# 基于前程无忧网站招聘信息的大数据分析

**作者**：潘禹帆，张紫微，范玮旻，葛港，黄旺辉，王彬，冯永

**单位**：重庆大学，计算机学院

**案例版权**：该案例归重庆大学计算机学院所有

**涉及的知识点**：分布式爬虫、Hadoop，Spark，大数据分析，数据可视化

**案例来源及案例真实性情况**：该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 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。本文案例基于Hadoop完全分布式平台，以Spark开源集群计算环境为补充，对前程无忧网站上2020年计算机行业相关的招聘信息进行了多方面的数据分析以及数据可视化。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。

**关键词**：招聘信息，Hadoop，Spark，大数据分析，数据可视化

## 1 引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的关键问题为基于前程无忧网站招聘信息的大数据分析，需引导学生进行的主要内容有：（1）搭建分布式的数据爬取、存储与处理平台；（2）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；（3）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。

## 2 背景介绍

信息时代，无论是公共数据还是个人数据都在快速增长，或者可以说是呈现井喷式爆发。数据已经渗透到生活和工作的方方面面，成为时代的特征。纽约证交所每天产生的交易数据大约在4TB至5TB之间；脸谱网（Facebook）存储的照片超过2400亿张，并以每月至少7PB的速度增长。据IDC发布《数据时代2025》的报告显示，全球每年产生的数据将从2018年的33ZB增长到175ZB，相当于每天产生491EB的数据[1]。

随着互联网+、智能+时代的到来，计算机专业人才越来受欢迎，近年来，高校的计算机专业也不断扩招。互联网公司需要哪些人才？哪一类职业更抢手？各类职业工作经验对应薪资分布如何？相信很多从事计算机行业的人员都十分关心这样的问题。前程无忧(jobs)招聘网是中国领先的专业招聘网站，经第三方监测机构iResearch等调查，其连续多年市场占有率第一，先后多次被评为“中国最具影响力的人才招聘网站”。该网站拥有全国最大的职位信息库，为计算机行业的从业人员提供了大量最新的招聘信息[2]。利用这些招聘信息，我们可以从工资、地区、学历、经验、职位要求、公司等多个方面进行数据分析，以可视化方式展示分析结果，为就业学生提供参考。

## 3 内容

**3.1 搭建大数据处理平台**

首先我们需要搭建本案例所需的数据爬取、存储与处理平台，该平台主要使用Scrapy-Redis分布式爬虫框架，MongoDB数据库，Hadoop+Spark大数据处理平台，具体的技术框架如下图所示：

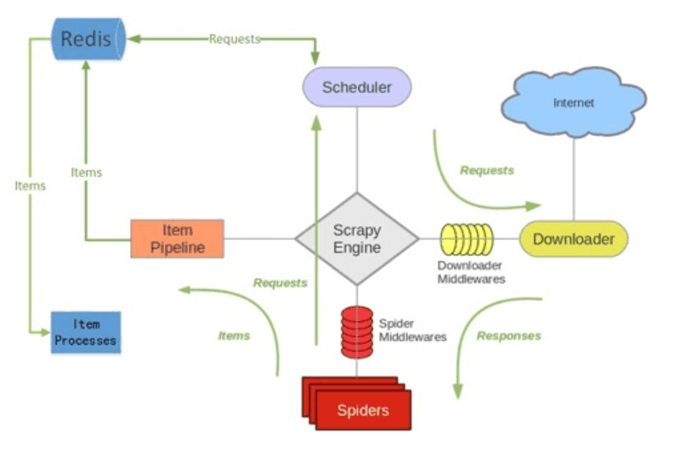


**3.2 数据爬取与处理**

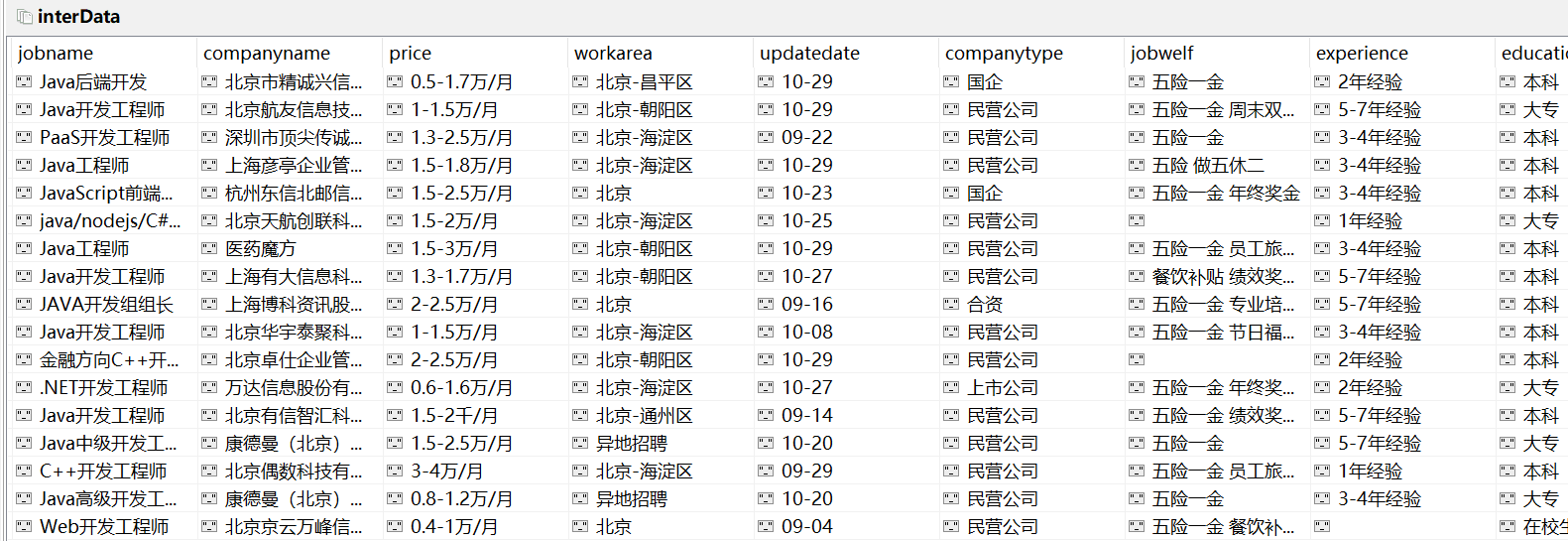
本案例使用Python编写分布式爬虫程序来实现数据集的收集。数据来源为前程无忧招聘网上的计算机行业招聘信息，主要爬取了前程无忧网站上包括北京，上海，广州，深圳在内的20多个全国热门城市在计算机行业招聘岗位，岗位信息包括：职位名称，薪资，待遇，经验、学历以及具体要求，公司名称、类型、领域、规模等相关信息。

本案例应用Scrapy-Redis分布式爬虫框架，从多台机器上抓取一个网站的数据，并将抓取到的数据写入到MongoDB数据库中。

首先Slave端从Master端（安装有Redis系统）获取任务（Request、url）后进行数据抓取，在Slave抓取数据的同时，产生的新任务Request会被提交给 Master 处理；Master端只有一个Redis数据库，负责将未处理的Request去重和任务分配，将处理后的Request加入待爬队列，并存储爬取的数据。Scrapy-Redis分布式爬虫框架如下图所示：



下图展示了部分爬取到的岗位数据：



爬取到的数据包含的字段含义及数据类型如下表所示：

|  |  |  |
| --- | --- | --- |
| 字段名 | 含义 | 数据类型 |
| jobname | 岗位名称 | String |
| companyname | 公司名称 | String |
| companytype | 公司性质 | String |
| jobwelf | 岗位福利 | String |
| experience | 工作经验要求 | String |
| education | 学历要求 | String |
| needpeople | 招聘人数 | String |
| demand | 岗位职责 | String |
| companysize | 公司规模 | String |
| companyind | 公司所属行业类型 | String |
| categories | 职位所属类型 | String |
| maxSalary | 月最高工资 | String |
| minSalary | 月最低工资 | String |
| city | 工作地点所属城市 | String |
| area | 工作地点所属区域 | String |
| province | 工作地点所属省份 | String |
| isFresh | 是否为应届生/在校生 | String |

经过上述的数据收集后，可以发现在数据集中含有一部分脏数据，主要是因为企业重复发送招聘信息造成的重复数据与和与默认值不匹配的字符串类型的错误数据。因此，我们需要对数据集进行清洗处理，主要步骤包含：

（1）根据职位名称，公司名称，工作地点，薪资，具体要求，学历，工作经验和所招人数对数据进行去重处理；

（2）处理price字段：price字段的格式不一致，有的是千/月，万/月，万/年，还有的是万以上以及元/天和空值。需要将其统一转换为千/月，并按薪资上限和薪资下限进行存储；

（3）处理workarea字段：将工作地点城市－区域根据－拆分成城市和区域两个字段，进行分别存储。若只有城市，则将区域设为空值；

（4）处理education字段：对所有数据添加一个字段isFresh，用于表示是否是“在校生/应届生”。如果某条数据的education字段包含“在校生/应届生”，则将其isFresh字段设为‘是’，否则将其设为‘否’，并去除education中的‘在校生/应届生’；

（5）处理province字段：为每条数据添加相应的province字段，将province字段的值更新为city字段表示城市所对应的省份。如果city为直辖市，则其对应province字段填写其直辖市城市名；

本案例数据集的使用方式为将MongoDB中经过清洗的数据导出为csv文件，然后上传至 HDFS文件系统中，进而可以用MapReduce对数据进行并行化地快速处理。

**3.3 数据分析与可视化**

对上述获取到的数据集进行分析的角度有非常多的选择，包括但不限于（1）平均工资相关分析：不同岗位的平均工资、不同城市的平均工资；（2）中值工资相关分析：工资和学历关系、工资和经验关系、工资和企业性质关系、工资和公司行业关系；（3）学历经验分析：热门岗位的学历分析、热门岗位的经验分析、不同岗位的学历需求、不同岗位的经验需求；（4）公司相关分析：不同行业公司的公司数量、不同性质企业的公司数量、不同规模公司的地域分布；（5）岗位相关分析：不同城市岗位总数、TopK岗位数量、大数据相关岗位的职位要求。

在本案例中，我们给出两组使用MapReduce进行统计分析的例子，并鼓励学生使用MapReduce的方式设计并行化程序完成其它角度下的大数据分析。

（1）岗位数量分析

在该部分中，我们需要统计不同城市的岗位数，统计不同规模公司的地域分布，按公司性质统计岗位数量，按公司的类型统计公司数量，按学历要求统计岗位数量，按经验统计岗位数量。以按公司的类型统计公司数量统计为例， Map方案如下：

定义集合，其中是数据库内*i*列所存储的元素。可得{‘民营公司’，‘上市公司’，‘国企’，‘合资’，‘外资（非欧美）’，‘外资（欧美）’，‘创业公司’，‘事业单位’，‘政府机关’，‘非营利组织’，‘外企代表处’}。对每次选取到的KEY实例key，设定其对应的value=1。

设计一个Reduce方案如下：

建立一个数据字典，数据字典中key分别对应KEY集合中的元素，对全部选取的key-value实例进行访问，若key值相同则将其所对应的value自增1。

（2）中值工资分析

在该部分在，我们需要分析中值工资和学历关系、中值工资和经验关系、中值工资和企业性质关系、中值工资和公司行业关系。选取中值工资是因为相比平均工资它更具有代表性，平均工资会因为个别数据的偏大或偏小影响最后平均工资结果，但是中值工资不会受个别特殊数据的影响，更具有一般性。以分析中值工资和经验的关系统计为例，设计Map方案如下：

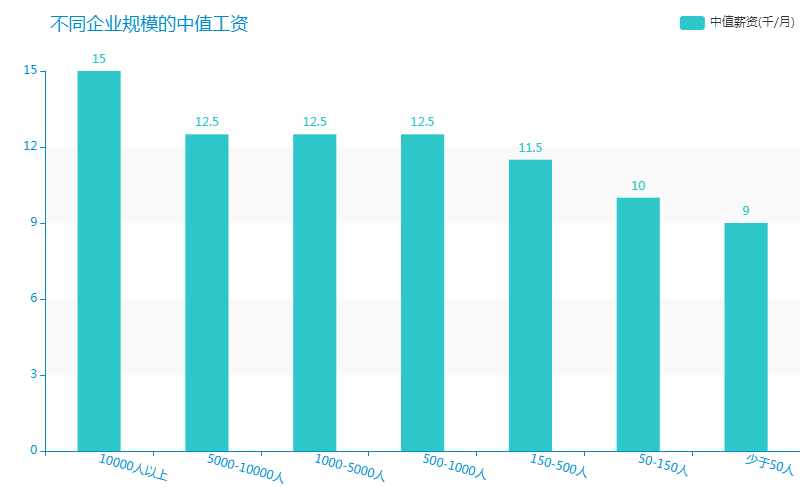
定义集合，其中是数据库内i列所存储的元素。i=10可得KEY={‘1-2年’，‘3-4年’，‘5-7年’，‘8-9年’，‘10年以上’，‘无需经验’}，对每次选取到的KEY实例key，由于工资是用最大工资maxSalary和最小工资minSalary区间表示，所以先由maxSalary和minSalary计算月平均工资，然后设定其对应的value值为这个平均工资。

设计Reduce方案如下：

建立一个数据字典，数据字典中key分别对应KEY集合中的元素，对全部选取的key-value实例进行访问，若key值相同则将其对应的value值加入到这个key值所对应的列表中，最后利用numpy库的中位数函数media()求相应的中值工资。

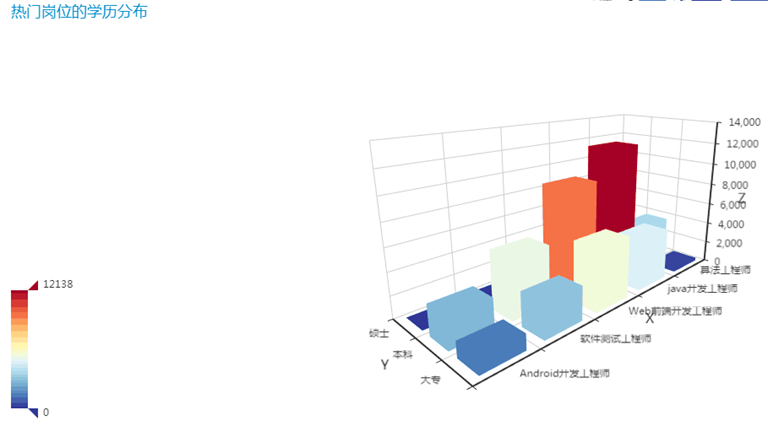
接下来，可使用Pyecharts将分析结果可视化，以便于展示招聘大数据，一下为数据可视化的一些例子：

（1）不同企业规模的中值工资



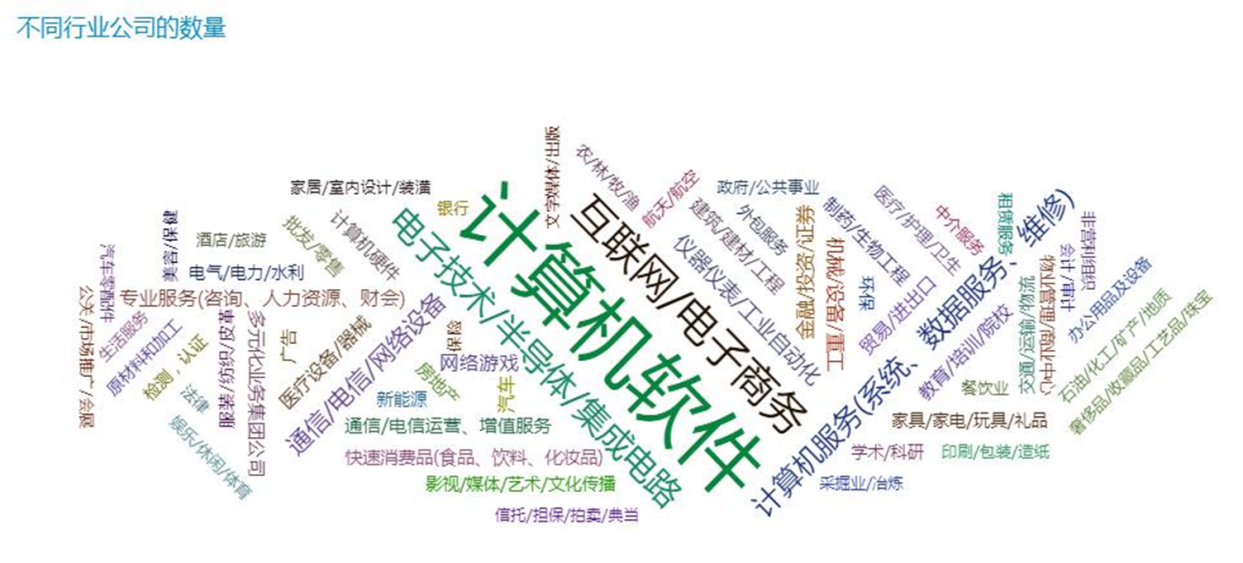
如上图所示，在中小型企业中（即公司人数在50-1000人左右），工资基本和企业规模基本成正相关关系，企业规模越大，则工资越高。但公司规模在1000-5000和5000-10000人规模的公司中值工资区别不大。其中，10000人以上企业工资最高。由此可见，万人以上规模的公司薪资水平高，择业时可优先考虑企业规模大的公司。

（2）热门岗位的学历分布



上图展示了热门岗位的学历分布：五个热门岗位需求最多的仍是本科生，大专学历中Web前端需最多的。求最多，本科学历中Java开发工程师需求最多，而硕士学历中算法工程师需求最多。由此可见算法相关岗位对学历要求较高，而开发相关岗位则对学历无过高要求。整体来看，本科生是目前几个热门岗位需求量

（3）不同行业公司数量



上图展示了不同行业公司的数量：在招聘计算机类岗位的公司中，占比最大的公司行业仍是计算机软件、互联网电子商务和电子技术等，但是公司行业类型多种多样，覆盖几乎所有行业，可见计算机相关行业就业面十分广泛。

## 4 小结

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，本案例的内容仅为指导性的过程，在实际教学中，可保持基本研究内容不变，鼓励学生引入其它的数据预处理、数据挖掘、机器学习方法完成任务，使用其它公开数据集进行实验研究，并考虑针对实际应用的进一步拓展。

## 附录

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

2. 本案例参考文献如下：

[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].