ZZ052–大数据应用与服务赛项试题04

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、电商的大数据分析工作，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：Flume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-flume-1.9.0-bin.tar.gz解压到/opt/module目录下，将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）完善相关配置设置，配置Flume环境变量，并使环境变量生效，执行命令flume-ng 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）启动Flume传输Hadoop日志（namenode或datanode日志），查看HDFS中/tmp/flume目录下生成的内容，将查看命令及结果（至少5条结果）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：Flink on Yarn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flink-1.14.0-bin-scala_2.12.tgz解压到路径/opt/module中(若路径不存在，则需新建)，将完整解压命令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）修改容器中/etc/profile文件，设置Flink环境变量并使环境变量生效。在容器中/opt目录下运行命令flink --version，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）开启Hadoop集群，在yarn上以per job模式（即Job分离模式，不采用Session模式）运行$FLINK_HOME/examples/batch/WordCount.jar，将运行结果最后10行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。

示例 ：

flink run -m yarn-cluster -p 2 -yjm 2G -ytm 2G $FLINK_HOME/examples/batch/WordCount.jar

（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:

（1）在hotel_all表中删除id为25的酒店数据;

（2）在comment_all表中将id为30的评分改为5。

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

有一份购物平台列表数据：商品ID、名称、价格、浏览量、销量、库存，并且存入到shopping.csv文件中。使用pandas读取shopping.csv并将读取的csv打印在IDE终端的截图复制粘贴至【Release\提交结果.docx】中对应的任务序号下。

2.子任务二：数据处理

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。同时为了正确保护消费者权益，对于刷单或僵尸商户要进行及时监管，你的小组为此对数据中异常数据进行处理。

相关数据文件中已经包含了数据采集阶段从购物网站爬取的数据集，需要通过编写代码或脚本完成对相关数据文件的清洗和整理。

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shopping.csv中库存小于 10 或库存大于 10000 的数据，并存入shop1.csv;

（2）将涉及“刷单”、“捡漏”等字段的数据删除，并存入shop2.csv;

（3）将商品中涉及“女装”字段的数据删除，并存入shop3.csv;

（4）将shopping.csv中手机价格为区间数据的，设置为价格区间的平均数，存入shop4.csv。

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。

（二）任务二：数据标注

1.子任务一：分类标注

使用Python工具库SnowNLP对手机商城评论数据model_comment.csv进行标注，获取情感倾向评分（sentiments）,具体的对情感倾向的标注规则如下:

（1）对分数大于等于0.6的评论数据标注为正向;

（2）对分数大于0.4小于0.6的评论数据标注为中性;

（3）对分数小于等于0.4的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model_sen.csv。具体格式如下:

编号	手机品牌	评论信息	情感倾向	编号
1	华为 HUAWEI	XXXXXX	正向	1

将model_sen.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）任务二：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）将mobile.txt文件上传至 HDFS 新建目录/input/中,查看文件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）移动HDFS上mobile.txt文件到/user/hive/warehouse目录下，查看文件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：处理异常数值

mobile.txt文件存储了用户购买行为数据，数据中有以下内容：

字段名称	字段说明	数据类型	示例
Model	型号	string	华为荣耀4A
Title	标题	string	华为 荣耀4A 双卡双待4G手机 白色 移动4G版标配
comment	评论	string	给我叔叔买的 价格合理

			功能完善 用着还OK等过段时间再来评价
member_level	会员等级	string	金牌会员
from_platform	购买平台	string	京东PC客户端
area	地区	string	辽宁
user_impression	用户印象	string	国民手机 信号稳定 外观漂亮 照相不错
color	颜色	string	金色
price	价格	float	699
type	网络类型	string	移动4G版 标配
time	时间	string	2019/3/29 23:25

编写 MapReduce 程序，实现以下功能：清除数据中分隔符混乱的，多于11个字段的数据，输出文件到HDFS；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

mobile.txt文件存储了用户购买行为数据，数据中有以下内容:

字段名称	字段说明	数据类型	示例
model	型号	string	华为荣耀4A

title	标题	string	华为 荣耀4A 双卡双待 4G手机 白色 移动4G版 标配
comment	评论	string	给我叔叔买的 价格合理 功能完善 用着还OK等过段时间再来评价
member_level	会员等级	string	金牌会员
from_platform	购买平台	string	京东PC客户端
area	地区	string	辽宁
user_impression	用户印象	string	国民手机 信号稳定 外观漂亮 照相不错
color	颜色	string	金色
price	价格	float	699
type	网络类型	string	移动4G版 标配
time	时间	string	2019/3/29 23:25

编写MapReduce程序，实现以下功能：根据user_impression这一字段，统计买家对商家销售的手机商品的印象，结果按照印象数降序排序，格式为：(user_impression,次数)，如：(性价比高,10)，结果保存至HDFS，在控制台读取HDFS文件输出各组人数，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

品牌价值和商品特性对用户的购物习惯有着重要的影响，不同的商品特性能够满足消费者不同的需求和偏好，消费者往往也会根据自己对品牌的认知和评价以及对商品特性的需求进行选择和购买决策。请编写程序或脚本根据模块二任务一子任务一采集到的数据文件shopping.csv进行处理，要求对商品名称进行分割，第一个元素作为对应商品品牌，其他元素作为对应特征，统计以下的相关信息，具体要求如下:

（1）对各品牌进行统计，进行正序排序展示前十名;

（2）对各商品特征进行统计，进行正序排序前六名;

（3）统计各品牌的销量，进行正序排序展示前五名。

将该3个统计结果在PyCharm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.子任务二：数据可视化

在购物平台上，各地区的商品购物信息能够反映一不同区域对于不同产品需求成都。例如同品牌商品在不同区域，其销售量和热销产品线、产品价格能够反映该地区人群的购物习惯。根据现有数据及给定参数完成手机商城销

售数据统计。

使用Python可视化库Matplotlib编写数据可视化的相关功能，所用数据为模块一任务一子任务一所采集到的shopping.csv数据，具体要求如下:

（1）用柱状图显示不同价格区间手机销售情况，了解大众消费情况;

（2）用饼图显示不同地区手机品牌销售统计占比。

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。

（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model_sen.csv评论情感分析功能，以月度为单位统计每月某品牌商户的正向、中性、负向评价数量，绘制折线图，并对此品牌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，并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model_sen.csv文件中的结果，通过Excel 生成报表信息方便品牌商户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）该品牌商户的评论正向、负向、中性的评论趋势

柱状图，按评论数量倒序排序;

（2）该品牌商户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。