数据清洗专业（中专）
人才培养方案

南京云创大数据科技股份有限公司
Nanjing Innovative Data Technologies,Inc.

2019 年 03 月

数据清洗专业（中专）

人才培养方案

一、招生对象及学制

（一）招生对象：全日制普通中学初中毕业生，招生方式为普通中考招生。

（二）学习年限：基本学制三年，实行弹性学制,学生在校时间原则上不能少于两年，总在校时间（含休学）不得超过六年。

二、培养目标

培养学生德、智、体、美全面发展，能够较快适应生产、建设、管理、服务等一线岗位需要的，面向交通、金融、医疗、安防等各行业的数据清洗、数据处理和大数据技术应用相关工作岗位。学生应掌握数据清洗与大数据技术应用理论基础知识，掌握数据清洗流程、数据清洗工具、数据转换与加载等基本技能，可熟练地对数据审查过程中发现的错误值、缺失值、异常值、可疑数据，选用适当方法进行“清理”，使“脏”数据变为“干净”数据。同时，需具备较高综合素质与良好职业素养，成为能够从事相关工作的发展型、复合型、创新型技术技能人才。

三、人才培养规格

大数据由大量结构化、半结构化和非结构化数据组成，因为来自多样化数据源，造成数据内容并不完美，存在着许多“脏数据”，即数据不完整有缺失、存在错误和重复的数据、数据的不一致和冲突等缺陷，需要经过数据采集、清洗、

存储、分析、建模、可视化等过程加工处理之后，才真正产生价值。所以，数据清洗工作是大数据处理过程中不可缺少的必要环节，该岗位能在数据中发现不准确、不完整或不合理数据，并对这些数据进行修补或移除，以提高数据整体质量，具有非常好的市场就业前景。因此，本专业毕业生在知识、能力和素质等方面应达到如下具体要求：

（一）基础素质

思想政治素质：毕业生身心健康，有良好的道德修养，尊重生命、遵纪守法、诚信友善、乐于奉献；有高尚的民族精神，积极弘扬传统文化，热爱祖国，崇尚集体主义精神；有坚定的理想信念，拥护中国特色社会主义，贯彻科学发展观、和谐社会理论和“四个全面”思想。

职业素质：具有良好的职业态度和职业道德修养，具有正确的择业观和创业观，具有职业操守、爱岗敬业、吃苦耐劳、诚实守信、办事公道、服务群众、奉献社会等美德，具有从事职业活动所必需的基本能力和管理素质，脚踏实地、严谨求实、勇于创新。

身心素质：具有一定的体育运动和生理卫生知识，养成良好的锻炼身体、讲究卫生的习惯，掌握一定的运动技能，达到国家规定的体育健康标准，具有坚韧不拔的毅力、积极乐观的态度、良好的人际关系及健全的人格品质。

（二）职业通用能力

1） 具有一定的英语阅读和写作能力。

2） 具有基本的软件程序编程能力。

3） 具有图像处理的基本能力和 Office 办公软件操作能力。

4） 通过对大数据技术应用理论基础知识的学习和实践，能够具备一定的大数据技术应用能力。

（三）职业岗位能力

数据清洗就是对原始数据进行重新审查和校验的过程，因此数据清洗工程师必须掌握如下技能：首先，掌握数据清洗的基本流程，包括分析数据并定义清洗规则、搜寻并标识错误实例、纠正发现的错误、干净数据回流和数据清洗的评判等；其次，熟练使用数据清洗工具，包括 Microsoft Excel 数据清洗、Kettle 软件、OpenRefine 软件、DataWrangler 软件和 Hawk 软件等开源 ETL 工具；再次，

掌握数据抽取、转换和装载的基本技术方法，熟练全量与增量抽取、数据清洗、数据校验、数据转换、数据审计、数据加载等环节的基本应用；最后，掌握大数据理论基础知识，了解大数据生态系统的技术框架，熟悉 Hadoop 的核心和扩展组件，具备大数据平台实践能力和程序设计能力，具备数据库基础理论知识和实际操作能力。

四、毕业标准

本专业学生在毕业审查时，要求同时达到以下条件：

（一）具有良好的思想道德和身体素质。

（二）各科成绩合格，取得的总学分达到 113 学分及以上。

（三）毕业设计答辩合格。

（四）取得下列人才认证证书之一：

1） 工业和信息化部颁发的《工业和信息化领域急需紧缺人才培养工程证书》。

2） 软件专业技术水平（资格）考试程序员或软件工程师证书。

3） 全国计算机等级二级合格证书。

五、课程设置和学分要求

课程类别与学分结构表

课程模块类别	课程学分	课程学时	占总学分比例(%)
公共基础课程	49	816+2 周	43.35
专业基础课程	23	368	20.35
专业核心课程	22	352	19.47
毕业设计与就业指导	19	48+16 周	16.83
总计	113	1584+18 周	100%

公共基础课程：包括语文、数学、公共英语、哲学与人生、政治经济与社会、职业道德与法律、安全教育、心理健康教育、公共体育等。

专业基础课程：包括计算机基本操作、计算机网络、SQL 数据库、Offcie 办公系统、网页制作、图像处理、Linux 系统与应用、C 语言程序设计等。

专业核心课程：包括 Python 语言、R 语言、云计算导论、大数据导论、大数据实践、数据清洗、数据挖掘基础等。

六、教学进度规划（含主要实践性教学环节、专业实验）

第一学年：上半学期			
课程名称	**学分**	**学时(周)**	**其中实验学时**
军训	2	2 周	2 周
语文（一）	4	64	
数学（一）	4	64	
公共英语（一）	4	64	
政治经济与社会	3	48	
安全教育	1	16	
公共体育（一）	1	32	
计算机基本操作	2	32	16
Offcie办公系统	3	48	24
说明：共计 24学分。			

第一学年：下半学期			
课程名称	**学分**	**学时(周)**	**其中实验学时**
语文（二）	4	64	
数学（二）	4	64	
公共英语（二）	4	64	
职业道德与法律	3	48	
公共体育（二）	1	32	
计算机网络	2	32	16
SQL 数据库	3	48	24

C语言程序设计	4	64	32
说明：共计 25学分。			

第二学年：上半学期			
课程名称	**学分**	**学时(周)**	**其中实验学时**
数学（三）	4	64	
公共英语（三）	4	64	
哲学与人生	2	32	
公共体育（三）	1	32	
Linux系统与应用	3	48	24
网页制作	3	48	24
Python语言（专业核心课程）	4	64	32
大数据导论（专业核心课程）	3	48	16
云计算导论（专业核心课程）	3	48	16
说明：共计 27学分。			

第二学年：下半学期			
课程名称	**学分**	**学时(周)**	**其中实验学时**
公共体育（四）	1	32	
心理健康教育	2	32	
图像处理	3	48	24
大数据实践（专业核心课程）	3	48	32
数据清洗（专业核心课程）	3	48	32
数据挖掘基础	3	48	32

（专业核心课程）			
R语言 （专业核心课程）	3	48	24
说明：共计 18学分。			

第三学年			
课程名称	学分	学时(周)	其中实验学时
就业指导与职业规划	2	32	
职业礼仪	1	16	
毕业设计	12	16周	16周
顶岗实习	4	-	-
说明：共计 19学分。			

七、专业核心课程教材推荐

数据标注作为新兴产业，如何实现教学与行业需求相吻合具有重要意义，南京云创大数据科技股份有限公司作为深耕云计算、大数据、人工智能行业多年的企业，在大数据、人工智能方面拥有丰富的实际项目经验和独到的行业见解。同时，经过与各高校多年的深入合作，南京云创大数据科技股份有限公司也更清楚高校育人与企业用人如何有效对接，故对相关专业课程的教材进行推荐，教材的绝大部分内容也是源于企业实际项目，更具有实践意义。

推荐的每本教材皆有配套的 PPT、视频、操作手册、源代码及原始数据，教师教学能更加轻松顺畅，学生也更能体会到实际企业项目的过程，提升教学质量。

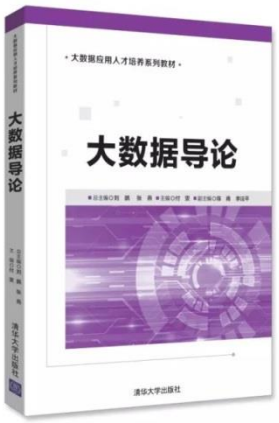
（一）专业核心课程教材推荐表

课程名称	学时数	推荐教材
云计算导论	48	《云计算导论》由刘鹏教授作为丛书总主编率领团队编写，清华大学出版社出版。
大数据导论	48	《大数据导论》由刘鹏教授作为丛书总主编率

		领团队编写，清华大学出版社出版。
数据清洗	48	《数据清洗》由刘鹏教授作为丛书总主编率领团队编写，清华大学出版社出版。
数据挖掘基础	48	《数据挖掘基础》由刘鹏教授作为丛书总主编率领团队编写，清华大学出版社出版。
大数据实践	48	《大数据实践》由刘鹏教授作为丛书总主编率领团队编写，清华大学出版社出版。
Python语言	64	《Python 语言》由刘鹏教授作为丛书总主编率领团队编写，清华大学出版社出版。
R语言	48	《R 语言》由刘鹏教授作为丛书总主编率领团队编写，清华大学出版社出版。

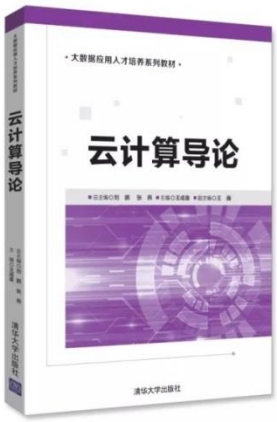
（二）推荐教材内容介绍

1. 《大数据导论》



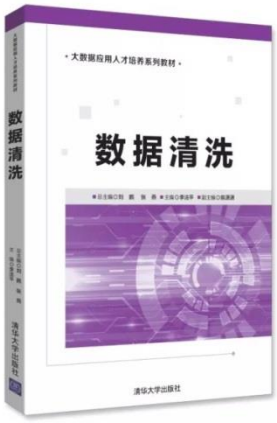
《大数据导论》是了解和学习大数据的基础条件，通过本书了解大数据基本概念，大数据的架构，大数据的采集方式和预处理，常用的 ETL 工具，简单熟悉数据仓库的构建模式，大数据的存储，数据挖掘的方法，以及大数据的可视化技术，从而更好的将大数据技术应用在各行业领域，更深入地开展大数据技术的应用研究。从基础开始，通过理论与实际案例相结合，帮助读者由浅入深进行学习，逐步清理大数据的核心技术和发展趋势。本书可以作为培养应用型人才的课程教材，也适用于初学入门者，对大数据基础理论有需求的广大读者。

2. 《云计算导论》



本书主要内容包括云计算的基本概念、发展现状、主要平台的部署及关键技术、虚拟化与容器技术、云计算的实用化、国内外云计算服务与大规模应用、环境云(envicloud.cn)和万物云(wanwuyun.com)典型行业应用介绍与剖析等内容，适用于应用型本科、高职高专院校的云计算课程和教学。本书的实验环境部署通过云创大数据实验平台(https://bd.cstor.cn)上远程开展。

3. **《数据清洗》**



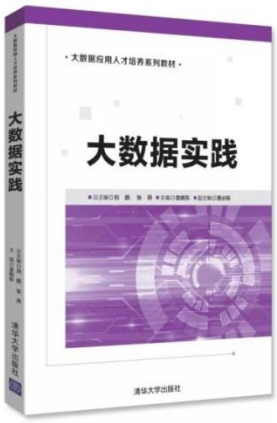
本书系统地讲解了数据清洗理论和实际应用，共分为 8 章：第 1 章主要介绍数据清洗的概念、任务和流程，数据标准化概念及数据仓库技术等；第 2 章主要介绍 Windows 和类 UNIX 操作系统下的数据常规格式、数据编码及数据类型转换等；第 3 章介绍 ETL 概念、数据清洗的技术路线、ETL 工具及 ETL 子系统等；第 4 章介绍了 Excel、Kettle、OpenRefine、DataWrangler 和 Hawk 的安装及使用等；第 5 章介绍 Kettle 下文本文件抽取、Web 数据抽取、数据库数据抽取及增量数据抽取等；第 6 章介绍数据清洗步骤、数据检验和数据错误处理，数据质量评估及数据加载；第 7 章介绍网页结构，利用网络爬虫技术进行数据采集，利用 JavaScript 技术进行行为日志数据采集等；第 8 章介绍 RDBMS 的数据清洗方法

和数据脱敏处理技术等。

4. 《数据挖掘基础》

本书介绍了数据挖掘的基本概念，包括数据挖掘的常用算法、常用工具、用途和应用场景及应用状况，讲述了常用数据挖掘方法，如分类、聚类、关联规则的概念、思想、典型算法、应用场景等。此外，本书还从实际应用从发，讲解了基于日志的大数据挖掘技术的原理、工具、应用场景和成功案例。通过以上内容的学习，读者将了解数据挖掘的基本概念、思想和算法，并掌握其应用要领。本书可以作为培养应用型人才的课程教材，也可作为相关开发人员的自学教材和参考手册。

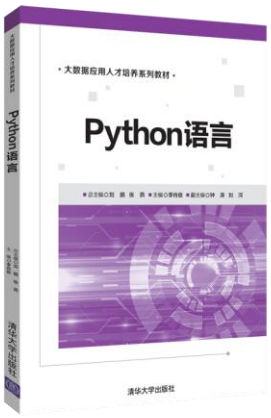
5. 《大数据实践》



本书内容涵盖了目前使用广泛的大数据处理系统 Hadoop 生态圈中的几大核心软件系统：分布式大数据处理系统 Hadoop、Hadoop 数据库 HBase、数据仓库工具 Hive、内存大数据计算框架 Spark 和 Spark SQL，详细介绍了它们的架构、工作原理、部署方法、常用配置、常用操作命令、SQL 引擎等内容。本书对上述几大系统的各种安装部署方式都给出了详细步骤，常用命令也都有具体示例介绍，

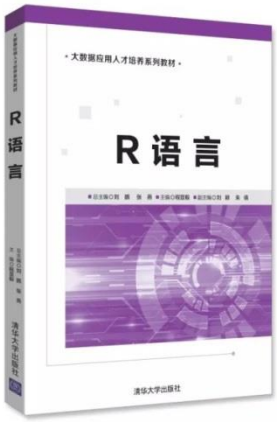
是一本实操性很强的工具书，能帮助初学者快速掌握和操作这几款常用的大数据处理系统。本书以浅显易懂的语言风格和图文并茂的操作示例引领读者迈入大数据实践之门，可以作为培养应用型人才的课程教材，也可作为相关开发人员的自学教材和参考手册。

6. 《Python 语言》



本书以 WIN10 和 Python3.6.5 搭建 Python 开发基础平台为起点，重点阐述 Python 语言的基础知识和三个典型的项目实战案例。全书以理论引导、案例驱动、上机实战为理念打造 Python 语言学习的新模式。具体内容分为两大部分：第一部分以 Python 编程语言基础知识普及为主，分别介绍了 Python3 概述、基本语法、流程控制、Python 组合数据类型、字符串与正则式、函数、模块、类和对象、异常处理、文件操作；第二部分：以项目实战为核心，以学以致用为导向，以切近生活的案例为依托，分别介绍 Python 爬虫项目实战、Python 数据可视化项目实战、Python 数据分析项目实战。

7. 《R 语言》



近年来，R 语言可谓是数据分析的热门语言，相关的资料五花八门，让读者

无所适从，本书力求用简洁、精炼、理论实践相结合的方式让大家快速掌握 R 语言。全书共 14 章，分为基础篇(第 2-10 章)，应用篇(第 11-12 章)和进阶篇(第 13-14 章)。基础篇按照数据分析过程，主要讨论了 R 数据结构、数据导入/导出、数据清洗、数据变换、可视化、高级语言编程和常用建模方法。应用篇通过对 5 个经典案例的分析，使读者能够把学到的 R 基础知识应用到解决实际问题，把数据变成价值。进阶篇解决如何用 R 处理大数据的一些技术。本书可以作为培养应用型人才的课程教材，也可作为数据分析爱好者的参考资料。

八、培训老师配备与要求

（一）计算机相关专业大学本科及以上学历，3 年或以上工作经验。

（二）3 年以上的大数据行业工作经验，并多次参与大数据项目管理或实施，具备丰富的项目经验。

（三）接受过大数据、软件开发、数据库等方面专业级培训，获得过相关的技能认证证书。

（四）从事过大数据相关系统开发的优先，有过大数据培训课程授课经验的优先。

九、成立专业教学指导委员会

专业教学指导委员会的成员主要来自学校领导、授课老师和行业技术专家。成立专业教学指导委员会的主要工作任务包括：

1） 组织和开展本专业教学领域的理论与实践研究。

2） 指导本学科专业建设、教材建设、教学改革、实训基地建设、实验室建设等工作。

3） 制定专业教学规范。

4） 制定教学质量标准。

5） 组织师资培训、教学研讨和信息交流等工作。

联系方式：

地址：南京市白下高新技术产业园中国云计算创新基地 A 栋 9 层
电话：400-8855-360　　传真：025-83708922
官方网站：http://www.cstor.cn　　微信公众号：cStor_cn