT

三、数据及代码

案例配备的数据和代码，其中数据是真实的企业数据，有利于进行实际的数据分析体验。数据和代码的部分截图，如图所示：

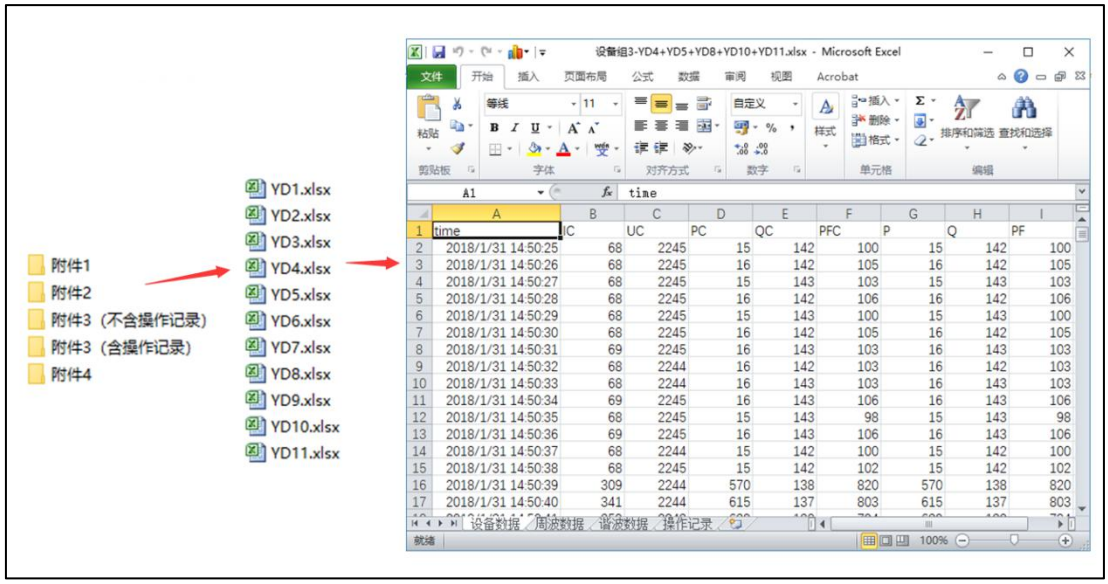


图 5-19 配套数据

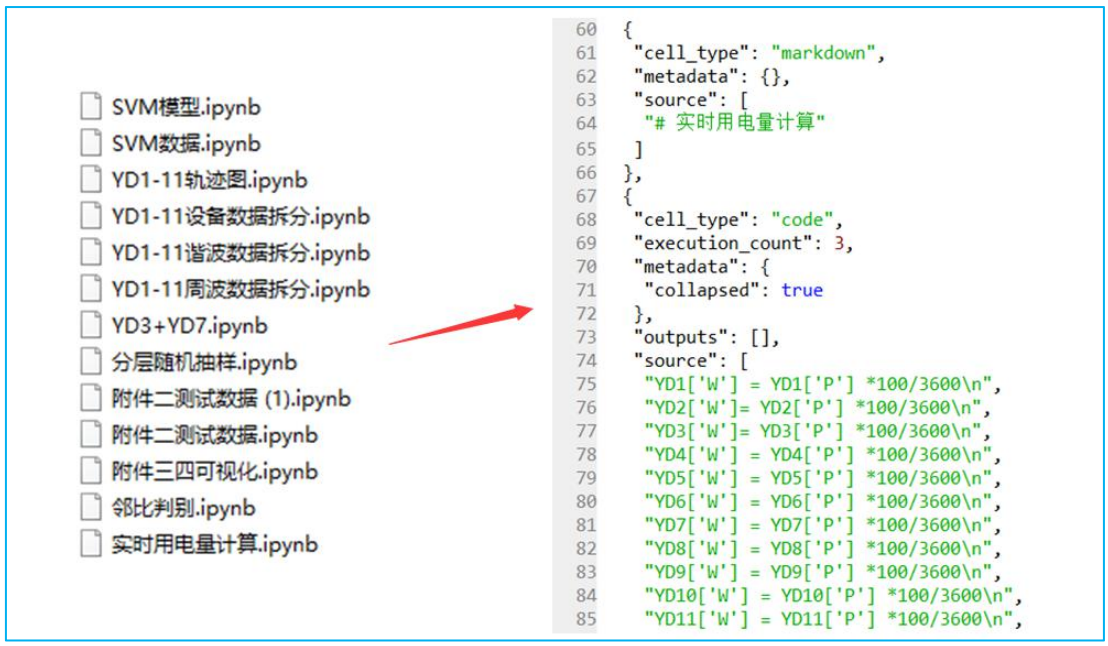


图 5-20 配套代码

6. 师资培训计划

为贯彻落实《国家职业教育改革实施方案》和《关于在院校实施“学历证书+若干职业技能等级证书”制度试点方案》（职教成〔2019〕6 号）、《教育部办公厅 国家发展改革委办公厅 财政部办公厅关于推进 1+X 证书制度试点工作的指导意见》等文件精神，依据《关于确认参与 1+X 证书制度试点的第三批职业教育培训评价组织及职业技能等级证书的通知》（教职所〔2020〕21 号）要求，为保障大数据应用开发（Python）职业技能等级证书高质量运营，实现标准化教学，根据教育部颁发的《教师专业标准（试行）》中关于教师培养、准入、培训、考核等工作的基本要求，提升试点院校教师的实践教学能力，特制定如下教师培训计划：

6.1. 培训对象

开设有“大数据应用开发（Python）1+X 职业技能等级证书”对应专业的院校相关老师、企业在职教师、企业技术人员（院校校外师资）、以及有兴趣了解“大数据应用开发（Python）1+X 职业技能等级证书”其他院校老师。

6.2. 培训目标

面向“大数据应用开发（Python）”职业技能等级证书试点院校的教师，培养一支高技能和优教能力相结合教师队伍，提升院校开展试点工作的整体师资水平，打造能满足教学与培训需求的教学创新团队，使其能独立实施并完成大数据应用开发（Python）职业技能等级证书初级中级课程的教学培训与考核评价工作，促进“大数据应用开发（Python）”职业技能等级证书的教育培训质量全面提升。

6.3. 培训时间

根据区域试点情况及教师人数具体商议培训时间，培训周期一般为 7 天。

6.4. 培训模块及形式

培训主要包括主题报告、核心技能串讲与实践、技能考核等模块。采取集中“项目任务式”培训，具体设有“主题讲座+现场对话”、“知识串讲+技能演练”、“项目实施+考核验收”等形式，培训时长为 5 天，共计 40 学时。

同时积极参与和支持“国培计划”和各地“省培计划”教师培训项目的培训实施工作（具体培训形式及培训要求以各地教育主管部门发布通知为准）

6.5. 培训内容

6.5.1. 大数据应用开发（Python）职业技能培训大纲（初级）

课程模块	课程内容
Python 编程基础	1 认识 Excel 2016 2 输入数据 2.1 输入订单号和菜品名称 2.2 输入价格和数量 2.3 输入日期和时间 3 美化工作表 3.1 合并单元格 3.2 设置边框

	3.3 调整行高与列宽 3.4 设置单元格底纹 58 3.5 突出显示为 1 的单元格 4 获取数据 4.1 获取北京市统计局网站数据 4.2 导入 MySQL 数据源的数据 5 对订单数据进行排序 5.1 根据单个关键字进行排序 5.2 根据自定义店铺所在地顺序进行排序 6 筛选订单数据的关键信息 6.1 根据颜色筛选店铺所在地 6.2 自定义筛选某些会员的消费数据 7 分类汇总每位会员的消费金额 7.1 分类汇总每位会员的消费金额的总额 7.2 分类汇总每位会员的消费金额的平均值 7.3 分页显示汇总结果 8 制作数据透视表 8.1 手动创建订单数据的透视表 8.2 编辑订单数据的透视表 9 使用数学函数处理某企业的营业数据 9.1 使用 PRODUCT 函数计算折后金额 9.2 使用 SUM 函数计算 8 月营业总额 9.3 使用 SUMIF 函数按条件计算 8 月 1 日营业总额 9.4 使用 QUOTIENT 函数计算 8 月平均每日营业额 9.5 将折后金额向下舍入到最接近的整数 10 项目实战分析商品的整体销售情况 10.1 商品销售额的环比分析 10.2 商品毛利率分析 10.3 商品销售量排行榜分析 10.4 商品单价区间的销售量分析
Power BI 数据分析与可视化	1 准备工作 1.1 安装 1.2Power BI 界面介绍 2 获取数据 2.1 获取 excel 数据 2.2 获取 web 数据 2.3 从数据库中获取数据 3.M 语言获取网络分页数据 4 数据预处理 4.1 数据的集成 4.2 数据的清洗 4.3 变换数据 4.4 数据归约 4.5 新建表与计算列 4.6 新建表间关系 4.7 新建度量值

	4.8 上下文操作 4.9 钻取操作 5 数据分析可视化 5.1 条形图-堆积条形图 5.2 条形图-簇状条形图 5.3 条形图-百分比堆积条形图 5.4 柱形图-堆积柱形图 5.5 柱形图-簇状柱形图 5.6 柱形图-百分比堆积柱形图 5.7 漏斗图 5.8 饼图 5.9 环形图 5.10 瀑布图 5.11 树状图 5.12 雷达图 5.13 散点图 5.14 折线图 5.15 箱线图 5.16 表 5.17 子弹图 5.18 仪表 5.19KPI 6 数据分析报表 6.1 分析报告 6.2 发布数据 7 仪表盘
Python 编程基础	1 准备工作 1.1 认识 Python 1.2 搭建 Python 环境 1.3 安装 PyCharm 并创建一个应声虫程序 2Python 基础知识 2.1 掌握 Python 固定语法 2.2 创建字符串变量并提取里面的数值 2.3 计算圆形的各参数 3Python 数据结构 3.1 创建一个列表（list）并进行增删改查操作 3.2 转换一个元组（tuple）并进行取值操作 3.3 创建一个字典（dict）并进行增删改查操作 3.4 将两个列表转换为集合（set）并进行集合运算 4 程序流程控制语句 4.1 实现考试成绩划分 4.2 实现一组数的连加与连乘 4.3 使用冒泡排序法排序 4.4 实训（猜数字游戏） 5 函数 5.1 自定义函数实现输出方差

	5.2 使用匿名函数添加列表元素 5.3 存储并导入函数模块 6 面向对象 6.1 认识面向对象编程 6.2 创建 Car 类 6.3 创建 Car 对象 6.4 迭代 Car 对象 6.5 产生 Land_Rover 对象（子类） 7 文件基础 7.1 认识文件 7.2 读取 txt 文件中的数据 7.3 保存数据为 csv 格式文件 7.4 认识 os 模块
新零售智能销售数据分析	1 数据的清洗 2 归约数据 3 数据建模 4 销售分析及可视化 5 库存分析及可视化 6 用户分析及可视化 7 数据部署
航空公司客户价值分析	1 背景与目标 1.1 案例背景 1.2 案例目标 2 数据预处理 2.1 数据读取 2.2 剔除票价为空的记录 2.3 剔除异常记录 3 特征构造 3.1RFM 模型介绍 3.2LRFMC 模型 3.3 构造入会时长特征 3.4 剩余特征构造 4K-Means 客户分群 4.1 使用 K-means 算法进行客户分群 4.2 获取 K-Means 聚类结果 4.3 聚类结果可视化 5 小结
学生校园消费行为分析	1 背景与目标 2 数据预处理 2.1 数据探索和数据预处理 2.2 数据关联 3 食堂就餐行为分析 3.1 分析可视化各食堂就餐人次的占比饼图 3.2 分析可视化工作日和非工作日食堂就餐时间曲线图 3.3 根据上述分析的结果，为食堂的运营提供建议 4.学生消费行为分析

	4.1 根据学生的整体校园消费数据，计算本月人均刷卡频次和人均消费，分析不同专业间不同性别学生群体的消费特点。 4.2 根据学生的整体校园消费行为，选择合适的特征，构建聚类模型，分析每一类学生群体的消费特点，为学校判定学生的经济状况提供参考意见。
疫情期间湘鄂省区物流数据分析	1 背景与目标 1.1 背景 1.2 数据说明 1.3 分析目标 1.4 分析步骤与流程 2 数据导入和预处理 2.1 导入数据 2.2 数据概况 2.3 清晰数据 2.4 数据插补 2.5 数据合并 2.6 数据规约 2.7 数据转化 2.8 保存数据 3 统计分析与可视化 4 总结
百货商场用户画像描绘与价值分析	1 背景与目标 2 数据探索与预处理 2.1 结合业务对数据进行探索并进行预处理 2.2 将会员信息表和销售流水表关联与合并 3 统计分析 3.1 分析会员的年龄构成、男女比例等基本信息 3.2 分析会员的总订单占比，总消费金额占比等消费情况 3.3 分别以季度和天为单位，分析不同时间段会员的消费时间偏好 4 会员用户画像 4.1 构建会员用户基本特征标签 4.2 构建会员用户业务特征标签 4.3 构建会员用户兴趣特征标签 4.4 建立用户画像 5 会员用户细分和营销方案制定 5.1 对会员用户进行精细划分并分析不同群体带来的价值差异 5.2 针对不同类型的群体制定相应的营销方案

6.5.2. 大数据应用开发（Python）职业技能培训大纲（中级）

课程模块	课程内容
Python 数据分析与应用	1Python 数据分析概述 1.1 认识数据分析 1.2 熟悉 Python 数据分析的工具 1.3 安装 Python3 的 Anaconda 发行版

	1.4 掌握 Jupyter Notebook 常用功能 2NumPy 数值计算基础 2.1 认识 NumPy 数组对象 ndarray 2.2 认识 NumPy 矩阵与通用函数 2.3 利用 NumPy 进行统计分析 3Matplotlib 数据可视化基础 3.1 了解绘图基础语法与常用参数 3.2 分析特征间的关系 3.3 分析特征内部数据分布与分散状况 4Pandas 统计分析基础 4.1 读写不同数据源的数据 4.2 掌握 DataFrame 的常用操作 4.3 转换与处理时间序列数据 4.4 使用分组聚合进行组内计算 4.5 创建透视表与交叉表 5 使用 Pandas 进行数据预处理 5.1 合并数据 5.2 清洗数据 5.3 标准化数据 5.4 转换数据 6 使用 scikit-learn 构建模型 6.1 使用 sklearn 转换器处理数据 6.2 构建并评价聚类模型 6.3 构建并评价分类模型 6.4 构建并评价回归模型
Python 数据可视化	1 准备工作环境 1.1Pyhon 数据可视化概述 1.2Matplotlib 绘图库介绍 1.3Matplotlib 散点图绘制 1.4Matplotlib 绘制折线图 1.5Matplotlib 绘制柱状图 1.6Matplotlib 绘制饼图 2 绘制并定制化图表 2.1Matplotlib 例子的背景介绍 2.2Matplotlib 例子：预处理 2.3Matplotlib 例子：销售额随时间变化的可视化 2.4Matplotlib 例子：星期与销售额关系 2.5Matplotlib 例子：时间销售额与订单量的关系分析 3 学习更多图表和定制化 3.1Pyecharts 简介与绘图逻辑说明 3.2Pyecharts 绘制散点图 3.3Pyecharts 绘制线图 3.4Pyecharts 绘制饼图 3.5Pyecharts 绘制柱状图 3.6Pyecharts 图形组合 4 创建 3D 可视化图表

	4.1Pyecharts 地理图表介绍 4.2Pyecharts 绘制地理散点图 4.3Pyecharts 绘制地理迁徙图 4.4Pyecharts 绘制广东区域图 5 用图像和地图绘制图表 5.1 绘制微信好友性别分布饼图 5.2 绘制微信好友地区分布地理图
Python 数据分析实训	1 探索 Iris 鸢尾花数据 1.1 将数据集存成变量 iris 创建数据框的列名称['sepal_length','sepal_width','petal_length','petal_width','class'] 1.2 数据框中有缺失值吗？ 1.3 将列 petal_length 的第 10 到 19 行设置为缺失值。 1.4 将 petal_lengt 缺失值全部替换为 1.0。 1.5 删除列 class。 1.6 将数据框前三行设置为缺失值。 1.7 删除有缺失值的行。 1.8 重新设置索引。 2 探索 Chipotle 快餐数据 2.1 将数据集存入一个名为 chipo 的数据框内 2.2 查看前 10 行内容 2.3 数据集中有多少个列(columns)？ 2.4 打印出全部的列名称 2.5 数据集的索引是怎样的？ 2.6 被下单数最多商品(item)是什么？ 2.7 在 item_name 这一列中，一共有多少种商品被下单？ 2.8 一共有多少个商品被下单？ 2.9 将 item_price 转换为浮点数 2.10 在该数据集对应的时期内，收入(revenue)是多少？ 2.11 在该数据集对应的时期内，一共有多少订单？ 2.12 每一单(order)对应的平均总价是多少？ 3 探索 Apple 公司股价数据 3.1 读取“appl_1980_2014.csv”数据并存为一个名叫 apple 的数据框。 3.2 查看每一列的数据类型。 3.3 将 Date 这个列转换为 datetime 类型。 3.4 将 Date 设置为索引。 3.5 有重复的日期吗？ 3.6 将 index 设置为升序。 3.7 找到每个月的最后一个交易日(businessday)。 3.8 数据集中最早的日期和最晚的日期相差多少天？ 3.9 在数据中一共有多少个月？ 3.10 按照时间顺序可视化 Adj Close 值。
市财政收入分析预测	1 背景与案例目标 1.1 财政收入预测背景介绍 1.2 数据基本情况介绍

	1.3 分析目标解读 1.4 项目流程介绍 2 相关系数分析 2.1 求解 person 相关系数 2.2person 相关系数解读 3Lasso 回归特征提取 3.1 了解 Lasso 回归方法 3.2Lasso 回归选取关键特征的实现 3.3Lasso 回归数据写出及相应解读 4 灰色预测模型 4.1 关键特征数据读取及准备 4.2GM11 特征值预测 4.3GM11 特征数据整理及写出 5 模型训练及预测 5.1 数据标准化 5.2 模型训练及预测 5.3 结果可视化
Python 网络爬虫技术	1Python 爬虫环境与爬虫简介 1.1Python 网络爬虫实战介 1.2 认识爬虫 1.3 认识反爬虫 1.4Python 爬虫环境 2 网页前端基础 2.1 概述 2.2HTTP 请求方法与过程 2.3 常见 HTTP 状态码 2.4HTTP 头部信息 2.5 认识 cookies 2.6 小结 3 简单静态网页爬取 3.1 静态网页爬取概述 3.2 使用 urllib3 实现 HTTP 请求 3.3 使用 requests 库实现 HTTP 请求 3.4 谷歌开发者工具介绍 3.5 正则表达式介绍 3.6 使用正则表达式获取网页标题信息 3.7 使用 XPath 进行网页解析 3.8 使用 BeautifulSoup 进行网页解析 3.9 数据存储 3.10 小结 4 常规动态网页爬取 4.1 常规动态网页爬取概述 4.2 逆向分析爬取动态网页 4.3 使用 Selenium 打开浏览对象 4.4Selenium 页面等待 4.5 使用 Selenium 获取图书信息

	4.6 小结 5 模拟登录 5.1 模拟登录概述 5.2 查找表单数据入口及提交数据 5.3 验证码人工处理与代理 IP 5.4 使用 POST 请求方法登录 5.5 使用浏览器 cookies 登录 5.6 基于表单登录的 cookies 登录 5.7 小结 6 终端协议分析 6.1 终端协议分析概述 6.2 了解 HTTPAnalyzer 工具 6.3 爬取音乐 PC 客户端数据 6.4 小结
热门电影影评数据爬取及分析	1 案例背景与挖掘目标 2 数据爬取 2.1 短评数据爬取介绍 2.2 安装 selenium 及配置 chromedriver 2.3 获取用户名 2.4 获取短评正文 2.5 设置 cookies 2.6 获取用户居住地和入会时间信息 2.7 单页数据整理 2.8 自定义获取单页数据的函数 2.9 判定网页是否已被加载 2.10 翻页爬取 2.11 代码整理及小结 3 评论数据处理 3.1 短评正文数据预处理 3.2 词频统计 3.3 绘制整体评论数据的词云图 3.4 好评差评词云图绘制及小结 4 评论数据统计及分析 4.1 评分分数分布统计 4.2 短评数量与日期的关系 4.3 短评数量与时刻的关系 4.4 不同评分数量与时间的关系 5 小结 5.1 评论最多的前十个城市 5.2 评分数量与城市的关系 5.3 总结
Hadoop 大数据开发基础	1Hadoop 简介、核心及生态系统 1.1 Hadoop 简介 1.2 Hadoop 核心组件 1.3Hadoop 生态系统 1.4Hadoop 应用场景

	2Hadoop 集群搭建 2.1 安装配置虚拟机 2.2 安装 Java 2.3 搭建 Hadoop 完全分布式集群 3Hadoop 基本操作 3.1 查看 Hadoop 集群的基本信息 3.2 上传文件到 HDFS 3.3 运行首个 MapReduce 3.4 管理多个 MapReduce 任务 4MapReduce 入门编程 4.1 使用 Eclipse 创建 MapReduce 工程 4.2 通过源码初识 MapReduce 编程 4.3 编程实现按日期统计访问次数 4.4 编程实现按访问次数排序 5MapReduce 编程进阶 5.1 筛选日志文件生成序列化文件 5.2Hadoop Java API 读取序列化日志文件 5.3 优化日志文件统计程序 5.4Eclipse 提交日志文件统计程序 6.Hadoop 案例 6.1 基于 KNN 的鸢尾花分类预测 6.2 基于 KMeans 的客户价值分析
网络招聘数据采集与大数据人才需求分析	1 背景与目标 2 数据爬取 2.1 信息爬取介绍 2.2 获取岗位名称数据 2.3 获取目录页的所有字段信息 2.4 获取二级网址的网页链接 2.5 获取二级网址的所有字段信息 2.6 对单一目录页中的所有二级网页信息进行抓取 2.7 将第一个目录页的数据进行保存 2.8 批量爬取及数据保存 3 数据处理 3.1 已爬取数据介绍 3.11 数据预处理小结 3.2 根据岗位名筛选招聘信息 3.3 统一岗位名称 3.4 根据工资列筛选数据 3.5 完成工资数据处理 3.6 工作地点字段处理 3.7 公司类型字段处理 3.8 行业字段数据处理 3.9 工作描述字段处理 3.10 公司规模字段处理 4 数据分析 4.1 热门招聘岗位可视化

	4.2 热门行业及公司招聘分析 4.3 热门岗位的工资水平 4.4 可视化综合分析 4.5 岗位技能分析 5 总结

6.5.3. 大数据应用开发（Python）职业技能培训大纲（高级）

课程模块	课程内容
Python 数据分析与挖掘	1 数据挖掘绪论 2 数据探索与预处理 3 回归分析(Regression Analysis) 3.1 基本形式 3.2 线性模型 3.3 逻辑回归 4 决策树(Decision Tree) 4.1 基本流程 4.2 划分选择 4.3 剪枝 5 人工神经网络(Artificial Neural Network) 5.1 神经元模型 5.2 感知机与多层网络 5.3 误差逆传播 5.4BP 神经网络 6 最近邻算法（KNN） 7 朴素贝叶斯分类(Naive Bayesian) 8 聚类分析(Cluster Analysis) 8.1 聚类任务 8.2 性能度量 8.3 距离计算 8.4 常用聚类算法 9 支持向量机(Support Vector Machine) 9.1 间隔与支持向量 9.2 对偶问题 9.3 核函数 9.4 软间隔与正则化
020 优惠券使用预测	1 背景与目标 2 数据说明 2.1 线下训练集数据介绍 2.2 线上训练集数据介绍 2.3 测试数据介绍 2.4 项目流程介绍 3 数据预处理 3.1 构建正样本 3.2 构建负样本

	3.3 构建样本标签 4 特征构建 4.1 特征构建介绍 4.2 处理 Discount_rate 列 4.3 特征 1-折扣率 4.4 特征 2-商户与用户之间的距离 5 模型训练 5.1 建模前数据准备 5.2 初级模型构建 5.3ROC 曲线与 AUC 值 5.4 模型性能评估 5.5 训练函数封装 5.6 模型预测 5.7 预测函数封装 6 特征完善 6.1 特征 3-优惠券流行度 6.2 特征 4-用户在商家中的消费次数 6.3 如何进行特征拼接 6.4 拼接训练集的特征 3&4 6.5 拼接测试及的特征 3&4 7 预测 7.1 模型训练 7.2 预测 7.3 代码整理 7.4 结果提交
TensorFlow2 实战	1 任务 1：构建一个线性模型 1.1tensorflow 介绍 1.2tensorflow2 常用数据类型和操作 1.3 初始化模型 1.4 构建损失函数 1.5 模型训练及可视化 1.6 使用高阶 API-keras 2 任务 2：mnist 手写数字识别 2.1 数据读取及探索 2.2 交叉熵 2.3 模型构建及训练 2.4 调用保存好的模型对新样本进行预测 3 作业-鸢尾花分类
深度学习原理及编程实现	1.1 深度神经网络-引言 2 卷积神经网络 CNN 2.1 浅层神经网络的局限 2.2 卷积操作 2.3 卷积操作的优势 2.4 池化及全连接 2.5 高维输入及多 filter 卷积 2.6 实现卷积操作

	2.7 实现池化操作 3 循环神经网络 RNN 3.1 循环神经网络简介 3.2 循环神经网络的常见结构 4 长短时记忆网络 LSTM 4.1LSTM 的三个门 4.2LSTM 三个门的计算示例 4.3 利用 RNN&LSTM 实现 mnist 手写数字识别
文本挖掘	1 自然语言处理简介 2 开源中文 NLP 系统介绍 3 中文分词介绍 4 机械分词法 5 机器学习算法分词 6NLP 概率图介绍 7jieba 分词演示 8 文本的 one-hot 表达 9tf-idf 权值策略实现 10 文本的 TF-IDF 表达 11 模型训练与预测
垃圾短信智能识别	1 背景与目标 2 数据探索 2.1 数据读取 2.2 数据抽取 3 数据预处理 3.1 去除短信中的 x 序列 3.2 结巴分词 3.3 去除停用词 3.4 数据预处理函数封装 3.5 垃圾短信的词频统计 3.6 词云图绘制 4 文本向量的表示 4.1 文本数据的向量化表达 4.2 获取训练样本的 tf-idf 权值向量 4.3 获取测试样本的 tf-idf 权值向量 5 模型训练及评价 6 小结
利用循环神经网络（RNN）对路透社新闻进行分类	1.项目背景与目标 2.数据探索分析 2.1 读取新闻数据 2.2 了解数据的基本情况 3.词嵌入（Word Embedding） 3.1word embedding 的基本概念 3.2word2vec 介绍 3.3CBOW 词向量训练过程 3.4《鹿鼎记》word2vec 介绍 3.5《鹿鼎记》word2vec 实现-读取数据

	3.6《鹿鼎记》word2vec 实现-数据预处理 3.7《鹿鼎记》word2vec 实现-词向量训练 3.8《鹿鼎记》word2vec 实现-获取词向量矩阵 4.构建模型 4.1 数据 padding 4.2 网络结构中的 Embedding 层 4.3 构建 RNN 网络模型 4.4 模型训练及评估 5.模型优化 5.1 词向量预训练 5.2 模型优化
动态人脸智能识别	1 案例背景及介绍 2 人脸识别流程及实现 2.1 人脸识别案例流程 2.2 工程文件说明 2.3 人脸采集 2.4 人脸检测 2.5 灰度处理 2.6 模型结构与训练 2.7 模型测试 2.8 模型应用：调用电脑摄像头采集数据 2.9 模型应用：人脸检测 2.10 模型应用：模型测试与展示 3 人脸识别拓展思考
基于深度学习的推荐系统受众性别预测	1 明确项目背景及目标 2 数据获取与探索分析 2.1 读取数据并查看数据规模 2.1 缺失值探索分析。 3 获取用户相应单击流数据 3.1 理解用户单击流相关概念 3.2 获取用户的各单击流数据 4 对各单击流数据进行探索 4.1 查看单击流的长度分布，并进行可视化 5 实现词嵌入（Word Embedding）操作 5.1 对用户单击流数据进行预处理 5.2 进行词向量训练（如获取素材 id 所有词的词向量矩阵） 5.3 对用户的单击流进行编码及 padding 操作 5.4 将词向量矩阵做相应排序并储存 5.5 将单击流数据转化为二维样本数据并储存 6 构建循环神经网络（RNN）序列模型 6.1 搭建网络 6.2 模型训练并储存 6.3 加载模型并预测 7 模型优化

6.6. 师资来源

培训师资团队由职教专家、行业（企业）一线专家和技术工程师、院校一线优秀教师共同组成。

（1） 行业（企业）师资主要来源：广东泰迪科技、广州智能装备研究院、网宿科技、广州思迈特、艾普工华科技、深圳市怡亚通等。

（2） 院校优秀教师主要来源：开设有大数据专业的院校中的一线知名骨干教师。

6.7. 培训证书

培训结束后安排考核，通过考核后，将颁发结业证书，并取得相应继续教育学时和学分。

6.8. 线上资源

为了充分满足院校教师随时随地学习以及知识补充的需要，还会开发专门的线上学习资源免费提供给院校教师，并分批次召开线上师资培训直播课。通过线上线下相结合，有效提升教师的教学能力和知识广度。

7. 考点申报

7.1. 考点申报条件

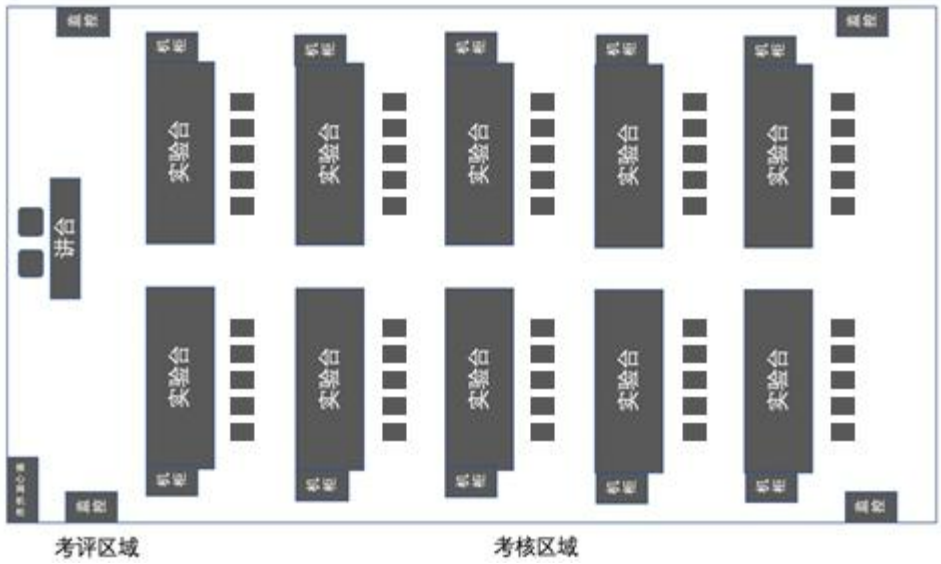
1. 申报条件：首先成为试点学校才能申请考核站点.
2. 满足基本条件：交通、安全、能实现封闭式管理
3. 考试用设施设备条件：课桌椅、考试机、巡查设备、网络设备等
4. 组织考试条件：考站负责人、监考人员、技术支持人员、视频监考员等.

7.2. 考点建设标准

为了顺利开展证书的考核工作，申请考核站点（以下简称考点）的单位需要满足大数据应用开发（Python）职业技能等级证书考点建设标准。 具体如下：

7.2.1. 考核场地

考核场地包括三个区域：考核区域、考评区域以及后台监控中心区域，示意图如下所示：



考核场地示意图

1. 考核区域

考核场地为参加考核人员的考试区域，安排有 60 个座位，可容纳 60 人进行考试。

2. 后台监控中心区域

后台监控中心区有监控设备，实时监控考核过程，可防止考核人员作弊等情况发生，并保障考核的公平性、公正性和公开性。

7.2.2. 考核设备

1，理论机考硬件环境

硬件需求	考试中心	考试终端
CPU	第四代智能英特尔酷睿 i5处理器及以上	双核及以上（2.0GHz）
内存	8.0GB及以上	2.0GB及以上
硬盘	10.0GB以上	10.0GB以上
显卡	集成显卡及以上	集成显卡及以上
分辨率	≥1366*768	≥1366*768
系统	Windows XP/Windows 7/Windows 8/Windows 8.1/Windows 10	

2，实操考核环境

名称	硬件配置	单位	数量
应用平台服务器	规格: 2U 机架式服务器 CPU：Intel Xeon 系列；双路CPU；主频2.0GHz；每颗12核心； 内存：160GB DDR4； 硬盘：2*240GB(SSD 企业级)；4*2TB（SATA 7.2K） 阵列卡：RAID控制器，缓存1GB，支持RAID 0/1/5/6/10/50；	台	2
AI计算服务器	4U 机架式GPU服务器 CPU：Intel Xeon 系列；双路CPU；主频2.2GHz；每颗12核心； 内存：DDR4 256GB（32GB*8） 硬盘：3*480GB（SSD 企业级）、2*240GB（SSD 企业级）、3*4TB(SATA 7.2K) GPU: Nvidia GPU RTX 2080Ti*4；显存11GB；每颗含4352个CUDA核心 网卡：万兆双口，含光模块 其它：冗余电源1100W，企业版IPMI	台	3
服务器机柜	42U标准机柜 19英寸加厚2米机柜	台	1

	机柜采用优质冷轧钢材质，脱脂喷塑工艺面板		
机架式KVM切换器	8合1视频接口，17英寸LCD屏 接口：USB口*2、PS2口*2、VGA口*8 金属结构、静电喷漆、抗磨防腐	台	1
管理交换机	万兆交换机 端口：24个10G SFP+端口，2个40G QSFP+ 规格：19英寸（标准机架） 速率：10Gbps/40Gbps 交换容量：1.28Tbps/12.8Tbps 包转发率：480Mpps 网络模块：SFP/SFP+万兆多模4个、光电转换模块2个	台	1
应用交换机	千兆交换机 端口：48个10/100/1000Base-T以太网端口，4个100/1000 Base-X SFP光口 规格：19英寸（标准机架） 交换容量：336Gbps 包转发率：78Mpps	台	2

7.2.3. 考核人员

1. 机考部分：每个考场配备两名监考人员，根据实际情况配备一定数量巡考人员。
2. 实操部分：每个考场配备两名评考员，根据实际情况配备一定数量巡考人员。

7.2.4. 考核站点保密管理制度建设

1．各方对其提供的知识产品拥有绝对的知识产权，各方均承诺尊重与保护对方的知识产权。

2．不得擅自解密、泄漏、传播题库内容给任何第三方。

7.2.5. 安全规范

1．考核站点应建立与公安、消防、司法行政、交通、卫生、食品、质检等相关部门的协调机制，保证安全，制定应急预案，及时处置突发事件。

2．考核站点应符合消防安全要求，需在显眼地方清晰标注安全通道位置，且保证安全通道无障碍通行。

3．考核站点的网线、电源线以及其他线路应符合安全布线要求。

4．考核站点需配备应急医生一名，以及一些常用急救药品。应急工具例如应急灯具等齐备并保障可以使用。

5．如果出现安全问题，在安保人员指挥下，迅速按紧急疏散路线撤离现场。