**《网络爬虫技术与应用》**

**课程标准**

课程代码：

课程类别： 专业核心课

课程属性：

学分/学时： 3学分/48学时

开课单位：

适用专业： 大数据技术与应用

制 订 人：

审 订 人：

北京新大陆时代教育科技有限公司

2020.07

# 一、课程概述

## 课程性质

本课程是高等职业学校大数据技术与应用及相关专业的专业核心课之一。

## 课程任务

本课程主要针对大数据爬虫工程师、网络爬虫工程师等岗位，面向就职于大数据相关科研机构及IT互联网企业、互联网转型的传统型企事业单位、政府部门等，根据业务需求，从事对数据系统或网络数据进行数据采集等工作的学生开设，目的为提升学生在上述工作领域中的专业知识的应用能力、锻炼完成不同工作任务的技能。

## 课程设计思路

本课程以高等职业学校大数据相关专业的学生就业为导向，将教学内容与工作岗位对专业人才的知识要求与技能要求结合起来，将项目实践提升到最重要的位置，按照“项目引导-项目分析-项目实施-项目总结”的组织结构进行课程设计。

本课程以大数据工程化处理于应用职业技能等级标准中所列举的数据采集的工作领域，本教材分为6个项目基本总结网络数据采集全部技术，分别是：“赛事网络数据采集、购票网络数据采集、音乐网络数据采集、房产网络数据采集、航班网络数据采集以及综合网络数据采集”通过6个项目，本课程系统地将工作任务对应职业技能，锻炼了学生在数据采集领域的工作能力。

课程突出了项目实践的重要性。在内容的编排上淡化了学科性，避免介绍过多偏深的理论，而注重理论在具体运用中的要点、方法和技术操作，并结合实际范例，逐层分析和利用网络爬虫技术进行实际项目应用。

## 前后续课程

|  |  |  |
| --- | --- | --- |
| **序号** | **前续课程名称** | **前续课程为本课程支撑的主要能力** |
| 1 | 《大数据ETL处理》 | 具备使用ETL工具进行数据处理的能力 |

|  |  |  |
| --- | --- | --- |
| **序号** | **后续课程名称** | **前续课程为本课程支撑的主要能力** |
| 1 | 《大数据查询技术与应用》 | 具备对数据进行即席查询或文件检索的能力 |

# 二、课程目标

## 总体目标

本课程要求学生掌握数据采集领域的工作能力，培养学生具备高等职业院校大数据相关专业所需要的大数据采集以及网络爬虫的基本知识和技能，能根据项目需求及大数据处理流程，从事网络数据采集的工作任务，具备对大数据领域出现的新技术、新思想进一步学习的能力。

通过本课程的学习，学生能够加深对大数据采集相关技术的理解，为进一步研究和从事大数据处理和开发方向工作提供良好的基础和参考。

## 具体目标

### 知识目标

* 1. 了解网络数据结构信息；
  2. 了解基本网页结构；
  3. 掌握使用图形化工具或者原生爬虫的方式进行单网页网络数据采集的方法；
  4. 掌握使用图形化工具或者原生爬虫的方式进行连接HDFS的方法；
  5. 掌握使用图形化工具或者原生爬虫的方式进行连接MySQL的方法；
  6. 掌握能使用图形化工具或者爬虫框架的方式进行分页网络数据采集的方法；
  7. 掌握能使用图形化工具或者爬虫框架的方式进行定义采集数据格式的方法；
  8. 掌握能使用图形化工具或者爬虫框架的方式进行遍历爬取网络数据的方法。

### 能力目标

* 1. 具备使用可视化工具或者脚本的方式进行获取网络数据源准确的格式信息能力；
  2. 具备使用Python编写原生爬虫脚本，进行爬取单个网页数据并将其保存为txt格式到本地的能力；
  3. 具备使用Python编写原生爬虫脚本，进行爬取可分页网页数据并将数据全量保存到MySQL数据库的能力；
  4. 具备使用Python编写爬虫框架，进行爬取单个网页数据并将其保存为csv格式到HDFS的能力；
  5. 具备使用Python编写爬虫框架，进行爬取分页网络数据并增量存储到MySQL的能力；
  6. 具备使用Python编写爬虫框架，进行遍历爬取网络数据并将其保存为Json格式文件到HDFS的能力；
  7. 具备使用Python编写爬虫框架，进行遍历爬取分页网络数据并将其定义多种指定类型数据文件保存到Hive中；
  8. 具备使用Crontab或者Azkaban将指定的采集脚本进行调度的能力。

### 素质目标

1. 培养谦虚、好学、勤于思考、认真做事的良好习惯---严谨的开发流程和正确编程思路；
2. 培养团队协作能力---相互沟通、互相帮助、共同学习、共同达到目标；
3. 提升自我展示能力---讲述、说明、表述和回答问题；
4. 培养自我学习能力---利用书籍或网络上的资料帮助解决实际问题。

# 三、课程内容及情境设计

本课程以赛事网络数据采集、购票网络数据采集、音乐网络数据采集、房产网络数据采集、航班网络数据采集以及综合网络数据采集，6个项目为载体，与企业合作设计选取15个项目任务，根据岗位工作任务要求，确定学习目标和学习任务内容；本课程采取项目驱动教学模式，以学生为主体，以任务为导向组织教学考核。

### 项目一 赛事网络数据采集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目序号** | 项目一 | | | |
| **项目名称** | 赛事网络数据采集 | | | |
| **培养能力** | 1. 根据赛事网页的原始代码，能分析网页结构并找到数据位置；  2. 根据赛事网页结构，能使用Python编写原生爬虫脚本，进行采集单网页网络数据；  3. 根据存储系统的导入方式，能将采集的数据定义为txt格式保存到本地的指定路径下。 | | | |
| **项目任务** | **知识要求** | **技能要求** | | **学时** |
| 网络数据采集 | 1. 了解赛事业务数据逻辑；  2. 掌握网络爬虫概念；  3. 掌握HTML基本格式；  4. 了解网络爬虫的工作流程；  5. 了解网络爬虫策略。 | 1. 能够观察网页源代码，获取数据位置信息；  2. 能够编写原生爬虫脚本进行单网页网络数据采集；  3. 能够使用Python将数据存储为txt格式到本地指定路径下。 | | 4 |
| **教学情境** | 项目驱动、实验演示、边讲边做、自学探究 | | **项目学时** | 4 |

### 项目二 购票网络数据采集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目序号** | 项目二 | | | |
| **项目名称** | 购票网络数据采集 | | | |
| **培养能力** | 1. 根据购票网页的原始代码，能分析网页结构并找到数据位置；  2. 根据购票网页结构，能使用Python编写原生爬虫脚本，进行采集分页网络数据；  3. 根据存储系统的导入方式，能将采集的数据全量保存到MySQL数据库的指定库表中。 | | | |
| **项目任务** | **知识要求** | **技能要求** | | **学时** |
| 网络数据采集 | 1. 了解购票业务数据逻辑；  2. 掌握Python库的使用；  3. 掌握PyMySQL语法；  4. 掌握带参数的GET请求。 | 1. 能够观察网页源代码，获取数据位置信息；  2. 能够编写原生爬虫脚本进行分页网络数据采集；  3. 能够使用Python将数据全量存储到MySQL数据库指定库表中。 | | 6 |
| **教学情境** | 项目驱动、实验演示、边讲边做、自学探究 | | **项目学时** | 6 |

### 项目三 音乐网络数据采集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目序号** | 项目三 | | | |
| **项目名称** | 音乐网络数据采集 | | | |
| **培养能力** | 1. 根据音乐网页的原始代码，能分析网页结构并找到数据位置；  2. 根据音乐网页结构，能使用Python编写爬虫框架，进行采集单网页网络数据；  3. 根据存储系统的导入方式，能将其保存为csv格式并存储到HDFS的指定路径下。 | | | |
| **项目任务** | **知识要求** | **技能要求** | | **学时** |
| 网络数据采集 | 1. 掌握爬虫框架详解；  2. 了解爬虫框架的优缺点；  3. 了解原生与框架的区别；  4. 掌握Spiders爬虫脚本概念（页面解析）；  5. 掌握Python连接HDFS方法。 | 1. 能够观察网页源代码，获取数据位置信息；  2. 能够编写爬虫框架进行单网页网络数据采集；  3. 能够使用Python将数据保存为csv格式并存储到HDFS指定路径下。 | | 8 |
| **教学情境** | 项目驱动、实验演示、边讲边做、自学探究 | | **项目学时** | 8 |

### 项目四 房产网络数据采集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目序号** | 项目四 | | | |
| **项目名称** | 房产网络数据采集 | | | |
| **培养能力** | 1. 根据房产网页的原始代码，能分析网页结构并找到数据位置；  2. 根据房产网页结构，能使用Python编写爬虫框架，进行采集分页网络数据；  3. 根据存储系统的导入方式，能将其增量存储到MySQL的指定库表中；  4. 根据系统资源需求，能将指定的作业进行定时调度。 | | | |
| **项目任务** | **知识要求** | **技能要求** | | **学时** |
| 网络数据采集 | 1. 了解分页数据结构详解；  2. 了解分页方式与方法；  3. 掌握初始化文件概念；  4. 掌握定义实体类概念与用法。 | 1. 能够观察网页源代码，获取数据位置信息；  2. 能够编写爬虫框架进行分页网络数据采集；  3. 能够使用Python将数据增量存储到MySQL的指定库表中。 | | 4 |
| 作业调度 | 1. 了解Crontab定时器；  2. 掌握Crontab详解；  3. 掌握Crontab定时器语法；  4. 掌握Shell脚本语法。 | 1. 能够根据时间要求，编写对应的启动脚本；  2. 能够使用Crontab进行定配置时间规则。 | | 4 |
| **教学情境** | 项目驱动、实验演示、边讲边做、自学探究 | | **项目学时** | 8 |

### 项目五 航班网络数据采集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目序号** | 第3篇章 | | | |
| **项目名称** | 航班网络数据采集 | | | |
| **培养能力** | 1. 根据航班网页的原始代码，能分析网页结构并找到数据位置；  2. 根据航班网页结构，能使用Python编写爬虫框架，进行遍历采集网络数据；  3. 根据存储系统的导入方式，能将其保存为Json格式文件到HDFS的指定路径下；  4. 根据系统资源需求，能将指定的作业进行定时调度。 | | | |
| **项目任务** | **知识要求** | **技能要求** | | **学时** |
| 网络数据采集 | 1. 了解航班业务数据逻辑；  2. 掌握复杂页面解析方法；  3. 了解中间件概念与用法；  4. 掌握下载器概念与用法；  5. 掌握项目配置概念与用法；  6. 初识Json。 | 1. 能够观察网页源代码，获取数据位置信息；  2. 能够编写爬虫框架进行遍历网络数据采集；  3. 能够使用Python将数据保存为Json格式文件到HDFS的指定路径下。 | | 6 |
| 作业调度 | 1. 了解作业调度理念；  2. 了解作业调度工具；  3. 了解作业调度优缺点。 | 1. 能够根据时间要求，编写对应的启动脚本；  2. 能够使用Crontab进行定配置时间规则。 | | 4 |
| **教学情境** | 项目驱动、实验演示、边讲边做、自学探究 | | **项目学时** | 10 |

### 项目六 综合网络数据采集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目序号** | 项目六 | | | |
| **项目名称** | 综合网络数据采集 | | | |
| **培养能力** | 1. 根据网页的原始代码，能分析网页结构并找到数据位置；  2. 根据航班网页结构，能使用Python编写爬虫框架，进行遍历采集分页网络数据；  3. 根据存储系统的导入方式，能将其定义多种指定类型数据文件保存到Hive的指定库表中；  4. 根据系统资源需求，能将指定的作业进行定时调度。 | | | |
| **项目任务** | **知识要求** | **技能要求** | | **学时** |
| 网络数据采集 | 1. 了解电商业务数据逻辑；  2. 了解Java爬取详解；  3. 了解反爬策略；  4. 掌握Pyhive语法；  5. 掌握如何连接Hive；  6. 掌握综合相关知识（如何汇总数据、如何根据多个不同类型的完整网络进行获取数据）；  7. 掌握HTTP请求包。 | 1. 能够观察网页源代码，获取数据位置信息；  2. 能够编写爬虫框架进行遍历爬取分页网络数据；  3. 能够使用Python将数据定义多种指定类型数据文件；  4. 能够将数据结果存储到Hive的指定库表中。 | | 8 |
| 作业调度 | 1. 掌握Azkaban调度规则；  2. 掌握打包与上传语法；  3. 掌握配置定时方法。 | 1.能够使用命令行脚本方式进行打包；  2. 能够使用Azkaban进行创建调度项目；  3. 能够使用Azkaban进行上传压缩包；  4. 能够使用Azkaban进行作业脚本调度。 | | 4 |
| **教学情境** | 项目驱动、实验演示、边讲边做、自学探究 | | **项目学时** | 12 |

# 四、教学实施建议

## 教学方法建议

为引导学生了解和掌握新知识，培养学生学习兴趣，加强学生对数据采集、数仓开发、数据处理以及数据应用的能力，实现与企业岗位工作“零距离对接”，本课程建议采用的特色教学方法有以下几种：

（1）“设计项目任务驱动”教学法

通过在真实的任务中探索学习，不断地提高学生成就感，更大地激发他们的求知欲望，逐步形成一个感知心智活动的良性循环，从而培养出独立探索、勇于开拓进取的创新能力。

（2）项目教学法

在教学中把知识与技能进行有机的结合，充分发掘学生的创造潜能，提高学生解决实际问题的综合能力，为学生零距离就业奠定基础。

（3）讨论式与启发式教学相结合

对于实践性强的内容，安排专题学生自学，然后由一个学生在课堂上讲述，大家再一起讨论、分析和评价，这样使每个学生都有兴趣积极参与，活跃课堂气氛，培养自学的能力。

（4）模拟和实际相结合的环境教学法

针对大数据处理与大数据应用中实际岗位的工作环境，综合运用了模拟环境和实际环境的教学方法。

## 教材选用或编写建议

**教材选用：**

本课程建议使用《网络爬虫技术与应用》一书展开教学。

**编写建议：**

本课程教材编写要以实际教学内容为依据，由专业教师和企业权威专家共同编写。教材设计思路应充分体现任务引领、实践导向，引入典型项目和产品案例，各个职业能力对应相应知识点、技能点，每个知识点、技能点应配备相应的二维码扩展资源，二维码资源的展现形式有文字、视频、动画等，教材形式选用活页式教材，开发和合理运用各类信息化、智能化教学资源，满足多样化需求，创新服务供给模式，服务学生终身学习。

## 考核评价建议

本课程评价采用“过程性评价和终结性评价相结合”的方式，各项评价所占权重如表1所示：

表1 课程评价所占权重分配表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 过程性评价（60%） | | | | | | 终结性评价（40%） | |
| 项目一 | 项目二 | 项目三 | 项目四 | 项目五 | 项目六 | 综合实操考试 | 平时考勤纪律 |
| 5% | 8% | 10% | 10% | 12% | 15% | 30% | 10% |

各个综合任务的评价指标与评分标准如表2所示：

表2 项目评价表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 姓名 | 考勤 | 任务完成情况 | 综合素质 | | | 小组评价 | 加分 | 合计 |
| 规范意识 | 合作沟通意识 | 踏实 |
|  |  |  |  |  |  |  |  |  |

## 教学环境建议

为了顺利开展本课程的教学工作，需要提供相应的硬件设备和软件工具，给学生一个真实的实践环境。（主要指教室环境、校内外实验实训设备配置建议，包括实践场地、实训设备配置要求等）

1.硬件设备

* 投影仪一套：实现展示课件、展示实例、演示操作等功能
* 教师\学生用机：50台以上
* 实训室面积：120平米以上

2.软件工具

* 操作系统：Windows 7/Windows 10及以上操作系统
* 电子教室软件：实现屏幕广播、访问控制等功能
* 局域网：实现教学广播等功能
* 实验平台：新大陆大数据教学实验平台