**2020年全国高校计算机能力挑战赛**

**大数据应用赛答题卡第三部分**

**队长姓名： 队长学校：**

# 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：

### 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：