大数据技术与应用赛项竞赛试题（样卷）

近年来随着 IT 产业的加速发展，全国各地对 IT 类的人才需求也越来越多，“ABC 公司”为了明确今后 IT 产业人才培养方向，在多地进行 IT 公司岗位情况调研分析。你所在的小组将承担模拟调研分析的任务，通过在招聘网站进行招聘信息的爬取，获取到公司名称、工作地点、岗位名称、招聘要求、招聘人数等信息，并通过对数据的清洗和分析，得出各地域招聘人数，“大数据”相关职位招聘数量，以绘制雷达图展示各地平均薪资情况。

为完成该项任务，你所在的小组计划选用在业界广泛应用的“Python 和 JAVA”语言，作为整个项目的基础语言，并综合利用 requests 模块、MapReduce、MySQL、Flask 开源框架、Jinja2 模板引擎和 ECharts 组件提高开发效率并实现项目要求，由于本次为模拟任务，总数据量不会过大，项目组计划使用分布式节点 Hadoop 模式，本次项目环境搭建采用服务器集群方式，配置了小规模的技术演示环境，通过在招聘网站上爬取到的相关信息，使用 requests 模块、Hive、Python、JAVA 等手段对数据进行爬取、清洗、整理、计算、表达、分析，力求实现对 IT 人才就业信息拥有更清晰的掌握。

请按照下面步骤完成本次技术展示任务，并提交技术报告。

任务一：Hadoop 相关组件安装部署（15 分）

当前环境中已安装 Hadoop 运行环境和 MySQL 数据库，相关安装信息如下表所示，请在此环境基础上按照相关操作步骤安装 Hive 组件。

1. 将指定路径下的 Hive 安装包解压并更名；
2. 设置 Hive 环境变量；
3. 编辑 Hive 相关配置文件；
4. 初始化 Hive 元数据；
5. 启动并保存输出结果。

任务二：数据采集与数据预处理（20 分）

1. 从指定招聘网站中抓取数据，提取有效数据项，并保存为 json 格式文件；
2. 设置 post 请求参数并将信息返回给变量 response；
3. 将提取数据转化成 json 格式，并赋值变量；
4. 用 with 函数创建 json 文件，通过 json 方法，写入 json 数据；
5. 爬取的数据需要导入 hadoop 平台进行数据清洗与分析，在 HDFS 文件系统中创建文件夹，并将 json 文件上传到该文件夹下。

任务三：数据清洗与分析（25 分）

1. 为便于数据分析与可视化，需要对爬取出的数据进行清洗，使用 Java 语言编写数据清洗的 MapReduce 程序；
2. 将清洗程序上传至 hadoop，并对 HDFS 的原始数据进行清洗；
3. 将清洗后的数据加载到 Hive 数据仓库中；
4. 通过运行 HQL 命令完成数据分析统计；

5. 在 hive 中执行 sql 脚本，并查看表中大数据核心技能的出现次数。

任务四：数据可视化（20 分）

为更好的将数据分析结果表达出来，需要对数据分析的结束进行可视化呈现，可视化呈现，本次数据可视化需要呈现三部分内容：

1. 按要求使用柱状图展示各城市招聘人数，并在前端显示。要求：

主标题：各地域招聘人数

副标题：（--招聘人数变化趋势）

横坐标：城市信息，纵坐标：招聘人数

输出柱状图

2. 按要求使用折线图展示“大数据”相关职位招聘数量差异，并在前端显示。要求：

主标题：大数据相关职位分析

副标题：（--招聘数量变化趋势）

横坐标：岗位名称，纵坐标：岗位数量

输出折线图

3. 通过雷达图展示各地平均薪资的情况。要求：

主标题：各地平均薪资

输出雷达图

任务五：完成分析报告（15 分）

请结合数据分析结果回答以下问题：

1. 根据分析结果说明大数据岗位所需要的主要技能包含哪些，为什么（4 分）
2. 根据分析结果说明各地大数据产业发展情况（4 分）
3. 根据市场需求分析，大数据行业的人才培养方向有哪些，为什么（4 分）
4. 请简述，今后大数据产业地域发展方向在哪里（3 分）
5. 竞赛结果提交要求：

1）任务成果需拷贝至提供的 U 盘中。在 U 盘中以 XX 工位号建一个文件夹（例如 01），将所有任务成果文档保存至该文件夹中。

2）竞赛提交的所有文档中不能出现参赛队信息和参赛选手信息，竞赛文档需要填写参赛队信息时以工位号代替（XX 代表工位号）。