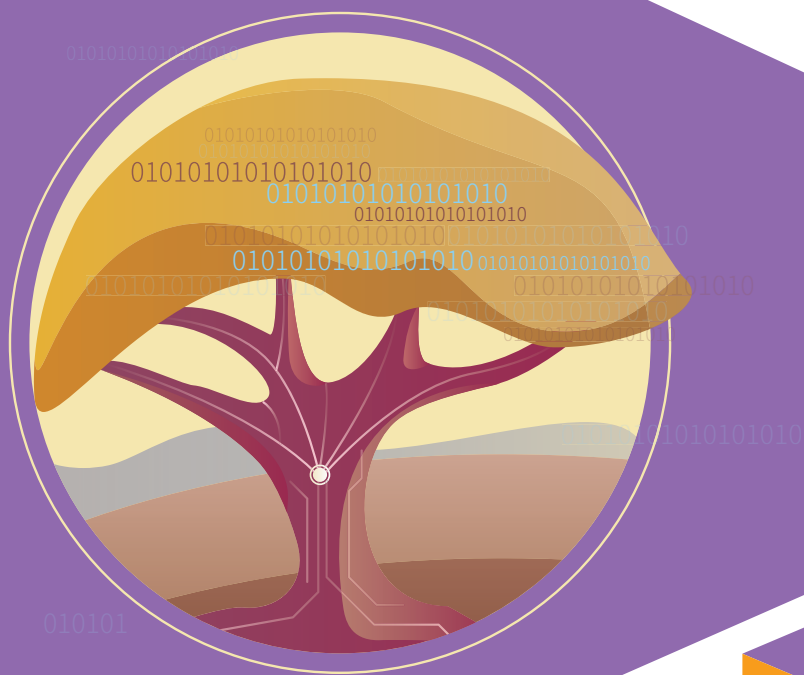




大数据 专业建设解决方案

MAJOR CONSTRUCTION
OF BIG DATA THE SOLUTION



Contents 目录

新大陆大数据产业背景	P01
大数据专业建设整体解决方案概览	P03
大数据认知与应用实训室	P04
1+X大数据工程化处理与应用职业技能等级证书	P07
新大陆“岗位就业导向”课程体系构建思路	P09
专业教学实训室	P11
大数据工程技术实训室	P16
实训室效果图	P20
合作案例	P20

新大陆大数据产业背景

立足通信行业近30年，全网市场布局广泛，全国软件行业收入50强



全网市场广泛布局

移动**26个**省

联通电信**14个**省

非运营商

全国软件行业收入50强

连续17次获得中国软件收入**50强**

服务全国**9亿**用户

企业连续3年信用等级评定为**AAA级**

拥有软件开发和系统集成等多项专业资质

产业标准

通过中国信通院“数据治理服务商能力成熟度”四级评估（全国仅三家）。

3项能力成功入选《数据治理产业图谱1.0》（DataAnts、DataEleph、数据治理服务商能力成熟度）



专业资质

CMMI L5

信息安全服务资质二级

涉及国家秘密的计算机信息系统集成甲级资质

CNAS实验室

高新技术企业

3A信用等级

ISO 9001、20000和27001体系

国家规划布局内重点企业

安全等保测评等



专利软著

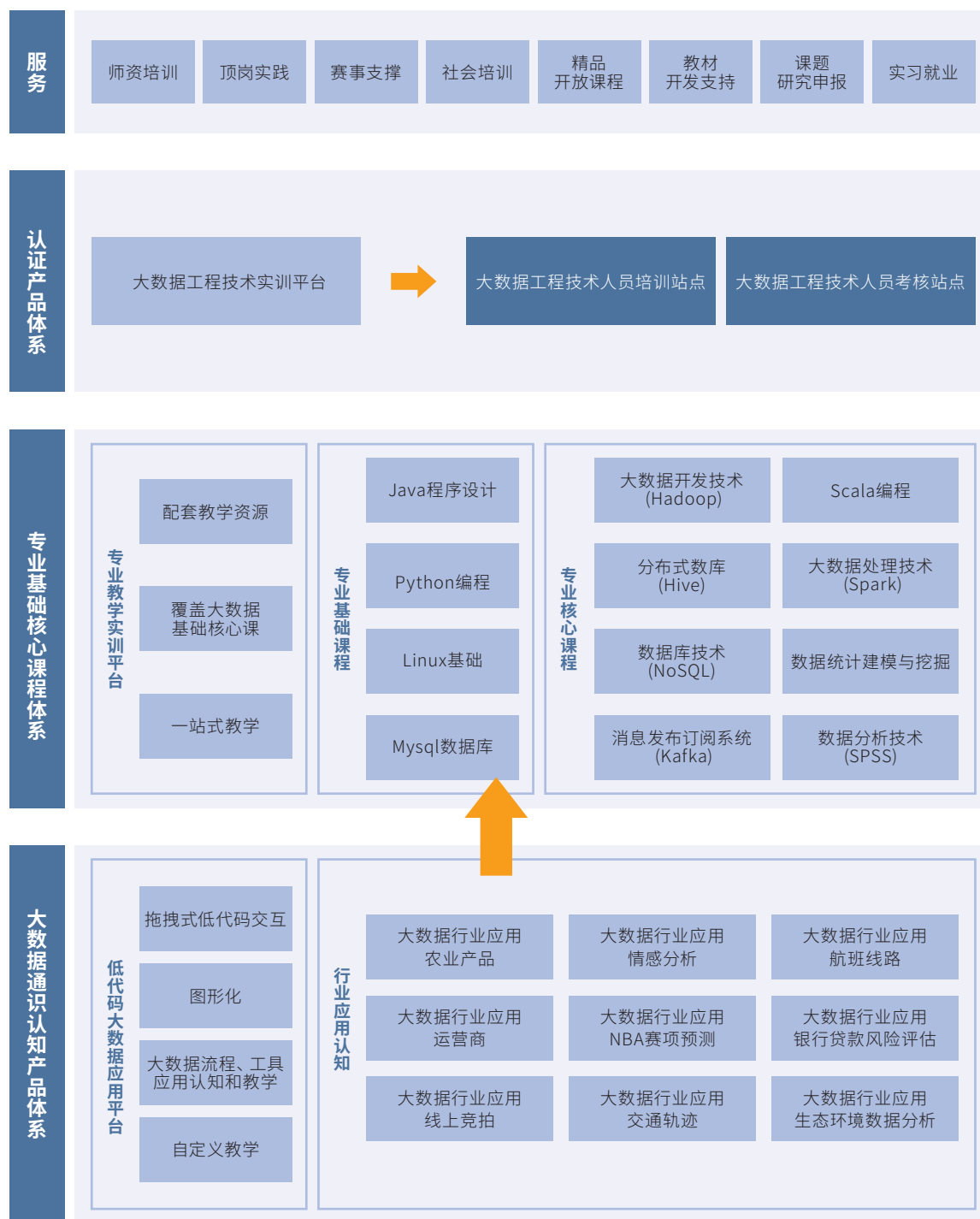
相关软件著作权500+项

相关专利共100+项

行业获奖100+项



大数据专业建设整体解决方案概览



大数据认知与应用实训室

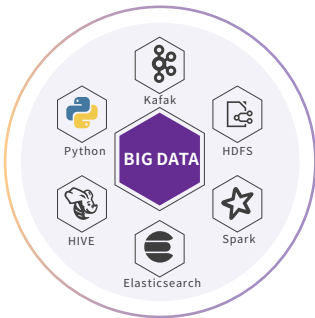
低代码大数据应用平台(VTS)

新大陆低代码大数据应用平台(VTS)汲取新大陆集团丰富的行业产品沉淀和实践经验,将真实的大数据行业中的项目案例进行教学转化、分解,基于对大数据的关键技术设计开发,用于大数据、计算机等相关专业的大数据技术的教学中。产品以“提升职业能力”为核心原则,为师生提供可视化操作的大数据项目实战环境。



● 可视化拖拽操作搭建实验

操作空间作为大数据实验的工作区域,通过拖拽工具组件进入操作空间并为其配置数据源与数据流,设置参数即可完成实验。同时也提供了整个数据实验的过程观察区域,将抽象的概念模型转化为易于理解的图形化模型展现在师生的面前,降低教学难度,提升教学效率。



● 丰富的组件工具库

提供大量可编程、调参的可视化“组件”,支持师生通过组件的拖拉拽来构建大数据的问题解决方案,并由浅入深的展开面向行业真实需求的大数据项目实践教学。



● 配套职业技能的实验资源

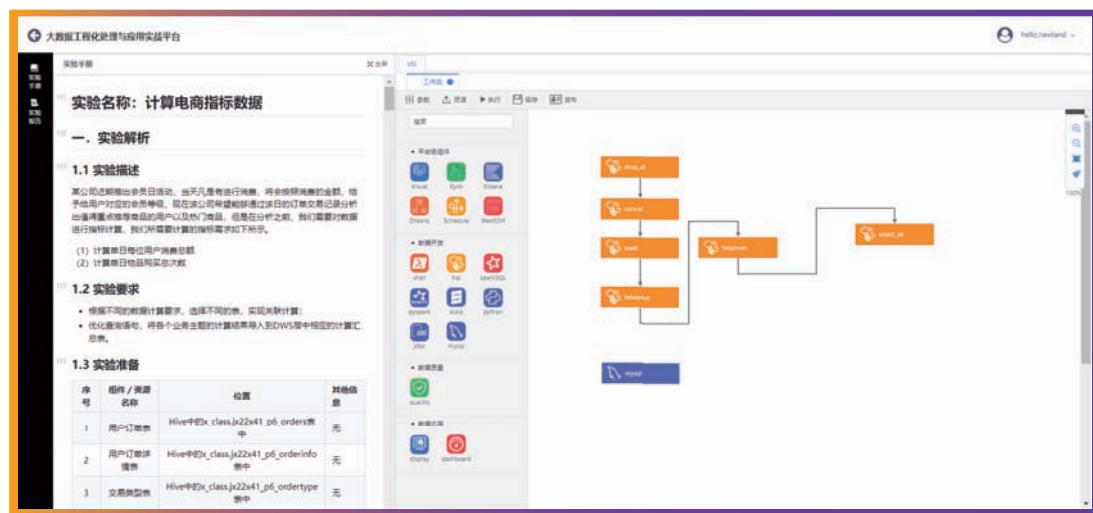
平台配套的实验资源基于大数据工程化处理与应用职业技能等级证书配套课程开发,支持数据采集、数仓开发、数据处理、数据应用工作领域的配套课程。提供实验任务书、实验手册、实验所需代码等资源。



● 真实的行业企业项目

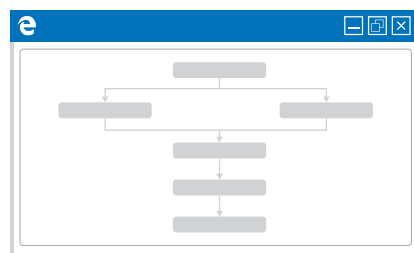
提供来自不同行业的企业真实业务数据(已脱敏)与项目书,学生结合不同行业的真实应用场景完成综合实践项目,提前了解行业需求,获取实战经验,提升未来岗位适应能力。

VTS产品架构



常规实验界面

VS



可视化实验界面

低代码大数据应用平台(VTS)采用插拔式的集成框架设计,可轻松接入各种数据应用系统,让大数据处理与应用工作变得简单。平台通过提供图形化拖拽工作流化的操作区域,支持数据采集、定时调度、数据清洗、数据分析挖掘、数据质量检测、数据输出应用与可视化展示等,包括数据处理与应用全流程场景的任务需求。通过组件的插拔式集成,让用户可以根据需要定制行业级大数据实验环境。

平台架构简图



VTTS产品特点

平台通过组件提供大数据行业中常用的各类环境与功能，以 workflow 串联打通所有已经集成的组件，在组件中编程即可完成实验的设置，并且一键实时执行。

平台组件库可支持更多组件拓展接入，可随着大数据行业更新加入新技术、新环境、新功能，保持领先。同时平台还提供了组件自定义的功能，师生可根据实际需要创建个性化的组件。



《大数据工程化处理与应用》职业技能等级证书认证考试平台

创新的可视化大数据实验平台，
轻松接入各种数据组件，让数据实验教学变得简洁轻松



1+X大数据工程化处理与应用职业技能等级证书

证书介绍



主要面向大数据行业、大数据处理与应用领域企事业单位的初级、中级大数据爬虫工程师、大数据ETL工程师、大数据开发工程师、大数据应用开发工程师、报表开发工程师等岗位，能根据业务需求，利用各种数据系统工具，编写程序完成工程化的离线或实时数据采集、数据处理与计算、开发、优化、数据查询和检索等技术工作。

本证书包括数据采集、数据处理、数据应用三大领域，从真实工作岗位拆解得出典型的工作任务，包括网络数据采集、离线数据采集、实时数据采集、作业调度、数据清洗、数据计算、数据派生、OLAP系统应用、搜索系统应用、报表系统应用等工作任务，内容涵盖行业中常用数据采集、处理与开发语言和工具，如Python、Sqoop、Hadoop、Hive、Kylin、ELK等。

职业技能登记证书的定义

本证书是面向大数据行业企业“数据处理”方向的核心岗位群要求，结合产业需求、行业经验与职业教育标准，适合职业院校学生学习与就业的高含金量证书。数据处理涵盖了大数据技术类的核心工作任务群实用性强、操作性好，适合职业教育阶段学习，也是就业面最广的大数据技术方向。

优势一

适用面广

证书要求涵盖业界主流组件工具，且不要求与特定厂商硬件强绑定，确保所学的泛用性。

优势二

课证融通

证书参考了专业教学标准中对大数据技术的要求，确保学生所学能够与专业课程相互融通。

优势三

易于教学

证书以各行业中的数据处理项目作为案例，以项目化形式提供师生生动、简便的教学形式。

教学模块



课时学分

技能要求

初级	建议课时：126 学分：8分 适用年级：中职/高职	主要面向大数据行业、大数据处理与应用企业，从事网络数据采集、离线数据采集、作业调度、数据清洗、数据计算、数据派生、OLAP系统应用、查询系统应用、报表系统应用等工作，根据项目文档及大数据处理流程，能使用图形化或可视化工具完成工程化的数据采集、数据抽取、数据转换、数据加载等能力。
中级	建议课时：160 学分：8分 适用年级：中职/高职/本科	主要面向大数据行业、大数据处理与应用企业，从事网络数据采集、离线数据采集、作业调度、数据清洗、数据计算、数据派生、OLAP系统应用、查询系统应用、报表系统应用等工作，根据业务需求，能编写脚本程序完成工程化的离线或网络数据采集、数据处理、数据查询和检索等能力。
高级	建议课时：176 学分：8分 适用年级：中职/高职/本科	主要面向大数据行业、大数据处理与应用企业，从事网络数据采集、离线数据采集、作业调度、实时数据采集、数据清洗、数据计算、数据派生、OLAP系统应用、查询系统应用、报表系统应用等工作，根据业务需求，能使用框架编写程序完成工程化的离线或实时数据采集、处理计算、网络爬虫架构设计、开发与优化和数据检索等能力。

新大陆 “岗位就业导向” 课程体系构建思路



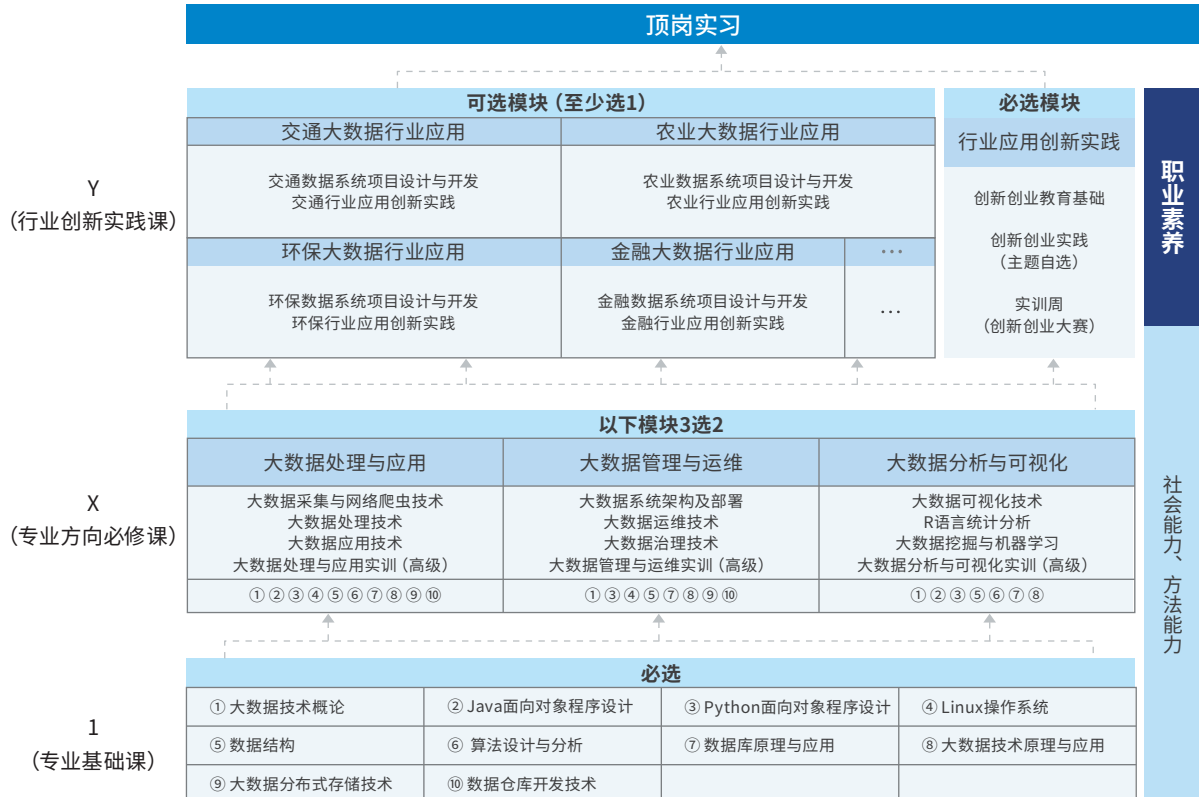
“1+X+Y” 课程体系 (中职试行版)



“1+X+Y”课程体系（高职试行版）



“1+X+Y”课程体系（本科试行版）



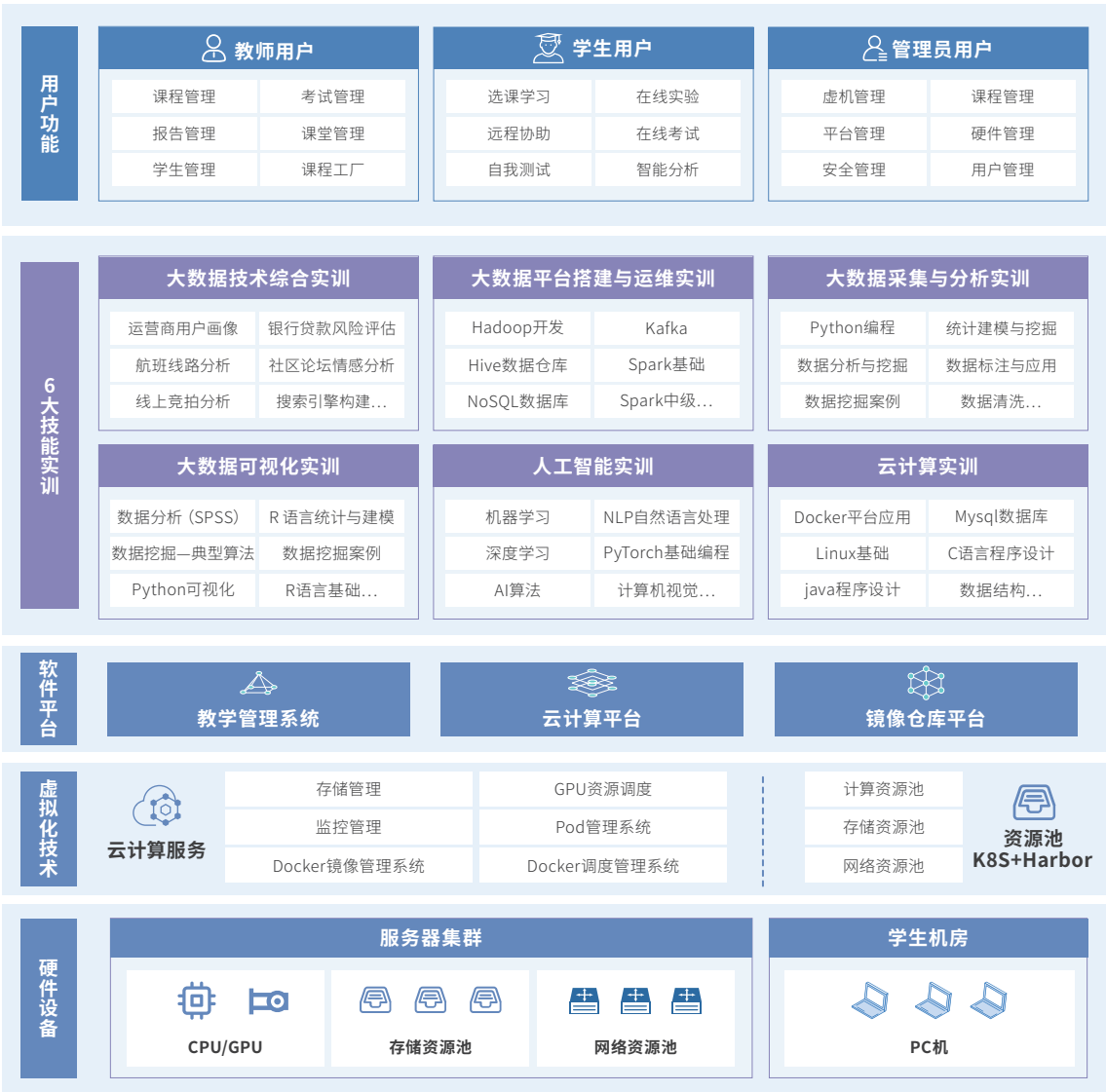
专业教学实训室

大数据专业教学实训平台

设计思路

依托集团丰富的行业经验，新大陆以大数据岗位所需技能为主要设计驱动，并汲取了丰富的行业产品沉淀和实践经验，为院校提供大数据基础教学实验室。针对大数据行业综合技能型人才培养，新大陆开发了独具特色的大数据教学实验平台，提供了稳定的实验环境，并依据人社部 and 工信部制定的《大数据工程技术人员国家职业技术技能标准（2021年版）》，打造了涵盖大数据专业知识理论、分析和处理技术、经典算法等多方面精品实验实训课程，并配备了涉及多个领域的项目案例和行业数据，能够有效地帮助学生掌握大数据专业所需的工程基础知识和专业知识，并具备就业能力和深造能力。

配套组成



大数据专业教学实训平台为学校专业教学实训提供虚拟化容器和环境，满足教师、学生通过环境所提供的组件及工具完成教学和实训任务。平台支持本地上传下载、支持课程及实训环境自主创建，满足多元化需求。



课程资源列表（部分）

序号	课程名称	序号	课程名称
01	大数据Hadoop基础	12	Spark高级案例
02	Hive离线分析	13	数据可视化技术
03	ZooKeeper分布式协调技术	14	java程序设计
04	HBase分布式数据库	15	Linux基础
05	工作流调度引擎Azkaban	16	数据库应用
06	Kafka消息发布订阅技术	17	Mysql数据库
07	Storm实时计算技术	18	数据分析与挖掘
08	Python编程	19	数据统计建模与挖掘
09	Scala语言	20	数据分析技术（SPSS）
10	Spark大数据处理基础	21	区块链理论与应用实战
11	Sqoop ETL技术	22	数据清洗

实验手册



实验十七 MapReduce 编程：检索特定偏好用户

1. 实验目的

- 掌握 *JavaSE* 基础
- 掌握 Hadoop 操作命令
- 掌握使用 *mapreduce* 编程模型实现检索特定偏好关键字的 *uid*
- 掌握在 *Eclipse* 中打包程序，并提交至 Hadoop 集群运行的步骤
- 熟悉在 HDFS 中查看程序运行结果文件

2. 实验环境

[详细讲解]

虚拟机配置：3 个（1 个 master，2 个 slave: slave01, slave02）

主节点 master 信息：

1. 操作系统：CentOS7.5
2. 软件包位置：/home/yspk/tpg
3. 数据所在位置：/home/yspk/experiment
4. 已安装软件：JDK1.8.1, Hadoop2.7.3, *Eclipse* 编辑器 Neon3.4.6.3

从节点 slave01 信息：

1. 操作系统：CentOS7.5
2. 软件包位置：/home/yspk/tpg
3. 数据所在位置：home/yspk/experiment

3、实验内容

- 在若干节点中，搭建 Hadoop 分布式集群，在搭建好的 Hadoop 分布式集群中借助 Eclipse 客户端搭建 Java 工程，导入支持 jar 包，使用 `mapreduce` 框架编程，编写 Mapper 类，编写 Reducer 类，编写 `main` 函数实现逻辑算法，以“随机数求和”的 `word`，利用 `mapreduce` 的特性对 `word` 进行数量并输出；将编译好 jar 包上传至实验环境，并在 Hadoop 集群上运行 jar 包查看输出。

实验内容流程图如下



```

graph LR
    A[在本地开发环境搭建 Hadoop 工程] --> B[打包成 jar 包并上传至 Hadoop 客户端]
    B --> C["在 Hadoop 集群上运行 jar 包  
(Hadoop 集群上安装 jar 包并运行 jar 包)"]
    C --> D["运行结果在 Hadoop 集群上输出并查看"]
    D --> E["将运行结果下载到本地查看"]
    E --> F["将运行结果上传至实验环境"]
    F --> G["在本地开发环境搭建 Hadoop 工程"]
  
```

图 3-1

4、实验关键点

- 注意需先启动集群，需要在 `master`, `slave0`, `slave02` 不同虚拟机上执行；
- 搭建 Hadoop 分布式集群时各个节点正确搭建；
- 搭建 Hadoop 分布式集群时大小写敏感问题；
- 注意工程编译后文件 jar 包需正确上传；
- 注意将编译好 jar 包上传到 Hadoop 集群；
- 注意本地和集群上得到正确输出；

```

5. 实验效果截图

展示结果为按照“ip地址”用户的 ipid 及搜索次数数据的部分截图。



```

```
[zkp@master ~]$ start-all.sh
Starting namenodes on [master]
master: starting namenode, logging to /home/zkpf/hadoop-2.7.3/logs/hadoop-zkpf-namenode-master.out
slave02: starting datanode, logging to /home/zkpf/hadoop-2.7.3/logs/hadoop-zkpf-datanode-slave02.out
slave1: starting datanode, logging to /home/zkpf/hadoop-2.7.3/logs/hadoop-zkpf-datanode-slave01.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/zkpf/hadoop-2.7.3/logs/hadoop-zkpf-secondarynamenode-master.out
starting yarn daemons
starting resourcemanager, logging to /home/zkpf/hadoop-2.7.3/logs/yarn-zkpf-resourcemanager-master.out
slave02: starting nodemanager, logging to /home/zkpf/hadoop-2.7.3/logs/yarn-zkpf-nodemanager-slave02.out
slave01: starting nodemanager, logging to /home/zkpf/hadoop-2.7.3/logs/yarn-zkpf-nodemanager-slave01.out
```

图 6-1

6.1.2 在 master 上运行 HDFS 命令 NameNode, SecondaryNameNode, ResourceManager 详细命令

```
[zkp@master ~]$ jps
5766 ResourceManager
5418 NameNode
5618 SecondaryNameNode
6045 Jps
```

图 6-2

图 6-4

6.2 打开 Eclipse，创建工程并导入依赖 jar 包

[jdk@master ~]\$ cd eclipse
[jdk@master eclipse]\$.eclipse &

6.2.1 右键单击 Eclipse 主界面空白处，依次选择 New...>Project，找到并选择 Java Project。

然后单击 Next 新建 Java 工程，依次填写为 hadoop，单击 Finish 完成创建，弹出提示框

询问是否将 Java Project 设置为默认工程，单击 是 即可

Select a wizard
Create a Java project

Wizards:

- My Java Project
- Java Project from Existing Ant Buildfile
- Plug-in Project
- General
- Eclipse Modeling Framework
- JUnit
- Other...

Next >

Create a Java Project
Create a Java project in the workspace, or in an external location.

Project name: helloworld

☒ Use default location

Location: Location...

JRE: ☒ Use an execution environment (JRE) ☐ Javac 1.8

图 6-8

6.2.3 在弹出的对话框中填入 Libraries 模块，并单击右侧的 Add External Libraries 按钮

6.2.4 将 hadoop 主目录中的 share/hadoop/common 中的 jar 包导入。然后再次点击 Add External JARs 按钮，将 share/hadoop/common/lib 目录中的所有 jar 包导入。



图 6-10



图 6-11

6.2.5 将 hadoop 主目录中的 share/hadoop/hdfs 中的 jar 都导入，在顶部点击 Add External JARs 按钮，将 share/hadoop/hdfs/lib 目录中的所有 jar 都导入。



图 6-12



大数据教学项目案例

行业	项目名称	项目方案及指导手册	项目源代码及代码注释	案例数据	系统安装部署文档
农业	大数据行业应用-农业市场大数据分析	✓	✓	✓	✓
运营商	大数据行业应用-运营商在线服务	✓	✓	✓	✓
电商	大数据行业应用-线上竞拍	✓	✓	✓	✓
舆情	大数据行业应用-情感分析	✓	✓	✓	✓
体育	大数据行业应用-NBA赛事预测	✓	✓	✓	✓
交通	大数据行业应用-交通轨迹	✓	✓	✓	✓
航空	大数据行业应用-航班线路	✓	✓	✓	✓
银行	大数据行业应用-银行贷款风险评估	✓	✓	✓	✓
互联网	大数据行业应用-搜索引擎构建	✓	✓	✓	✓
环境	大数据行业应用-生态环境数据分析	✓	✓	✓	✓

运营商在线服务平台



运营商在线服务

► 项目背景

随着通信行业的普及，以及人们对网络的需求越来越大，因此运营商的一些在线服务需求也越来越大。对于客户体验来说，电子渠道提供了一个足不出户办理业务的便捷方式，对于运营商来说，电子渠道低成本分流了实体渠道的业务压力，将线下渠道的人力资源从低价值的业务办理中释放出来。

► 项目核心知识点

HDFS、Spark、Scala、MySQL、echarts。

► 项目背景

- 申告量统计分析
- 处理量统计分析
- 回单量统计分析
- 回单率统计分析
- 重障量统计分析
- 重障率统计分析
- 及时量统计分析
- 及时率统计分析
- 成功率统计分析
- 用户多维度统计分析
- 区域服务团队监
- 热点故障区域分布监控

► 项目实训方案

架构简介：本系统架构采用模块化设计，分为数据预处理、数据存储模块、数据处理模块、数据ETL模块、结果展现模块。

模块名称：

- 数据预处理模块：将运营商给的历史数据进行格式转换。
- 数据存储模块：将清洗后的数据存储到HDFS上。

农业市场信息大数据分析平台



全国农业市场信息大数据分析平台

► 项目背景

农业大数据是融合了农业地域性、季节性、多样性、周期性等自身特征后产生的来源广泛、类型多样、结构复杂、具有潜在价值，并难以应用通常方法处理和分析的数据集合。它保留了大数据自身具有的规模巨大 (volume)、类型多样 (variety)、价值密度低 (value)、处理速度快 (velocity)、精确度高 (veracity) 和复杂度 (complexity) 等基本特征，并使农业内部的信息流得到了延展和深化。为了不断推进农业经济的优化，实现可持续的产业发展和区域产业结构优化，进一步推动智慧农业的建设进程，需要全面及时掌握农业的发展动态，这需要依托农业大数据及相关大数据分析处理技术，建设一个农业大数据分析应用平台---农业大数据平台来支撑。

► 项目核心知识点

HDFS、MapReduce、Hive、Sqoop、Hue、MySQL

► 功能模块

- 分析当天品类、品种、价格、省份的分布情况。
- 分析各个省份、各个市场、各个品种、各个品类最近5天的价格并且统计所选省份市场品种品类的当天的品种、均价、市场、时间信息。
- 分析各个年份各个月份下各个品种的价格。

► 项目实训方案

架构简介：本系统架构采用模块化设计，分为数据爬取模块、数据存储模块、数据预处理模块、数据计算模块、结果展现模块
模块名称：

- 数据爬取模块：用于网页数据的爬取，通过分析页面标签获取所需要的内容，并输出到本地文件中。
- 数据存储模块：负责存储从网页上爬取的源数据。
- 数据预处理模块：利用MapReduce计算框架对数据进行清洗操作。
- 数据计算模块：利用MapReduce计算框架对数据进行批量计算，用Hive对数据进行分析，实现数据监控、抓取量查询、数据查询、区域行情、品种对比、价格预测等。
- 结果展示模块：通过利用柠檬数图实现对结果的展示。

大数据工程技术实训室

大数据工程技术实训平台

人力资源社会保障部办公厅关于印发《专业技术人员知识更新工程数字技术工程师培育项目实施办法》的通知

发布日期：2021年10月28日 来源：【人社部】

各省、自治区、直辖市及新疆生产建设兵团人力资源社会保障厅（局），中央商务区党委人才发展局，国务院有关部门，直属机构人事部门，有关行业协会、企业、事业单位：

党的十九届五中全会提出，发展战略性新兴产业、推动互联网、大数据、人工智能等同各产业深度融合，推动先进制造业集群发展，构建一批各具特色、优势互补、结构合理的战略性新兴产业增长引擎，为加快数字技术技能人才培训，支持战略性新兴产业发展，助力数字经济和实体经济深度融合，我部根据《专业技术人员知识更新工程实施方案》和《人力资源社会保障部办公厅关于加强新职业培训工作的通知》，研究制定了《专业技术人员知识更新工程数字技术工程师培育项目实施办法》，现将实施办法印发给你们，请结合实际认真贯彻落实。

各地区、各有关部门要高度重视数字技术技能人才培训和评价工作，加强组织协调和沟通衔接，统筹谋划，科学部署，提高有序培育工作，不断提升评价质量，加强政策保障，结合实际研究制定配套实施办法，先施相关政策待遇，加强跟踪和评价全过程监管，构建政府监管、机构自律、社会监督的管理体系，加大宣传力，做好政策解读，加大宣传典型经验和做法，强化示范引领作用，请各单位对实施过程中如有问题，请及时与国家专业技术人员知识更新工程指导小组办公室（人力资源社会保障部专业技术人员管理司）联系。

任务目标

支撑专业技术人员知识更新工程，**每年**培养培训数字技术技能人才**8万人**左右

首批项目

首批证书：智能制造工程技术、区块链工程技术、**大数据工程技术**

站点信息

评价机构：**1家** 工业和信息化部教育与考试中心
培训机构：**31家**

站点要求

现阶段平均下来每个培训机构**每年**培育**1000人**

五、政策保障

各地人力资源社会保障部门要坚持以为为本，建立健全培养与使用、评价与激励相衔接的制度保障，将取得专业技术等级证书和职业技能等级证书人员信息纳入人才统计范围，兑现相关政策。

（一）参加数字技术工程师培育项目取得的相应学时记入《专业技术人员继续教育证书》，当年度全国有效。

（二）取得高级专业技术等级证书的，可作为申报高级职称评审的重要参考；取得中级、初级专业技术等级证书的，可纳入各地各部门中级、初级职称认定范围。具体职称认定或衔接办法由各地各部门结合实际自行研究制定。

（三）取得培训合格证书的，按照有关规定申领职业培训补贴。各地人力资源社会保障部门负责将培训机构目录、评价机构目录等信息纳入本地职业技能提升行动“两目录一系统”，制定完善职业培训补贴政策。

设计思路

基于《大数据工程技术人员国家职业技术技能标准（2021年版）》，新大陆工程技术人员实验室围绕大数据工程技术知识体系及应用，集成了大数据行业内企业常用的大数据技术及组件工具，为院校提供面向大数据基础技术、大数据处理与应用技术、大数据分析与管理技术、大数据管理技术等所需的各项实训和职业技术技能培训。实训资源来自多个行业，旨在为学生带来真实的实践经验和 workflows，帮助学生面向职场，面向一线，不断提高其职业胜任力。

配套教材

大数据工程技术人员（初级） 大数据基础技术	9787512910195
大数据工程技术人员（初级） 大数据处理与应用	9787512916852
大数据工程技术人员（初级） 大数据管理	9787512912144
大数据工程技术人员（初级） 大数据分析与管理	9787512916821

大数据工程技术人员实训平台介绍



实训资源包内容介绍

大数据工程技术人员实训资源包

含专业基础实训、大数据处理方向实训、大数据分析方向实训、大数据管理方向实训等。

大数据工程技术人员（初级）实验案例

实验模块	实验模块	实验内容
基础模块	大数据服务器系统搭建	服务器系统通常指的是安装在服务器硬件之上的操作系统，作为大数据系统的软件基础平台。相比于个人电脑上的桌面操作系统，亦或是安装于移动设备的嵌入式操作系统，服务器操作系统作为大数据软件系统的核心支撑，需要提供额外的管理、配置、稳定、安全等维度的能力。实验要求掌握基于硬件系统规划服务器系统部署方案并通过脚本自动化部署，完成高可用及容灾配置，将各大数据组件联通。
	大数据存储系统搭建	本实验主要包括Hadoop分布式集群、关系型数据库MySQL、非关系型数据库HBase及其操作、Hive数据仓库的部署与运维操作等内容。学生应在基本操作基础上重点掌握NoSQL及Hive数据仓库基本操作方法、优化处理方法等内容。
	大数据作业开发系统搭建	大数据技术突飞猛进，不同厂商推出了各种快速便捷的大数据处理引擎。无论是Hadoop、Storm，还是后来的Spark、Flink，没有哪一个框架可以完全支持所有的应用场景，这也表明不可能有任何一个框架可以完全取代另一个。实验将重点介绍MapReduce、Spark与Flink这三个常用大数据处理引擎。
	大数据传输系统搭建	实验将以大数据实际项目中，数据迁移框架Sqoop的使用以及实时数据采集的方法作为主要内容，包括日志监控及消息队列传输原理，实时数据采集框架Flume及消息中间件Kafka的机制等要点。
	大数据查询系统搭建	实验主要介绍以实际工作中联机数据分析和多维数据分析为场景，涉及数据的采集及数据可视化业务，同时介绍数据访问权限的问题。
	大数据安全系统搭建	本项目以真实工作中，构建大数据所需的集群安全系统作为项目内容，项目主要面向保障数据系统的访问安全，针对不同级别的安全需求，如用户权限、数据权限以及平台安全风险等方面内容，构建完整的大数据安全系统。
处理模块	网络数据处理	随着互联网技术的快速发展及大数据技术应用的普及，网站积累的数据也越来越多，如何从海量数据中发掘有价值的信息一直是大数据从业者所关注的问题。大数据技术通过对网站数据进行有针对性的采集、提取和分析，以获取更多有价值的信息，充分实现了网络数据价值和利益的最大化，为商业决策提供了有效的手段。
	离线数据处理	本实验以一个简易的电商系统为例，讲解电商系统中离线数据采集、处理与分析的基本流程。
	实时数据处理	本项目以网站点击流数据分析为例，运用大数据实时计算的基本思想，讲解网站点击流的生成、处理和实时分析的基本流程。
	交互式数据处理	本项目实验重点讲解使用Kylin对Hive中的数据进行实时聚合查询的方式。
	基于ElasticSearch的数据检索	使用ElasticSearch创建索引对图书信息进行各种类型的检索。
分析模块	BI数据分析	项目以实际工作中使用BI工具进行数据分析的项目内容为研究对象，项目内容包括使用网页版BI工具作为数据分析工具，导入数据仓库中的数据集市层数据，进行数据关联、格式调整，并选择展示字段进行数据可视化图表制作，最终完成BI数据分析。
	数据统计分析	本项目中所使用的数据来自目前正在进行的对美国马萨诸塞州弗雷明翰镇居民的心血管研究项目，其目的是分析和预测患者是否在未来的10年里有患冠心病的风险。
	数据挖掘建模	项目为电影推荐项目，该项目旨在实现movielens网站的个性化电影推荐功能。当前的项目任务为根据网站中所提供的电影数据信息，用以训练推荐模型，并最终实现对不同用户的不同推荐结果。
	数据可视化开发	本项目中，以创建全球气象仪表盘的方式，将后台数据接口中所发出的数据信息呈现在Web页面上。构建前端页面的开发工具众多，本节使用IDEA工具来构建前端网页。
管理模块	元数据管理	本项目来自某电商公司，通过其业务应用软件数据上报到服务器的形式，收集了很多用JSON格式的用户行为数据。营销人员提出需求：计算用户每日活跃指数，即从用户浏览商品个数、评论数、点赞数、广告点击数、商品消息推送数等方面，全面了解每位用户每日使用此应用软件的活跃程度，并对活跃指数高的用户给予一定的奖励或优惠。

大数据工程化处理拓展资源包

含基于Requestium的空气质量数据采集、基于Sqoop的外贸离线数据采集、基于Hive的气象数据清洗计算、基于Hive的电商数据指标计算派生、基于Hive的共享单车数据处理、基于Kylin的出租车数据应用、基于ElasticSearch的影评数据应用等实训内容资源。

大数据工程化处理与应用实验案例

实验案例	实验内容
Requestium的空气质量数据采集	现在某环保机构希望通过分析空气质量数据，为社会提供呼吸道疾病发病率预测、出行建议等服务。目标网站中包含了2014年到2015年的空气质量数据，为了能够爬取网站中的空气质量数据，爬虫开发工程师小新需要根据数据在网页结构中的位置，使用Requestium框架编写爬虫脚本，遍历爬取该网站所有的目标数据，并将爬取结果以文本格式存储在HDFS分布式文件系统中。
基于Sqoop的外贸离线数据采集	该公司通过海关授权数据交易商处获取到了三份原始数据，分别是：定期增加的贸易清单数据、贸易方式数据、地区编号数据。其中贸易清单数据已通过数据接口采集到了业务数据库 MySQL 中，而贸易方式数据和地区编号数据则分别以 TXT 格式和JSON 格式存储在 HDFS 中。这些数据零散在各个数据系统中，为方便进行后续的数据处理与分析，该公司的数据采集人员计划使用 Sqoop 和Hive 等工具将这些离线数据完整的采集到数据仓库 Hive 中的 ODS 层中。
基于Hive的气象数据清洗计算	由于原始数据包含了近 120 年的气象监测数据，数据量非常庞大，因此该公司选择了 1936 年至 1940 年的数据进行技术预研，数据采集人员将历年气象观测数据以 ORC 文件格式存储在数据仓库 Hive 的 ODS 层中，并根据年份建立了分区。除此之外，还有三张辅助说明的数据表，需通过Hive完成气象数据清洗及计算。
基于Hive的电商数据指标计算派生	某公司近期推出会员日活动，当天凡是有进行消费，将会按照消费的金额，给予给用户对应的会员等级，现在该公司希望能够通过该日的订单交易记录分析出值得重点推荐商品的用户以及热门商品，但是在分析之前，需要对数据进行指标计算，所需要计算的指标需求如下所示。（1）计算单日每位用户消费总额（2）计算单日物品购买总次数。
基于Hive的共享单车数据处理	采集部门已将共享单车数据从业务数据库中采集到了数据仓库 Hive 的贴源层，该数据记录了本市在 2013 年至 2016 年期间的近 50 万条订单信息。为了方便后续对数据的计算和派生，现需要对数据根据不同的数据处理业务需求进行清洗操作，并从中抽取以下数据表：（1）站点维度表（2）会员维度表（3）订单事实表。
基于Kvlin的出租车数据应用	数据处理人员已经将出租车数据进行了清洗操作，并以 ORC 格式存放在数据仓库 Hive 的 DW 层中。该数据包括一张订单事实表和一张地区信息维度表，存储了本市从 2013 年 1 月至 2016 年 10 月的出租车行驶记录。为了实现秒级查询，该公司决定使用 OLAP 工具 Kylin 对进行即席计算，并将查询结果保存。在技术预研阶段，项目经理计划通过 Kylin，提供以下几个计算指标数据助力公司发展战略：（1）统计 2015 年全市所有出租车运营公司每个月的盈利情况；（2）任选某一周，统计这一周本市居民每天乘坐出租车出行的频率情况；（3）任选某一天，统计当天各个时段出租车的投放量和热点地区。
基于ElasticSearch的影评数据应用	数据处理人员已将电影的影评分数和用户信息整合成一张宽表，存放于 Hive数据仓库中。在实现关键字搜索功能时，开发人员发现传统数据库系统对全文检索的支持性不佳，因此该公司决定使用 ElasticSearch（以下简称为 ES）作为全文检索引擎。在技术预研阶段，项目经理计划在搭载 ES 的分析平台 Kibana 上实现以下三个功能，用于评估功能的可行性：（1）当用户输入任意关键字时，可以根据相关性高低显示搜索结果；（2）可以根据用户的个人信息，向用户推送符合其偏好的高分电影；（3）该搜索系统可以统计出观众们最喜欢的十大电影类型。

实训室效果图



合作案例





新大陆教育公众号

联系我们 CONTACT US

地址：福建省福州市马尾区儒江西路1号新大陆科技园

北京市海淀区首体南路22号国兴大厦15层

网站 (WEB) : www.newland-edu.com

服务热线 (TEL) : 400-860-0591

新大陆教育QQ交流群: 70412420