**2020年全国高校计算机能力挑战赛**

**大数据应用赛答题卡第二部分**

**队长姓名： 队长学校：**

# 第二部分：大数据处理部分（Mapreduce）（20分）

数据文件 维修换件清单1.csv，维修换件清单2.csv，两个个的结构是一样的 分别为某工业公司维修部门的维修记录，有2个不同的组进行记录，2个文件里面有重复统计的数据，数据文件中的每行都是一个数据。 

## 任务一、请使用MapReduce对两个进行合并，合并后维修记录数据进行去重。去重后输出数据为weixiuhuanjian，并把重复的输出到 chongfudata,并对去除重复后的数据统计总行数,统计重复数据的行数（10分）

源码：

* + 1. 除重后文件中的数据内容截图：
    2. 重复文件中的数据内容截图
    3. 统计除重后数据总行数 截图
    4. 统计重复行数 截图
    5. 对数据文件合并的相关代码截图
    6. 对合并后的数据进行去重的相关代码截图
    7. 统计行数的相关代码截图

## 任务二、请使用MapReduce统计 维修换件清单1.csv中的每个设备，维修的换件数量和维修费用 ，并输出格式 为 设备名称，换件数量，维修费用（5分）。

源码：

1）结果截图：

1. 代码截图：

## 任务三、请使用MapReduce统计 维修换件清单1.csv中的每年 维修费用最高的前三条记录（5分）。返回格式：年 ，机台名称，维修费用

源码：

1）结果截图：

2）代码截图：

# 第三部分：大数据分析（25分）

Hive作为基于hadoop的海量数据处理工具之一，在企业中有广泛应用。可以将复杂的MapReduce计算逻辑使用SQL逻辑处理。

数据：

数据说明：

Product：产品表

product\_id, 产品ID

product\_name, 产品名称

unit\_price, 产品单价

inventory 产品库存

salse： 销售表

order\_id, 订单ID

product\_id, 产品ID

customer\_id,客户ID

year, 年

month,月

day, 日

freight\_charges,运费

unit\_price,单价

quantity 销售数量

## 任务一、hive创建数据库（10分）

### 1、创建一个数据库，以你的用户名命名，创建成功后使用use命令切换为该库，并执行set hive.cli.print.current.db=true;截图作为答案（3分）

1）创建数据库命令截图：（2分）

2）执行命令结果截图：（1分）

### 2、在HIVE中创建需要的表，并写出建表语句，表结构如下：（3分）

(1).产品表(product)

|  |  |  |
| --- | --- | --- |
| 字段 | 字段描述 | 字段类型 |
| product\_id | 产品ID | String |
| product\_name | 产品名称 | String |
| unit\_price | 产品单价 | decimal(20,3) |
| inventory | 产品库存 | decimal(20,3) |

1. 销售表(salse)

|  |  |  |
| --- | --- | --- |
| 字段 | 字段描述 | 字段类型 |
| order\_id | 订单ID | String |
| product\_id | 产品ID | String |
| customer\_id | 客户ID | String |
| year | 年 | Int |
| month | 月 | Int |
| day | 日 | Int |
| freight\_charges | 运费 | decimal(20,3) |
| unit\_price | 单价 | decimal(20,3) |
| quantity | 销售数量 | decimal(20,3) |

1. 创建product表 语句截图：（1分）
2. 创建salse表 语句截图：（1分）

2）创建成功，执行show tables；截图；（1分）

### 3、将数据加载到表中，写出加载数据的语句(4分)

数据文件名称分别与表名称对应，（字段分隔符为|），注意，linux系统和hive默认不支持中文，要自己设置成utf-8编码才可以支持中文。

1) 导入product表 语句截图：（1分）

2）导入salse表 语句截图：（1分）

3）执行 select \* from product 并截图（其中table\_name 为前面创建的表名）（1分）

4）执行 select \* from salse并截图（1分）

## 任务二、hive数据分析（15分）

### 查询每个产品的平均运费，平均销售额(2分)

sql语句截图：

sql语句结果截图：

### 查询选购了3种产品以上（包含3种产品）的客户ID (3分)

sql语句截图：

sql语句结果截图：

### 使用所建的表统计 各产品总销售数量，总销售额（单价\*销售数量），按照总销售额降序排列（4分）。 输出的字段 包含

|  |  |
| --- | --- |
| 字段 | 字段描述 |
| product\_name | 产品名称 |
| Total\_ quantity | 总销量 |
| Total\_salse | 总销售额 |

1）sql语句截图：

1. sql语句结果截图：

### 使用所建的表统计每年销售额最多的4个产品信息（6分）。 输出的字段 包含 产品id，年，总销售额

1）sql语句截图：

2）sql语句结果截图：

# 第四部分：大数据可视化（本题目不限制语言，图形包）（15分）

数据：



数据描述：

* 价格---租金 单位元/月
* 面积---房屋实际面积 单位：㎡
* 编号---房屋的编号
* 户型---房屋的总体户型情况
* 楼层---房屋所在小区总共有多少层 其中出租屋在多少层
* 朝向---房屋的方向
* 位置1---房屋所在的区
* 位置2---房屋所在街道
* 小区---房屋所在小区名称

## 任务一、户型占比分析（饼图）（3分）

公寓户型占比图，分析各种不同的户型的数量在总数的占比图形（使用饼图表示）

代码截图：

结果截图：

分析结论（1分）：

## 任务二、行政区域房屋数量占比分析（饼图）（3分）

公寓数在不同的行政区域占比图，分析不同占比图形（使用饼图表示）并分析

代码截图：

结果截图：

分析结论（1分）：

## 任务三、公寓的价格分布图形（直方图）（4分）

用直方图画出，公寓的价格分布（直方图）并分析工具价格的分布情况给出分析结论

代码截图：

结果截图：

分析结论（1分）：

## 任务四、公寓面积对比分析（柱状图）（5分）

通过柱状图 分析不同面积的数量。并给出面积的分析结论

代码截图：

结果截图：

分析结论（1分）：

# 第五部分 数据分析文档（5分）

主要数据处理的处理方式，数据分析的思路，数据可视化部分的呈现思路。

文档描述：

# 第六部分：综合题（15分）

## 任务一、Spark技术栈有哪些组件，每个组件都有什么功能，适合什么应用场景。（5分）

**答：**

## 任务二、本题目使用spark进行数据分析，不限制语言（可以使用pyspark，Rspark，scala等语言）（10分）

数据 

数据说明：

|  |  |
| --- | --- |
| 字段 | 字段说明 |
| positionName | 职位名称 |
| salary | 薪水 |
| workYear | 工作年限 |
| city | 城市 |
| companyShortName | 公司简称 |
| companySize | 公司规模 |
| district | 所在区 |
| financeStage | 融资阶段 |
| industryField | 所在领域 |
| thirdType | 职位类型 |
| resumeProcessDay | 简历日处理 |
| resumeProcessRate | 简历处理率 |
|  |  |

### 拆分字段salary -> min\_salary,max\_salry，并且取薪资的整数 （2分）

如果salary字段是否包含类似'10k-20k\*15薪'的值 如果存在 则把 \*15这样的类似数据给去掉，保留10K-20K; 则拆分为 min\_salary,max\_salry 分别是 10，20，

如果salary字段 类似10-20K 则拆分为 min\_salary,max\_salry 分别是 10，20，

如果 salay字段为空值或者 面议，则拆分为 min\_salary,max\_salry 分别是 -1，-1，

拆分后字段 表结构为

|  |
| --- |
| 字段 |
| positionName |
| salary |
| min\_salary |
| max\_salry |
| workYear |
| city |
| companyShortName |
| companySize |
| district |
| financeStage |
| industryField |
| thirdType |
| resumeProcessDay |
| resumeProcessRate |
|  |

源码：

代码截图：

结果截图：

### 分析不同城市 薪资概况（工资的最小平均值，最大平均值）（2分）

返回结果包含 城市名称，薪资最小平均值avg（min\_salry），最大平均值avg（max\_salry）

源码：

代码截图：

结果截图：

### 分析不同城市 不同工作年限公司职位需求数量（2分）

返回结果包含 城市名称，工作年限，需求数量，

源码：

代码截图：

结果截图：

### 分析获得每个城市，在 职位名称中包含 “数据分析“” 职位需求top3公司名称（4分）

返回结果包含 城市，职位名称，公司名称，需求数量，

源码：

代码截图：

结果截图：