2022重庆市职业院校技能大赛

高职组

“大数据技术与应用”

赛项赛卷（3卷）

任

务

书

参赛队编号：

背景描述

## 大数据时代背景下，电商经营模式发生很大改变。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。商家可以对客户的消费行为信息数据进行收集和整理，比如消费者购买产品的花费、选择产品的渠道、偏好产品的类型、产品回购周期、购买产品的目的、消费者家庭背景、工作和生活环境、个人消费观和价值观等。通过数据追踪，知道顾客从哪儿来，是看了某网站投放的广告还是通过朋友推荐链接，是新访客还是老用户，喜欢浏览什么产品，购物车有无商品，是否清空，还有每一笔交易记录，精准锁定一定年龄、收入、对产品有兴趣的顾客，对顾客进行分组、标签化，通过不同标签组合运用，获得不同目标群体，以此开展精准推送。

因数据驱动的零售新时代已经到来，没有大数据，我们无法为消费者提供这些体验，为完成电商的大数据分析工作，你所在的小组将应用大数据技术，以Scala作为整个项目的基础开发语言，基于大数据平台综合利用Spark、Flink、Vue.js等技术，对数据进行处理、分析及可视化呈现，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。

## 模块A：大数据平台环境搭建（10分）

### 任务一：Hadoop 完全分布式安装配置

本环节需要使用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，具体部署要求如下:

1. 将Master节点JDK安装包解压并移动到/usr/java路径(若路径不存在，则需新建)，将命令复制并粘贴至对应报告中;
2. 修改/etc/profile文件，设置JDK环境变量，配置完毕后在master节点分别执行“java”和“javac”命令，将命令行执行结果分别截图并粘贴至对应报告中;
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，使用绝对路径从master复制JDK解压后的安装文件到slave1、slave2节点，并配置相关环境变量，将全部复制命令复制并粘贴至对应报告中;
4. 在Master将Hadoop解压到/opt目录下，并将解压包分发至slave1.slave2中,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复制粘贴至对应报告中；
5. 启动Hadoop集群，查看master节点jps进程，将查看结果复制粘贴至对应报告中。

### 任务二：Flink on Yarn 安装配置

本环节需要使用root用户完成相关配置，已安装Hadoop及需要配置前置环境，具体部署要求如下：

1. 将Flink包解压到路径/opt目录下，将完整命令复制粘贴至对应报告中;
2. 修改/etc/profile文件，设置Flink环境变量，并使环境变量生效将环境变量配置内容复制粘贴至对应报告中;
3. 开启Hadoop集群，在yarn上以per job模式运行 $FLINK\_HOME/examples/batch/WordCount.jar，将运行结果复制粘贴至对应报告中。

### 任务三：Spark on Yarn安装配置

本环节需要使用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、将scala包解压到/usr/路径，配置环境变量使其生效，将完整命令复制粘贴至对应报告中；

2、配置/etc/profile文件，设置Spark环境变量，并使环境变量生效将环境变量配置内容复制粘贴至对应报告中;

3、完成on yarn相关配置，使用spark on yarn 的模式提交$SPARK\_HOME/examples/jars/spark-examples\_2.11-2.1.1.jar 运行的主类为org.apache.spark.examples.SparkPi，将运行结果截图粘贴至对应报告中。

## 模块B：离线数据抽取 （15分）

### 任务一：全量数据抽取

编写Scala工程代码，将MySQL的shtd\_store库中表CUSTOMER、NATION、PART、PARTSUPP、REGION、SUPPLIER的数据全量抽取到Hive的ods库中对应表customer，nation，part，partsupp，region，supplier中。

1. 抽取shtd\_store库中CUSTOMER的全量数据进入Hive的ods库中表customer。字段排序，类型不变，同时添加静态分区，分区字段类型为String，且值为某日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.customer命令，将结果截图复制粘贴至对应报告中；
2. 抽取shtd\_store库中NATION的全量数据进入Hive的ods库中表nation。字段排序，类型不变，同时添加静态分区，分区字段类型为String，且值为某日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.nation命令，将结果截图复制粘贴至对应报告中；
3. 抽取shtd\_store库中PART的全量数据进入Hive的ods库中表part。字段排序，类型不变，同时添加静态分区，分区字段类型为String，且值为某日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.part命令，将结果截图复制粘贴至对应报告中；
4. 抽取shtd\_store库中PARTSUPP的全量数据进入Hive的ods库中表partsupp。字段排序，类型不变，同时添加静态分区，分区字段类型为String，且值为某日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.partsupp命令，将结果截图复制粘贴至对应报告中；
5. 抽取shtd\_store库中REGION的全量数据进入Hive的ods库中表region。字段排序，类型不变，同时添加静态分区，分区字段类型为String，且值为某日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.region命令，将结果截图复制粘贴至对应报告中；
6. 抽取shtd\_store库中SUPPLIER的全量数据进入Hive的ods库中表supplier。字段排序，类型不变，同时添加静态分区，分区字段类型为String，且值为某日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.supplier命令，将结果截图复制粘贴至对应报告中。

### 任务二：增量数据抽取

编写Scala工程代码，将MySQL的shtd\_store库中表ORDERS、LINEITEM增量抽取到Hive的ods库中对应表ORDERS，LINEITEM中。

1. 抽取shtd\_store库中ORDERS的增量数据进入Hive的ods库中表orders，要求只取XXXX年XX月XX日及之后的数据，根据ORDERS表中ORDERKEY作为增量字段，只将新增的数据抽入，字段类型不变，同时添加动态分区，分区字段类型为String，且值为ORDERDATE字段的内容。并在hive cli执行select count(distinct(dealdate)) from ods.orders命令，将结果截图复制粘贴至对应报告中；
2. 抽取shtd\_store库中LINEITEM的增量数据进入Hive的ods库中表lineitem，根据LINEITEM表中orderkey作为增量字段，只将新增的数据抽入，字段类型不变，同时添加静态分区，分区字段类型为String，且值为某日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.lineitem命令，将结果截图复制粘贴至对应报告中。

## 模块C：离线数据统计（20分）

### 任务一：数据清洗

编写Scala工程代码，将ods库中表customer，nation，region，supplier全量抽取到Hive的dwd库中对应表中。表中有涉及到timestamp类型的，均要求按照yyyy-MM-dd HH:mm:ss，不记录毫秒数，若原数据中只有年月日，则在时分秒的位置添加00:00:00，添加之后使其符合yyyy-MM-dd HH:mm:ss。

1. 将ods库中customer表数据抽取到dwd库中dim\_customer的分区表，分区字段为etldate且值与ods库的相对应表该值相等，并添加dwd\_insert\_user、dwd\_insert\_time、dwd\_modify\_user、dwd\_modify\_time四列,其中dwd\_insert\_user、dwd\_modify\_user均填写“user1”，dwd\_insert\_time、dwd\_modify\_time均填写当前操作时间，并进行数据类型转换。在hive cli中按照cust\_key顺序排序，查询dim\_customer前1条数据，将结果内容复制粘贴至对应报告中；
2. 将ods库中nation表数据抽取到dwd库中dim\_nation的分区表，分区字段为etldate且值与ods库的相对应表该值相等，并添加dwd\_insert\_user、dwd\_insert\_time、dwd\_modify\_user、dwd\_modify\_time四列，其中dwd\_insert\_user、dwd\_modify\_user均填写“user1”，dwd\_insert\_time、dwd\_modify\_time均填写当前操作时间，并进行数据类型转换。在hive cli中按照nation\_key顺序排序，查询dim\_nation前1条数据，将结果内容复制粘贴至对应报告中；
3. 将ods库中region表数据抽取到dwd库中dim\_region的分区表，分区字段为etldate且值与ods库的相对应表该值相等，并添加dwd\_insert\_user、dwd\_insert\_time、dwd\_modify\_user、dwd\_modify\_time四列，其中 dwd\_insert\_user、dwd\_modify\_user均填写“user1”，dwd\_insert\_time、dwd\_modify\_time均填写当前操作时间，并进行数据类型转换。在hive cli中按照region\_key顺序排序，查询dim\_region表前1条数据，将结果内容复制粘贴至对应报告中；
4. 将ods库中orders表数据抽取到dwd库中fact\_orders的分区表，分区字段为etldate且值与ods库的相对应表该值相等，并添加dwd\_insert\_user、dwd\_insert\_time、dwd\_modify\_user、dwd\_modify\_time四列，其中dwd\_insert\_user、dwd\_modify\_user均填写“user1”，dwd\_insert\_time、dwd\_modify\_time均填写当前操作时间，并进行数据类型转换。在执行hive cli执行select count(distinct(dealdate)) from dwd.fact\_orders命令，将结果内容复制粘贴至对应报告中；
5. 待任务4完成以后，需删除ods.orders中的分区，仅保留最近的三个分区。并在hive cli执行show partitions ods.orders命令，将结果截图粘贴至对应报告中；
6. 将ods库中lineitem表数据抽取到dwd库中fact\_lineitem的分区表，分区字段为etldate且值与ods库的相对应表该值相等，抽取的条件为根据orderkey和partkey进行去重，并添加dwd\_insert\_user、dwd\_insert\_time、dwd\_modify\_user、dwd\_modify\_time四列，其中dwd\_insert\_user、dwd\_modify\_user均填写“user1”，dwd\_insert\_time、dwd\_modify\_time均填写当前操作时间，并进行数据类型转换。在hive cli执行show partitions dwd.fact\_lineitem命令，将结果截图粘贴至对应报告中。

### 任务二：指标计算

1. 编写Scala工程代码，根据dwd层表统计每人每月下单的数量和下单的总金额，存入dws层的customer\_consumption\_day\_aggr表（表结构如下）中，然后在hive cli中按照cust\_key，totalconsumption, totalorder均为逆序排序，查询出前5条，将SQL语句与执行结果截图粘贴至对应报告中;

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 中文含义 | 备注 |
| cust\_key | int | 客户key |  |
| cust\_name | string | 客户名称 |  |
| totalconsumption | double | 消费总额 | 当月消费订单总额 |
| totalorder | int | 订单总数 | 当月订单总数 |
| year | int | 年 | 订单产生的年 |
| month | int | 月 | 订单产生的月 |

1. 编写Scala工程代码，根据dwd层表统计连续两个月下单的用户，订单发生时间限制为大于等于XXXX年之后的数据，存入MySQL数据库shtd\_store的usercontinueorder表中。然后在Linux的MySQL命令行中根据订单总数、消费总额、客户主键逆序排序，查询出前5条，将SQL语句与执行结果截图粘贴至对应报告中。

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 中文含义 | 备注 |
| custkey | int | 客户主键 |  |
| custname | text | 客户名称 |  |
| month | text | 月 | 记录当前月和下月，用下划线‘\_’相连  例如：202201\_202202  表示2022年1月到2月用户连续下单。 |
| totalconsumption | double | 消费总额 | 连续两月的订单总额 |
| totalorder | int | 订单总数 | 连续两月的订单总数 |

## 模块D：数据采集与实时计算（20分）

### 任务一：实时数据采集

1. 在Master节点使用Flume采集XXXX端口的socket数据，将数据存入到Kafka的自定义Topic中，将Flume的配置截图粘贴至对应报告中。

### 任务二：使用Flink处理Kafka中的数据

1. 编写Scala工程代码，使用Flink消费kafka中的数据，统计实时订单总额，将key设置成totalprice存入redis中。使用redis cli以get key方式获取totalprice值，将结果截图粘贴至对应报告中。

## 模块E：数据可视化（20分）

### 任务一：用条形图展示消费额最高的国家

编写Vue工程代码，根据接口，用条形图展示XXXX年消费额最高的5个国家，同时将用于图表展示的数据结构在浏览器的console中进行打印输出，将图表可视化结果和浏览器console打印结果分别截图并粘贴至对应报告中。

### 任务二：用饼状图展示各地区（即大洲）消费能力

编写Vue工程代码，根据接口，用饼状图展示XXXX年第一季度各大洲的消费总额占比，同时将用于图表展示的数据结构在浏览器的console中进行打印输出，将图表可视化结果和浏览器console打印结果分别截图并粘贴至对应报告中。

### 任务三：折线图展示总消费额变化

编写Vue工程代码，根据接口，用折线图展示XXXX年上半年商城总消费额的变化情况，同时将用于图表展示的数据结构在浏览器的console中进行打印输出，将图表可视化结果和浏览器console打印结果分别截图并粘贴至对应报告中。

## 模块F：综合分析（10分）

### 任务一：Hive分区表

Hive分区表的作用是什么？将内容编写至对应报告中。

### 任务二：简要描述任务过程中的问题并进行总结

将内容编写至对应报告中。