ZZ052–大数据应用与服务赛项试题 03

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，为完成互联网酒店的大数据分析工作，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：Hadoop 完全分布式安装配置

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）从 Master 中的/opt/software 目录下将文件 hadoop-3.1.3.tar.gz 、 jdk-8u191-linux-x64.tar.gz 安装包解压到/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客户端桌面【Release\提交结果.docx】中对应的任务序号下;

（2）修改 Master 中/etc/profile 文件，设置 JDK 环境变量并使其生效，配置完毕后在 Master 节点分别执行“java -version”和“javac”命令，将命令行执行结果分别截图并粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下;

（3）请完成 host 相关配置，将三个节点分别命名为 master、slave1、slave2，并做免密登录，用 scp 命令并使用绝对路径从 Master 复制 JDK 解压后的安装文件到 slave1、slave2 节点（若路径不存在，则需新建），并配置 slave1、slave2 相关环境变量，将全部 scp 复制 JDK 的命令复制并粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下;

（4）在 Master 将 Hadoop 解压到/opt/module(若路径不

存在，则需新建)目录下，并将解压包分发至 slave1、slave2 中，其中 master、slave1、slave2 节点均作为 datanode，配置好相关环境，初始化 Hadoop 环境 namenode，将初始化命令及初始化结果截图（截取初始化结果日志最后 20 行即可）粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下;

（5）启动 Hadoop 集群（包括 hdfs 和 yarn），使用 jps 命令查看 Master 节点与 slave1 节点的 Java 进程，将 jps 命令与结果截图粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

2. 子任务二：Kafka 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）从 Master 中的/opt/software 目录下将文件 apache-zookeeper-3.5.7-bin.tar.gz 、 kafka_2.12-2.4.1.tgz 解压到/opt/module 目录下，将 Kafka 解压命令复制并粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下;

（2）配置好 zookeeper，其中 zookeeper 使用集群模式，分别将 master、slave1、slave2 作为其节点（若 zookpeer 已安装配置好，则无需再次配置），配置好 Kafka 的环境变量，使用 kafka-server-start.sh --version 查看 Kafka 的版本内容，并将命令和结果截图粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下;

（3）完善其他配置并分发 Kafka 文件到 slave1、slave2 中，并在每个节点启动 Kafka，创建 Topic，其中 Topic 名称为 installtopic，分区数为 2，副本数为 2，将创建命令和创建成果截图粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

3. 子任务三：Hive 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）从 Master 中的/opt/software 目录下将文件 apache-hive-3.1.2-bin.tar.gz、mysql-connector-java-5.1.37.jar 解压到/opt/module 目录下，将命令复制并粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下;

（2）设置 Hive 环境变量，并使环境变量生效，执行命令 hive --version 并将命令与结果截图粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下;

（3）完成相关配置并添加所依赖包，将 MySQL 数据库作为 Hive 元数据库。初始化 Hive 元数据，并通过 schematool 相关命令执行初始化，将初始化结果截图（范围为命令执行结束的最后 10 行）粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（二）任务二：数据库配置维护

1. 子任务一：数据库配置

（1）配置服务端 MySQL 数据库的远程连接。

（2）初始化 MySQL 数据库系统，将完整命令及初始化成功的截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（3）配置 root 用户允许任意 ip 连接，将完整命令截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（4）通过 root 用户登录 MySQL 数据库系统，查看 mysql 库下的所有表，将完整命令及执行命令后的结果的截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（5）输入命令以创建新的用户。完整命令及执行命令后的结果的截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（6）授予新用户访问数据的权限。完整命令及执行命令后的结果的截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（7）刷新权限。完整命令及执行命令后的结果的截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

2. 子任务二：创建相关表

（1）根据以下数据字段在 MySQL 数据库中创建酒店表（hotel）。酒店表字段如下：

字段	类型	中文含义	备注

Id	int	酒店编号	
hotel_name	varchar	酒店名称	
City	varchar	城市	
Province	varchar	省份	
Level	varchar	星级	
room_num	int	房间数	
Score	double	评分	
shopping	varchar	评论数	

（2）根据以下数据字段在 MySQL 数据库中创建评论表（comment）。评论表字段如下:

字段	类型	中文含义	备注
Id	int	评论编号	
Name	varchar	酒店名称	
Commentator	varchar	评论人	
Score	double	评分	
comment_time	datetime	评论时间	
Content	varchar	评论内容	

将这两个 SQL 建表语句分别截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

3. 子任务三：维护数据表

根据已给到的 sql 文件将这两份数据导入任意自己创建的数据库中，并对其中的数据进行如下操作:

在 comment_all 表中将 id 为 30 的评分改为 5;

在 hotel_all 表中统计各商圈的酒店总数。

将这两个 SQL 语句分别截图复制粘贴至客户端桌面【Release\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

有一份酒店详情列表数据: 省份、住宿场所名称、城市、商圈、是否为客栈、星级、房间数、评论数、评分、城市平均订单、城市平均间夜、城市平均实住订单、城市平均实住间夜、住宿场所订单、住宿场所总间夜、住宿场所实住订单、住宿场所实住间夜、住宿场所直销订单、住宿场所直销间夜、住宿场所直销实住间夜、住宿场所直销拒单、城市直销订单、城市实住订单、城市直销拒单率，并且存入到 hotel.csv 文件中。使用 pandas 读取 hotel.csv 并将读取的 csv 打印在 IDE 终端的截图复制粘贴至客户端桌面【Release\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

现已从相关网站及平台获取到原始数据集，为保障用户隐私和行业敏感信息，已进行数据脱敏。数据脱敏是指对某些敏感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。在涉及客户安全数据或者一些商业性敏感数据的情况、不违反系统规则条件下，对真实数据进行改造并提供测试使用，如身份证号、手机号等个人信息都需要进行数据脱敏。

相关数据文件中已经包含了数据采集阶段从企业消费平台网站上爬取的数据集，其中包含了来自不同城市的多家

住宿场所的销售信息，你的小组需要通过编写代码或脚本完成对相关数据文件中住宿场所销售管理数据的清洗和整理。

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 hotel.csv 中商圈为空的数据并且存入 hotel2_c1_N.csv,N 为删除的数据条数;

（2）删除 hotel.csv 中缺失值大于 3 个的数据列并且存入 hotel2_c2_N.csv,N 为删除的数据列变量名，多列时用下划线“_”间隔无顺序要求;

（3）将 hotel.csv 中评分为空的数据设置为 0 并且存入 hotel2_c3.csv;

（4）将 hotel.csv 中评分为空的数据设置为总平均评分并且存入 hotel2_c4_N.csv，N 为总平均评分保留一位小数。

将该 4 个文件名截一张图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

3. 子任务三：使用 Excel 进行数据清洗

现有一个信息化项目，项目小组分别针对一幢建筑的四、五两层进行了设备调研，并分工撰写了针对实验基础设施和网络计算设施两个大类的设备预算，设备预算按照楼层和类别分别存储在四个 Excel 文件里面。

你的小组需要通过 Excel 对这四个文件进行合并和处理。

具体要求如下:

（1）合并四个 Excel 文件到一个 Excel 文件中;

将四个 Excel 文件合并加载的截图（不用下拉）复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

通过列拆分出楼层和一级分类两个字段;

（2）将拆分两个字段的结果分别进行截图（不用下拉）复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（3）通过添加一列实现合计(万元)字段将合计数转化为万元单位;

将添加列的定义界面进行截图（不用下拉）复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（4）通过删除不需要进行统计的列和调整列顺序，四个 Excel 合并后的列顺序为：楼层、房间号、一级分类、二级分类、品名、单位、单价(元)、数量、合计(万元)。

将包含最终列的结果和非数据行计数界面进行截图（不用下拉）复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（二）任务二：数据标注

使用 SnowNLP 对酒店评论数据 hotel_comment.csv 进行标注，获取情感倾向评分（sentiments）,具体的标注规则如下:

（1）对情感倾向分数大于等于 0.6 评论数据标注为正向;

（2）对情感倾向分数大于 0.4 小于 0.6 评论数据为中性;

（3）对情感倾向分数小于等于 0.4 评论数据标注为负向。

根据采集到的评论信息，给出三类标注好的数据，存入 standard.csv。具体格式如下:

编号	酒店名称	评论信息	情感倾向	备注
1	全季酒店	XXXXXX	中性	

将 standard.csv 打开后直接截图（不用下拉）复制粘贴至客户端桌面【Release\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，查看目录命令截图粘贴至客户端桌面【Release\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，查看目录权限截图粘贴至客户端桌面【Release\提交结果.docx】中

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/root/下，结果截图粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

2. 子任务二：处理异常数据

user_info.csv 文件存储了电商互联网平台上收集的用户数据，数据中有以下内容:

id: 主键非空，bigint 类型，长度为 20

login_name: 用户名，varchar 类型，长度 200

nick_name: 用户昵称，varchar 类型，长度 200

passwd: 密码，varchar 类型，长度 200

name: 姓名，varchar 类型，长度 200

phone_num:手机号，varchar 类型，长度 200

email:邮箱，varchar 类型，长度 200

head_img:头像，varchar 类型，长度 200

user_level:用户级别，varchar 类型，长度 200

birthday:用户生日，date 类型，长度 0，格式为 YYYY-MM-DD

gender:性别，varchar 类型，长度 1

create_time:创建时间，datetime 类型，格式为 yyyy-MM-dd HH:mm:ss

operate_time:修改时间，datetime 类型，格式为 yyyy-MM-dd HH:mm:ss

编写 MapReduce 程序，实现以下功能：将 user_info.csv 数据的分隔符“，”转换为“|”，输出文件到 HDFS，然后在在控制台按顺序打印输出前 10 条数据，将结果截图粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

3. 子任务三：数据统计

order_info.csv 文件存储了电商互联网平台上收集的订单信息表数据，数据中有以下内容:

id：主键非空，bigint 类型，长度为 20

consignee：收货人，varchar 类型，长度 100

consignee_tel：收件人电话，varchar 类型，长度 20

final_total_amount：总金额，decimal 类型，长度 16

order_status：订单状态，varchar 类型，长度 20

user_id:用户 id，bigint 类型，长度 20

delivery_address:送货地址，varchar 类型，长度 1000

order_comment:订单备注，varchar 类型，长度 200

out_trade_no:订单交易编号（第三方支付用），varchar 类型，长度 50

trade_body:订单描述（第三方支付用），varchar 类型，长度 200

create_time:创建时间，datetime 类型，格式为 yyyy-MM-dd HH:mm:ss

operate_time:操作时间，datetime 类型，格式为 yyyy-MM-dd HH:mm:ss

expire_time:失效时间，datetime 类型，格式为 yyyy-MM-dd HH:mm:ss

tracking_no:物流单编号，varchar 类型，长度 100

parent_order_id:父订单编号，bigint 类型，长度 20

img_url:图片路径，varchar 类型，长度 200

province_id:省份 id，int 类型，长度 20

benefit_reduce_amount:优惠金额，decimal 类型，长度 16

original_total_amount:原价金额，decimal 类型，长度 16

feight_fee:运费，decimal 类型，长度 16

编写 MapReduce 程序，实现以下功能： 对于 order_status 这一字段统计每种状态的订单总数，将结果写入 HDFS，在控制台读取 HDFS 文件，将结果截图粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

四、模块三：业务数据分析与可视化

（一）任务一：数据分析与可视化

1.子任务一：使用 Python 进行数据分析和可视化

（1）数据分析

城市游客接纳能力是城市规划建设中的重要指标，其中城市的酒店房间数量是城市游客接纳能力的关键要素。请编写程序或脚本根据模块二任务一中子任务一所采集到的数据文件 hotel.csv 统计以下的相关信息，具体要求如下:

（1）分别统计各个商圈的的酒店总数，进行倒序排序展示前五名;

（2）统计各个商圈酒店的平均房间数，进行正序排序展示前五名;

（3）统计所有五星级酒店的平均评分。

将该 3 个统计结果在 IDE 的控制台中打印并分别截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（2）数据可视化

在企业消费平台上，各地区的酒店信息能够反映一个地区商业活动的密集程度。例如酒店总量多的城市大都具有强烈的吸纳外来人员的能力，订单数量能够反映该地区的有较多的商业往来。根据现有数据及给定参数完成酒店数据统计。

使用 Python 代码编写数据可视化的相关功能，所用数据为模块二任务一中子任务一所采集到的 hotel.csv 数据，具体

要求如下:

（1）用柱状图显示各个商圈的酒店总数;

（2）用折线图显示各星级酒店平均评分走势。

将该 2 个可视化图表分别截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

2. 子任务二：使用 Excel 进行数据分析和可视化

在模块二任务一中子任务三中处理好的数据作为待分析数据。

使用数据透视图表完成针对四、五层设备预算的统计分析，具体要求如下:

（1）使用表中一级分类、二级分类作为行统计项，楼层作为列统计项，合计(万元)做为统计量。

将数据透视表字段配置截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（2）将透视表按照合计(万元)的降序排列，一级分类以表格形式显示项目标签，汇总的合计(万元)数据使用保留小数点两位。

将数据透视表截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。

（3）与数据透视表搭配的数据透视图使用柱状图，不需要纵坐标轴和网格线，为柱状图每组数据添加数据标签，设置数据透视图的标题为“信息化项目统计图表 单位(万元)”

将数据透视图截图复制粘贴至客户端桌面【Release\提交

结果.docx】中对应的任务序号下。

（二）任务二：业务分析与方案设计

1. 子任务一：业务分析

完成模块二任务二已标注数据 standard.csv 评论情感分析功能，以月度为单位统计每月该酒店的正向、中性、负向评价数量，绘制折线图，并对酒店的发展趋势作出简要分析。将图表截图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下，在其下方编写发展趋势分析。

2. 子任务二：报表分析

根据模块二任务二已标注数据 standard.csv 文件中的结果，通过 Excel 生成报表信息方便酒店运营方在后续服务中进行优化，及时准确的把握用户体验，具体要求如下：

（1）该酒店的评论正向、负向、中性的评论趋势柱状图，按评论数量倒序排序；

（2）该酒店的整体评价趋势数量饼状图。

将两张图表截一张图复制粘贴至客户端桌面【Release\提交结果.docx】中对应的任务序号下。