**背景描述**

近年来随着IT产业的快速发展，全国各地对IT类的人才需求数量也在不断增多，“XHS集团”为了明确今后IT产业人才培养方向，在多个省份进行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，并通过数据的清洗和分析，最终分析出当前IT产业热门岗位、大数据相关岗位从业人员数量等信息，并进行可视化呈现。

本次为模拟任务，项目组计划使用分布式节点Hadoop模式，环境搭建采用服务器集群方式，通过在招聘网站上爬取到的相关信息，实现对数据进行爬取、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。

作为该项目组的技术人员，你们是这次技术方案展示的核心成员，请按照下面步骤完成本次技术展示任务，并提交技术报告，祝你们成功。

# 任务一：Hadoop平台及组件的部署管理（15分）

1. 将指定路径下的Hive安装包解压到指定目录下；
2. 把解压后的apache-hive-1.1.0-bin文件夹更名为 hive；
3. 设置Hive环境变量，并使环境变量只对当前root用户生效；
4. 将Hive安装目录下hive-default.xml.template文件更名为 hive-site.xml；
5. 在Hive安装目录里创建临时文件夹；
6. 新建并配置hive-site.xml文件，实现“Hive元存储”的存储位置为MySQL数据库；
7. 初始化Hive元数据；
8. 启动Hive。

# 任务二、数据采集（15分）

1. 从各类招聘网站中抓取数据，提取数据项至少包括以下字段：“公司名称”、“工作城市”、“工作要求”、“招聘人数”、工资情况”（格式：‘底薪-上限’）、“name”(岗位名称)、“detail”(职位详情)，并保存；
2. 爬取的数据需要导入hadoop平台进行数据清洗与分析，将数据保存至HDFS系统中。

# 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，清洗出不包含空字段的职位信息数据，清洗后的职位数据各字段间用竖线“,”分割。数据清洗采用MapReduce程序。**

1. 使用编写数据清洗的MapReduce程序（原始数据中已包含职位描述的分词结果）；
2. 将编写好的文件上传并对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。

**2、数据清洗需要将清洗程序jar包上传至hadoop平台上进行运行，并将清洗后结果保存至hive数据库中，以便后续应用。**

1. 将jar包上传至hadoop平台的/root目录下；
2. 执行mapreduce任务，并写出执行命令；

3）将执行成功后的数据保存在HDFS文件系统的/Clean目录下；

1. 将清洗后数据存储到hive中task数据库的cleandata表中。
2. **运行HQL命令，完成以下数据的分析统计**

1）统计各岗位招聘数量，将结果写入cleantable表中；

2）查询“数据”相关岗位的技能要求，将查询结果写入table\_bigdata表中；

3）在hive中执行keycount.sql脚本，并查看keycount表中所有核心技能的出现次数。

注：核心技能关键词如下：c++、Scala、FFlume、Flink、ETL、数学、数据仓库、Hbase、Hadoop、Python、Java、Kafka、Storm、Linux、Hive、Spark。

# 任务四、数据可视化（20分）

数据可视化呈现内容如下：

1. 使用柱状图展示指定方向当前招聘数量最多的前十个热门职位；
2. 使用折线图展示“大数据”相关职位招聘数量差异；
3. 通过词云的方式对“大数据”相关职位的知识技能要求进行可视化展示。

# 任务五：综合分析（15）

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，并阐述详细理由；
2. 根据市场需求分析IT产业工程人才培养方向有哪些，并阐述详细理由；
3. 根据市场需求分析，大数据行业的人才培养方向有哪些，并阐述详细理由；
4. 请简述，“XHS集团”今后IT产业人才培养方向是什么。