ZZ052-大数据应用与服务赛项试题 10

一、背景描述

当今时代，数据正在迅速膨胀并变大，一天之中，互联网产生的全部内容可以达到 EB 级别，能够轻松刻满 1.68 亿张光盘。在商业、经济及其它领域中，决策将日益基于数据和分析而作出，而并非基于经验和直觉。那么，要怎样基于大数据做出正确的决策呢？大数据首先需要解决的问题就是数据存储的问题，由于数据量非常之大，想通过传统单一的节点的存储显得力不从心，搭建分布式的文件存储系统成为了一个完美的解决方案。解决了数据存储的问题，我们需要从数据中提取有用信息，通过数据分析手段让数据发挥出真正的价值。但往往采集的原始数据中包含了一些无用数据以及噪声数据，如果直接基于这些脏数据进行分析，往往会让分析结果产生偏差甚至错误，从而造成决策上的失准。因此，我们有必要对这些原始数据进行清洗，以保证其数据准确性、完整性和可用性，提高数据的质量。在解决脏数据的困扰后，我们需要采取各种数据分析手段，提取数据中的价值，得到可靠的结果，并以图表等直观的方式将分析结果进行展现。然后从业务层面对分析结果进行分析和解释，从而指引我们做出正确的决策，真正获取“数据财富”。

1

气候变化正在迅速地改变地球。随着全球气温不断升高、海平面上升、极端天气事件频繁发生，人们对于地球的未来更加担忧。为了更好地了解气候变化的趋势、预测未来天气趋势，指引相关部门尽早做出举措以应对气候变化，保护人类赖以生存的家园，你的团队将运用大数据技术对天气数据进行分析及决策。搭建大数据平台集群环境以应对海量天气数据的存储，结合数据库的毫秒级的响应，为天气决策系统提供数据存储及查询保障。通过数据清洗技术，去除数据中的噪音，提高数据质量。通过数据标注技术，结合业务认知，对数据进行分类标注，为后续通过人工智能算法模型决策奠定基础。通过各种数据分析技术，让看似杂乱无章的数据，变得灵动，找出天气变化的内在规律。通过数据可视化技术，让数据分析结果及天气变化规律以一种最为直观的方式呈现。最后从业务层面对天气数据分析结果进行分析及解释，使气象学家更好的了解气候变化，并做出精准决策应对气候问题。你们作为该大数据小组的技术人员，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\任务 D 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\任务 D 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\任务 D 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

3

可）粘贴至客户端桌面【Release\任务 D 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\任务 D 提交结果.docx】中对应的任务序号下。

2.子任务二：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 安装包解压到/opt/module 目录下，将命令复制并粘贴至客户端桌面【Release\任务 D 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\任务 D 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\任务 D 提

交结果.docx】中对应的任务序号下。

3.子任务三：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 安装包解压到 /opt/module 目录下，将解压命令复制并粘贴至客户端桌面【Release\任务 D 提交结果.docx】中对应的任务序号下;

2、完善相关配置，设置 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\任务 D 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\任务 D 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。

5

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\任务 C 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\任务 C 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\任务 C 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\任务 C 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\任务 C 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\任务 C 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	datetim	评论时间	

7

_time	e		
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\任务 C 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\任务 C 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 1：数据获取

网站解析，利用 Chrome 查看网页源码，分析企业消费平台网站后台网页结构。

1、打开企业消费平台网站后台，在网页中右键点击检查，或者 F12 快捷键，查看元素页面;

2、检查网站：浏览网站源码查看所需内容。

从企业消费平台网站中爬取需要数据，按照要求使用 Python 语言编写爬虫代码，爬取指定数据项，并对结果数据

8

集进行数据探索、以及必要的数据处理操作。

具体步骤如下:

1）使用 Scrapy 框架创建爬虫项目

2）构建爬虫请求

3）按要求定义相关字段

4）获取有效数据

5）将爬取到的数据保存到指定位置

具体要求如下:

爬取酒店详情列表数据: 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。请将网页解析部分的代码截图复制粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对应的任务序号下。

2.子任务二：HDFS 文件上传下载

子本任务需要使用 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 Master 中的/root/目录下新建一个文件夹: result，

将创建文件夹命令与结果截图粘贴至客户端桌面【Release\任务 E 提交结果.docx】中对应的任务序号下;

2、使用 HDFS 命令，将 Master 下: /root 目录下新建的文件夹: result 上传到 HDFS 指定目录下: /根目录下; 并且使用 HDFS 命令查看目录; 将 HDFS 上传，查看命令截图粘贴至客户端桌面【Release\任务 E 提交结果.docx】中对应的任务序号下;

3、使用 HDFS 命令，将 HDFS 目录下的/result 文件夹下载到 Master 指定目录下:根目录下; 将下载文件夹命令与结果截图粘贴至客户端桌面【Release\任务 E 提交结果.docx】中对应的任务序号下。

3.子任务三：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\任务 A 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:

对情感倾向分数大于等于 0.6 的评论数据标注为正向;

对情感倾向分数大于 0.4 小于 0.6 的评论数据标注为中性;

对情感倾向分数小于等于 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

11

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\任务 A 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：处理异常值数据

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

12

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\任务 E 提交结果.docx】中对应的任务序号下。

2.子任务二：数据统计

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 Spark 程序，实现以下功能：对于 gender 这一字段统计电商消费人数男女数量，将结果写入 HDFS 中，格式为：(性别,人数)，如：(男,10)，在控制台读取 HDFS 文件输出男女各多少人，将结果截图粘贴至客户端桌面【Release\任务 E 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

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任务 A 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、统计各个商圈所有酒店的平均评分排名，进行倒序排序展示前五名;

3、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;

将该 3 个统计结果在 PyCharm 的控制台中打印并分别截

2.子任务二：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，所用数据为任务 A 所采集到的 hotel.csv 数据，具体要求如下:

用柱状图显示各个商圈的酒店总数;

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\任务 B 提交结果.docx】中对应的任务序号下。

（二）任务二：业务分析

1.子任务一：业务分析

完成任务 A 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\任务 B 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

根据任务 A 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\任务 B 提交结果.docx】中对应的任务序号下。