# 基于前程无忧网站招聘信息的大数据分析案例教学指导手册

**教学适用的课程：**《大数据架构与技术》

**教学适用的专业：**电子信息

**教学目的与用途：**学习实际分布式数据爬取技术；掌握数据清洗、数据分析与可视化的相关知识点；掌握大数据处理与分析的基本流程。

**教学内容：**

(1) 理论依据：本案例基于前程无忧网站招聘信息进行大数据分析。案例整合了多种的数据分析与数据挖掘方法，对大数据相关课程的教学有较好的参考价值。

(2) 涉及知识点：分布式爬虫、Hadoop，Spark，大数据分析，数据可视化。

(3) 分析路径：首先部署Hadoop+Spark大数据处理平台；然后使用Scrapy-Redis分布式爬虫框架，爬取前程无忧网站上的招聘信息，存储至MongoDB数据库并进行清洗处理；最后，使对招聘信息进行数据分析与可视化展示，从微观上挖掘并展示企业招聘需求的详细信息，从宏观上分析计算机相关行业招聘现状。

**启发思考题：**

(1) 考虑使用一些其它数据挖掘或机器学习算法，以其他角度从前程无忧招聘网站的大量数据中挖掘招聘的深度信息，并能寻找到一定的应用价值。

(2) 本案例给出的数据爬取、数据挖掘、以及可视化展示为分离的三种任务，考虑将这三个阶段的任务进行聚合，构建出招聘信息大数据系统，可以实时爬取数据进行处理，并展示最新招聘信息动态。

**建议课堂计划：**

1. 时间安排：3课时
2. 学生学习准备：了解常见的数据爬取、分布式计算、数据可视化的理论知识。
3. 分组及讨论内容：分组讨论启发思考题。
4. 案例开场白：随着互联网行业的兴起，计算机行业相关的招聘信息不胜枚举。前程无忧作为全国性权威人才招聘网站，拥有全国最大的职位信息库， 它提供了最新最全最准确的信息，为企业和求职者提供人才招聘、求职等在内的全方位人才资源服务。在本案例中。我们将使用大数据的相关技术与方法对前程无忧网站上的大量招聘信息进行爬取与分析，从微观上挖掘并展示企业招聘需求的详细信息，从宏观上分析计算机相关行业招聘现状。
5. 结束总结：本案例基于前程无忧网站招聘信息，进行计算机相关岗位的大数据分析研究。在该案例中，我们主要接触了分布式数据爬取技术、基于Hadoop与Spark的分布式大数据处理技术、以及使用可视化绘图工具展示大数据分析的结果。需要注意的是，课程所学内容是有限的，希望对该问题感兴趣的同学可以自行搜集相关资料，在数据爬取、数据处理、数据展示上引入更多方法，或思考创新性的大数据应用场景。
6. 案例引导建议：在教师简要分析应用场景后，鼓励学生自主考虑如何使用所学知识解决该场景问题。

**参考文献：**

[1] IDC. Data Age 2025[EB/OL]. [2018-11]. https://www.seagate.com/cn/zh/our-story/data-age-2025

[2] 刘军. Hadoop大数据处理[M]. 北京: 人民邮电出版社, 2013.

[3] Singh R, Kaur P J. Analyzing performance of Apache Tez and MapReduce with hadoop multinode cluster on Amazon cloud[J]. Journal of Big Data, 2016, 3(1): 1.

[4] 李宁. 基于HDFS的小文件存储模型研究与实现[D].

**其他教学支持材料：**

1. 本案例提供配套的PPT、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。