2023-2024 年广东省职业院校技能大赛

中职组大数据应用与服务赛项

样

题

2

一、背景描述

大数据时代背景下，人们生活习惯发生了很多改变。在传统运营模式中，缺乏数据积累，人们在做出一些决策行为过程中，更多是凭借个人经验和直觉，发展路径比较自我封闭。而大数据时代，为人们提供一种全新的思路，通过大量的数据分析得出的结果将更加现实和准确。平台可以根据用户的浏览，点击，评论等行为信息数据进行收集和整理。通过大量用户的行为可以对某一个产品进行比较准确客观的评分和评价，或者进行相应的用户画像，将产品推荐给喜欢该产品的用户进行相应的消费。

因数据驱动的大数据时代已经到来，没有大数据，我们无法为用户提供大部分服务，为完成中国汽车分车型每月销售量数据分析工作，你所在的小组将应用大数据技术，通过 Python 语言以数据采集为基础，将采集的数据进行相应处理，并且进行数据标注、数据分析与可视化、通过大数据业务分析方法实现相应数据分析。运行维护数据库系统保障存储数据的安全性。通过运用相关大数据工具软件解决具体业务问题。你们作为该小组的技术人员，请按照下面任务完成本次工作。

二、模块一：平台搭建与运维

（一）任务一：大数据平台搭建

1. 子任务一：基础环境准备

本任务需要使用 root 用户完成相关配置，安装 Hadoop 需要配置前置环境。命令中要求使用绝对路径，具体要求如下：

（1）配置三个节点的主机名，分别为 master、slave1、slave2，然后修改三个节点的 hosts 文件，使得三个节点之间可以通过主机名访问，在 master 上将执行命令 cat /etc/hosts 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下；

（2）将 /opt/software 目录下将文件 jdk-8u191-linux-x64.tar.gz 安装包（若 slave1、slave2 节点不存在以上文件则需从 master 节点复制）解压到 /opt/module 路径中（若路径不存在，则需新建），将 JDK 解压命令复制并粘贴至【提交结果.docx】中对应的任务序号下；

（3）在 /etc/profile 文件中配置 JDK 环境变量 JAVA_HOME 和 PATH 的值，并让配置文件立即生效，将在 master 上 /etc/profile 中新增的内容复制并粘贴至【提交结果.docx】中对应的任务序号下；

（4）查看 JDK 版本，检测 JDK 是否安装成功，在 master 上将执行命令 java -vserion 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下；

（5）创建 hadoop 用户并设置密码，为 hadoop 用户添加管理员权限。在 master 上将执行命令 grep 'hadoop' /etc/sudoers 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下；

（6）关闭防火墙，设置开机不自动启动防火墙，在 master 上将执行命令 systemctl status fireawlld 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下；

（7）配置三个节点的 SSH 免密登录，在 master 上通过 SSH 连接 slave1 和 slave2 来验证。

2. 子任务二：Hadoop 完全分布式安装配置

本任务需要使用 root 用户和 hadoop 用户完成相关配置，使用三个节点完成 Hadoop 完全分布式安装配置。命令中要求使用绝对路径，具体要求如下：

（1）在 master 节点中的 /opt/software 目录下将文件 hadoop-3.3.6.tar.gz 安装包解压到 /opt/module 路径中，将 hadoop 安装包解压命令复制并粘贴至【提交结果.docx】中对应的任务序号下；

（2）在 master 节点中将解压的 Hadoop 安装目录重命名为 hadoop ，并修改该目录下的所有文件的所属者为 hadoop，所属组为 hadoop，将修改所属者的完整命令复制并粘贴至【提交结果.docx】中对应的任务序号下；

（3）在 master 节点中使用 hadoop 用户依次配置 hadoop-env.sh、core-site.xml、hdfs-site.xml、mapred-site.xml、yarn-site.xml、masters 和 workers 配置文件，Hadoop 集群部署规划如下表，将 yarn-site.xml 文件内容复制并粘贴至【提交结果.docx】中对应的任务序号下；

服务器	master	slave1	slave2
DHFS	NameNode		
HDFS	SecondaryNameNode		
HDFS	DataNode	DataNode	DataNode
YARN	ResourceManager		
YARN	NodeManager	NodeManager	NodeManager
历史日志服务器	JobHistoryServer		

（4）在 master 节点中使用 scp 命令将配置完的 hadoop 安装目录直接拷贝至 slave1 和 slave2 节点，将完整的 scp 命令复制并粘贴至【提交结果.docx】中对应的任务序号下；

（5）在 slave1 和 slave2 节点中将 hadoop 安装目录的所有文件的所属者为 hadoop，所属组为 hadoop。

（6）在三个节点的 /etc/profile 文件中配置 Hadoop 环境变量 HADOOP_HOME 和 PATH 的值，并让配置文件立即生效，将 master 节点中 /etc/profile 文件新增的内容复制并粘贴至【提交结果.docx】中对应的任务序号下；

（7） 在 master 节点中初始化 Hadoop 环境 namenode，将初始化命令及初始化结果（截取初始化结果日志最后 20 行即可）粘贴至【提交结果.docx】中对应的任务序号下；

（8） 在 master 节点中依次启动 HDFS、YARN 集群和历史服务。在 master 上将执行命令 jps 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下；

（9） 在 slave1 查看 Java 进程情况。在 slave1 上将执行命令 jps 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下。

3. 子任务三：Zookeeper 集群安装配置

本任务需要使用 root 用户完成相关配置，已安装 Hadoop 及需要配置前置环境，具体要求如下：

（1）在 master 节点将 /opt/software 目录下的 apache-zookeeper-3.8.3-bin.tar.gz 包解压到 /opt/module 路径下， 将解压命令复制并粘贴至【提交结果.docx】中对应的任务序号下；

（2）把解压后的 apache-zookeeper-3.8.3-bin 文件夹更名为 zookeeper-3.8.3，将命令复制并粘贴至【提交结果.docx】中对应的任务序号下；

（3）设置 zookeeper 环境变量，将新增的环境变量内容复制并粘贴至【提交结果.docx】中对应的任务序号下；

（4）创建 zookeeper 配置文件 zoo.cfg 并配置 master、slave1、slave2 三个节点的集群配置，其中 dataDir 参数设置为 /opt/module/zookeeper-3.8.3/data ，提交 zoo.cfg 配置内容至【提交结果.docx】中对应的任务序号下；

（5）在 master 节点上创建文件 myid 用于标识服务器序号，并将文件内容设置为 1；

（6）在 master 节点上将配置的 zookeeper 环境变量文件及 zookeeper 解压包拷贝到 slave1、slave2 节点，提交命令至【提交结果.docx】中对应的任务序号下；

（7）在 slave1 节点上修改 myid 文件内容修改为 2，在 slave2 节点上修改 myid 文件内容修改为 3，提交命令和结果截图粘贴至【提交结果.docx】中对应的任务序号下；

（8）在 master 节点、slave1 节点、slave2 节点分别启动 zookeeper，提交命令和结果截图粘贴至【提交结果.docx】中对应的任务序号下；

（9）在 master 节点、slave1 节点、slave2 节点分别查看 zookeeper 的状态，提交命令和结果截图粘贴至【提交结果.docx】中对应的任务序号下；

（10）在 master 节点查看 Java 进程，提交命令和结果截图粘贴至【提交结果.docx】中对应的任务序号下。

（二）任务二：数据库服务器的安装与运维

1. 子任务一：MySQL 安装配置

本任务需要使用 rpm 工具安装 MySQL 并初始化，具体要求如下：

（1） 在 master 节点中的 /opt/software 目录下将 MySQL 5.7.44 安装包解压到 /opt/module 目录下；

（2）在 master 节点中使用 rpm -ivh 依次安装 mysql-community-common、mysql-community-libs、mysql-community-libs-compat 、 mysql-community-client 和 mysql-community-server 包，将所有命令复制粘贴至【提交结果.docx】 中对应的任务序号下；

（3）在 master 节点中启动数据库系统并初始化 MySQL 数据库系统，将完整命令复制粘贴至【提交结果.docx】 中对应的任务序号下；

2.子任务二：MySQL 运维

本任务需要在成功安装 MySQL 的前提，对 MySQL 进行运维操作，具体要求如下：

（1）在 MySQL 中创建一个新的数据库 namedb，并将创建命令复制粘贴至【提交结果.docx】中对应的任务序号下；

（2）为 namedb 数据库创建一个新用户 username，并只授予该数据库的访问权限，将创建用户及授权命令和结果截图至【提交结果.docx】中对应的任务序号下；

（3）显示 MySQL 数据库的当前版本信息，并将显示版本的命令和结果截图至【提交结果.docx】中对应的任务序号下；

（4）查看并修改 MySQL 服务器的默认字符集为 utf8mb4，将查看和修改命令及和结果截图至【提交结果.docx】中对应的任务序号下；

（5）备份 namedb 数据库，并将备份命令和结果截图至【提交结果.docx】中对应的任务序号下；

3.子任务三：数据表的创建及维护

（1）根据以下数据字段在 namedb 数据库中创建一个学生表（student）。学生表字段如下：

字段	类型	中文含义
id	int	学号
name	varchar	姓名
age	int	年龄
major	varchar	专业

（2）根据以下数据字段在 namedb 数据库中创建一个课程表（course）。课程表字段如下：

字段	类型	中文含义
id	int	课程 ID
course_name	varchar	课程名称
credit	int	学分

将这两个 SQL 建表语句和结果分别复制粘贴至【提交结果.docx】中对应的任务序号下；

（3）为 student 表添加一个新的字段 email（varchar 类型），并将 SQL 语句粘贴至【提交结果.docx】中对应的任务序号下；

（4）编写 SQL 查询，在 student 表中查找所有专业为“计算机科学”的学生，并将查询语句粘贴至【提交结果.docx】中对应的任务序号下。

三、模块二：数据获取与处理

（一）任务一：数据获取与清洗

1. 子任务一：数据获取

有一份中国汽车销售数据：年份、月份、排名、车型、厂商、销量、售价（万元）。

并且已存入到 sale_car.csv 文件中，请使用 pandas 读取 sale_car.csv 并将数据集的前 10 行打印在 IDE 终端的截图复制粘贴至【提交结果.docx】中对应的任务序号下。

2. 子任务二：使用 Python 进行数据清洗

请使用 pandas 库加载并分析相关数据集，根据题目规定要求使用 pandas 库实现数据处理，具体要求如下：

（1）删除销量为空的记录，并将结果存储为 cleaned_data_c1_N.csv，N 为删除的数据条数。

（2）删除销量为负数的记录，并将结果存储为 cleaned_data_c2_N.csv，N 为删除的数据条数。

（3）删除销售月份格式不正确的记录，并存储为 cleaned_data_c3_N.csv，N 为修改的数据条数。

（4）删除车型名称重复的记录，并将结果存储为 cleaned_data_c4_N.csv，N 为删除的数据条数。

（5）输出车型或厂商字段为空的记录，并存储为 cleaned_data_c5_N.csv，N 为替换的数据条数。

将该 5 个文件名截一张图复制粘贴至【提交结果.docx】中对应的任务序号下。

（二）任务二：数据标注

1. 子任务一：车型豪华等级标注

使用 Python 编写脚本，根据车辆的售价（售价平均值）将汽车分为不同的豪华等级。具体的分类要求如下：

（1）**经济型**：售价低于 10 万元；

（2）**中档型**：售价在 10 万元至 30 万元之间；

（3）**豪华型**：售价超过 30 万元；

在数据集中新增一列**“豪华等级”**，根据上述标准对每种车型进行豪华等级标注，存入 luxury_level_mark.csv 文件中。具体格式如下：

编号	车型名称	售价平均值（万元）	豪华等级
1	轩逸	13.735	中档型

将 luxury_level_mark.csv 打开后的直接截图（不用下拉）复制粘贴至【提交结果.docx】中对应的任务序号下。

2．子任务二：销售等级标注

使用 Python 编写脚本，基于每月销售量，为每种车型分配一个销售等级。具体的分类要求如下：

（1）**A 级**：月销售量超过 10000 辆；

（2）**B 级**：月销售量在 5000 至 9999 辆之间；

（3）**C 级**：月销售量低于 5000 辆；

在数据集中新增一列**“**销售等级**”**，根据上述标准为每种车型进行销售等级标注。存入 sales_level_mark.csv 文件中。具体格式如下：

编号	车型名称	销售量	销售等级
1	轩逸	61170	A 级

将 sales_level_mark.csv 打开后的直接截图（不用下拉）复制粘贴至【提交结果.docx】中对应的任务序号下。

（三）任务三：数据统计

1．子任务一：HDFS 文件上传下载

本任务需要使用 Hadoop、HDFS 命令，已安装 Hadoop 及需要配置前置环境，具体要求如下：

（1）在 HDFS 目录下新建目录 /file2_1，将新建目录的完整命令粘贴至【提交结果.docx】中对应的任务序号下；

（2）修改权限，赋予目录 /file2_1 最高 777 权限，将修改目录权限的完整命令粘贴至【提交结果.docx】对应的任务序号下；

（3）下载 HDFS 新建目录 /file2_1，到本地容器 master 指定目录 /tmp 下，将完整命令粘贴至【提交 结果.docx】中对应的任务序号下。

2. 子任务二：计算输入文件中的单词数

本任务需要使用 Hadoop 默认提供的 wordcount 示例来完成单词数统计任务，具体要求如下：

（1）在 HDFS 上创建 /user/hadoop/input 目录；

（2）在 master 节点将 /var/log/dmesg 文件上传到 HDFS 的 /user/hadoop/input 目录下；

（3）使用 Hadoop 中提供的 wordcount 示例对 HDFS 上的 dmesg 文件进行单词统计，并将统计结果存储到 HDFS 的 /user/hadoop/output 目录下；

（4）查看 HDFS 中的 /user/hadoop/output 单词数统计结果并将结果前十行截图粘贴至【提交结果.docx】中对应的任务序号下。

3. 子任务三：数独解算器

本任务需要使用 Hadoop 默认提供的 sudoku 示例来完成数独题目的解题任务，具体要求如下：

（1）使用 Hadoop 提供的 sudoku 示例计算以下数独题目：

```
8 5 ? 3 9 ? ? ? ?
? ? 2 ? ? ? ? ? ?
? ? 6 ? 1 ? ? ? 2
? ? 4 ? ? 3 ? 5 9
? ? 8 9 ? 1 4 ? ?
3 2 ? 4 ? ? 8 ? ?
9 ? ? ? 8 ? 5 ? ?
? ? ? ? ? ? 2 ? ?
? ? ? ? 4 5 ? 7 8
```

（2）将数独解题结果截图粘贴至【提交结果.docx】中对应的任务序号下。

四、模块三：业务分析与可视化

（一）任务一：数据分析与可视化

1. 子任务一：数据分析

在这个任务中，我们将运用 Python 对汽车销售数据进行深入分析，以揭示市场的关键趋势和洞察。参赛者需要运用 Python 的数据处理和分析库，如 Pandas 来完成以下任务：

（1）分析 2022 年每个月的汽车总销量，并找出销量最高的月份；

（2）计算每个厂商的年度总销量，并进行倒序排序展示前五名；

（3）计算不同价格区间（0-15 万、15-30 万、30-60 万、60 万以上）的车型销量，车型售价取平均值，并找出最受欢迎的价格区间；

（4）分析不同车型的销售趋势，找出年度销量增长最快的车型；

（5）筛选出售价在 10 万元以下的车型，并统计这些车型的总销量；

将该 5 个统计结果在 IDE 的控制台中打印并分别截图复制粘贴至【提交结果.docx】中对应的任务序号下。

2. 子任务二：数据可视化

在这个任务中，参赛者将使用 Matplotlib 库来创建直观、互动的图表，以揭示数据中的关键模式和趋势。具体要求如下：

（1）使用柱状图展示 2022 年每个月的汽车总销量，每个柱子代表一个月份，其高度表示该月的汽车总销量；

（2）创建条形图比较不同厂商的年度总销量，每个条形代表一个品牌，其长度表示该品牌在一年中的总销量；

（3）制作散点图探索车型售价（车型售价取平均值）与销量之间的关系，每个点代表一个车型，其位置根据该车型的售价和销量确定；

（4）使用饼图展示不同车型在总销量中的占比，每个饼图的切片代表一个车型，其大小表示该车型在总销量中的份额；

将该 4 个可视化图表分别截图复制粘贴至 【提交结果.docx】中对应的任务序号下。

（二）任务二：业务分析

业务分析在汽车销售市场中至关重要，它可以帮助理解客户需求、市场趋势和产品定位。本任务中，我们将使用 Python 对汽车销售数据进行深入的业务分析，目的是识别市场的主要特征，并基于数据提出营销策略。

使用提供的汽车销售数据集，计算以下指标：

（1）**最受欢迎的车型**：根据年度销量，确定哪种车型最受欢迎；

（2）**销量与售价的关系**：分析售价与销量之间的关系，找出是否存在售价与销量的相关性；

（3）**市场细分分析**：根据销量和价格区间，识别不同市场细分（如经济型、中档型、豪华型）的表现；

根据上述分析结果，撰写一段简短的描述，提出至少两条针对汽车销售市场的营销策略建议。将内容复制粘贴至桌面【提交结果.docx】中对应的任务序号下。