**2025年河北省职业院校学生技能大赛**

**“大数据应用开发”高职组**

**赛项样卷**

**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：大数据平台搭建（容器环境）（10分）

**环境说明：**

|  |
| --- |
| **服务端登录地址详见各模块服务端说明。**  **补充说明：**宿主机可通过Asbru工具或SSH客户端进行SSH访问；  所有模块中应用命令必须采用绝对路径；  进入Master节点的方式为  docker exec -it master /bin/bash  进入Slave1节点的方式为  docker exec -it slave1 /bin/bash  进入Slave2节点的方式为  docker exec -it slave2 /bin/bash  三个容器节点的root密码均为123456 |

### 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. 将容器Master节点JDK安装包解压并移动到/opt/module路径中(若路径不存在，则需新建)，将命令复制并粘贴至客户端桌面【Release\模块A提交结果.docx】中对应的任务序号下;
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 -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\模块A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，将全部复制命令复制并粘贴至客户端桌面【Release\模块A提交结果.docx】中对应的任务序号下;
4. 在容器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（截取初始化结果结束倒数20行即可）复制粘贴至客户端桌面【Release\模块A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命令与结果截图复制粘贴至客户端桌面【Release\模块A提交结果.docx】中对应的任务序号下。

### 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，需要配置前置环境，具体要求如下：

1. 从宿主机/opt目录下将文件apache-zookeeper-3.5.7-bin.tar.gz、kafka\_2.12-2.4.1.tgz复制到容器Master中的/opt/software路径中（若路径不存在，则需新建），将Master节点Zookeeper，Kafka安装包解压到/opt/module目录下，将Kafka解压命令复制并粘贴至客户端桌面【Release\任务A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ZooKeeper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\任务A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\任务A提交结果.docx】中对应的任务序号下。

### 任务三：ClickHouse单机部署

本环节需要使用root用户完成相关配置，具体要求如下：

1. 将容器Master节点ClickHouse相关安装包解压到/opt/module/clickhouse目录下（若路径不存在，则需新建），将全部解压命令复制并粘贴至客户端桌面【Release\模块A提交结果.docx】中对应的任务序号下；
2. 执行启动各个相关脚本，将全部启动命令复制并将执行结果（截取结果最后倒数15行即可）截图粘贴至客户端桌面【Release\模块A提交结果.docx】中对应的任务序号下；
3. 设置远程访问并移除默认监听文件（listen.xml），同时由于9000端口被Hadoop占用，需要将clickhouse的端口更改为9001，将上述要求的设置远程访问配置文件配置截图、端口更改后配置文件配置复制并粘贴至客户端桌面【Release\模块A提交结果.docx】中对应的任务序号下；
4. 启动clickhouse，启动后查看clickhouse运行状态，并将启动命令复制、查看运行状态命令复制并将执行结果截图粘贴至客户端桌面【Release\模块A提交结果.docx】中对应的任务序号下。

## 模块B：数据采集（15分）

**环境说明：**

|  |
| --- |
| **服务端登录地址详见各模块服务端说明。**  **补充说明：**各节点可通过Asbru工具或SSH客户端进行SSH访问；  主节点MySQL数据库用户名/密码：root/123456（已配置远程连接）；  Hive的配置文件位于主节点/opt/module/hive-3.1.2/conf/  Spark任务在Yarn上用Client运行，方便观察日志；  建议使用gson解析json数据。 |

### 任务一：离线数据采集

编写Scala工程代码，将MySQL的db库中表table1、table2、table3、table4、table5、table6、table7、table8、table9、table10、table11的数据增量抽取到Hive的ods库中对应表table1、table2、table3、table4、table5、table6、table7、table8、table9、table10、table11中。

1. 抽取db库中table1的增量数据进入Hive的ods库中表table1。根据ods.table1表中modified\_time作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用hive cli执行show partitions ods.table1命令，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下；
2. 抽取db库中table2的增量数据进入Hive的ods库中表table2。根据ods.table2表中modified\_time作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用hive cli执行show partitions ods.table2命令，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下；
3. 抽取db库中table3的增量数据进入Hive的ods库中表table3，根据ods.table3表中modified\_time作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用hive cli执行show partitions ods.table3命令，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下；
4. 抽取db库中table4的增量数据进入Hive的ods库中表table4，增量字段取ods.table4表中time1、time2中的最大者，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用Hive Cli查询最新分区数据总条数，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下；
5. 抽取db库中table5的增量数据进入Hive的ods库中表table5，根据ods.table5表中modified\_time作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用hive cli执行show partitions ods.table5命令，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下。
6. 抽取db库中table6的增量数据进入Hive的ods库中表table6，根据ods.table6表中modified\_time作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用hive cli执行show partitions ods.table6命令，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下；
7. 抽取db库中table7的增量数据进入Hive的ods库中表table7，根据ods.table7表中modified\_time作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用hive cli执行show partitions ods.table7命令，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下；
8. 抽取db库中table8的增量数据进入Hive的ods库中表table8，根据ods.table8表中modified\_time作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用hive cli执行show partitions ods.table8命令，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下；
9. 抽取db库中table9的增量数据进入Hive的ods库中表table9，根据ods.table9表中modified\_time作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用hive cli执行show partitions ods.table9命令，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下；
10. 抽取db库中table10的增量数据进入Hive的ods库中表table10，根据ods.table10表中modified\_time作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用hive cli执行show partitions ods.table10命令，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下；
11. 抽取db库中table11的增量数据进入Hive的ods库中表table11，根据ods.table11表中modified\_time作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。使用hive cli执行show partitions ods.table11命令，将执行结果截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下。

### 任务二：实时数据采集

1. 在主节点使用Flume采集实时数据生成器XXXXX端口的socket数据，将数据存入到Kafka的Topic中，使用Kafka自带的消费者消费Topic中的数据，查看Topic中的前1条数据的结果，将查看命令与结果完整的截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下；
2. 实时脚本启动后，在主节点进入到maxwell-1.29.0的解压后目录下，配置相关文件并启动，读取主节点MySQL数据的binlog日志到Kafka的Topic中。使用Kafka自带的消费者消费Topic中的数据，查看Topic中的前1条数据的结果，将查看命令与结果完整的截图粘贴至客户端桌面【Release\模块B提交结果.docx】中对应的任务序号下。

## 模块C：实时数据处理（25分）

**环境说明：**

|  |
| --- |
| **服务端登录地址详见各模块服务端说明。**  **补充说明：**各节点可通过Asbru工具或SSH客户端进行SSH访问；  主节点MySQL数据库用户名/密码：root/123456（已配置远程连接）；  Flink任务在Yarn上用per job模式（即Job分离模式，不采用Session模式），方便Yarn回收资源；  建议使用gson解析json数据。 |

### 任务一：实时数据清洗

编写Java代码，使用Flink消费Kafka中Topic的数据并进行相应的数据统计计算。

1. 使用Flink消费Kafka中topic为topic1的数据，根据数据中不同的表将数据分别分发至kafka的DWD层的Topic中。使用Kafka自带的消费者消费Topic的前1条数据，将结果截图粘贴至客户端桌面【Release\模块C提交结果.docx】中对应的任务序号下；
2. 使用Flink消费Kafka中topic为topic2的数据，将数据分别分发至kafka的DWD层的Topic中。使用Kafka自带的消费者消费Topic的前1条数据，将结果截图粘贴至客户端桌面【Release\模块C提交结果.docx】中对应的任务序号下;
3. 在任务1、2进行的同时，需要将相关数据表备份至HBase中，使用HBase Shell查看ods:table1表的任意2条数据，查看字段为key1与key2、查看ods:table2表的任意2条数据，查看字段为key1与key2、查看ods:table3表的任意2条数据，查看字段为key1与key2。将结果分别截图粘贴至客户端桌面【Release\模块C提交结果.docx】中对应的任务序号下。

### 任务二：实时指标计算

编写Java工程代码，使用Flink消费Kafka中dwd层的Topic数据,表结构与离线数据表结构相同。

1. 使用Flink消费kafka中的数据，统计商品的UV和PV，将结果写入HBase中的表ads：online\_uv\_pv中。使用Hive cli查询ads.pv\_uv\_result表，查询出10条数据,将结果截图粘贴至客户端桌面【Release\模块C提交结果.docx】中对应的任务序号下;
2. 使用Flink消费kafka中的数据，统计商城每分钟的GMV，将结果存入redis中，key为store\_gmv，使用redis cli以get key方式获取store\_gmv值，将每次截图粘贴至客户端桌面【Release\模块C提交结果.docx】中对应的任务序号下（每分钟查询一次，至少查询3次）。

## 模块D：离线数据处理（20分）

**环境说明：**

|  |
| --- |
| **服务端登录地址详见各模块服务端说明。**  **补充说明：**各节点可通过Asbru工具或SSH客户端进行SSH访问；  主节点MySQL数据库用户名/密码：root/123456（已配置远程连接）；  Hive的配置文件位于主节点/opt/module/hive-3.1.2/conf/  Spark任务在Yarn上用Client运行，方便观察日志；  ClickHouse的jdbc连接端口8123，用户名/密码：default/123456，命令行客户端（tcp）端口9001；  建议使用gson解析json数据。 |

### 任务一：离线数据清洗

编写Scala工程代码，将ods库中表table1、table2、table3、table4、table5、table6、table7、table8、table9、table10、table11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. 抽取ods库中表table1最新分区数据，并结合dim\_table1最新分区现有的数据，根据id合并数据到dwd库中dim\_table1的分区表，分区字段为etl\_date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”。若该条记录第一次进入数仓dwd层则dwd\_insert\_time、dwd\_modify\_time均存当前操作时间，并进行数据类型转换。若该数据在进入dwd层时发生了合并修改，则dwd\_insert\_time时间不变，dwd\_modify\_time存当前操作时间，其余列存最新的值。使用hive cli查询modified\_time为XXXX年XX月XX号当天的数据，并按照id进行升序排序，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
2. 抽取ods库中表table2最新分区数据，并结合dim\_table2最新分区现有的数据，根据id合并数据到dwd库中dim\_table2的分区表，分区字段为etl\_date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”。若该条记录第一次进入数仓dwd层则dwd\_insert\_time、dwd\_modify\_time均存当前操作时间，并进行数据类型转换。若该数据在进入dwd层时发生了合并修改，则dwd\_insert\_time时间不变，dwd\_modify\_time存当前操作时间，其余列存最新的值。使用hive cli执行show partitions dwd.dim\_table2命令，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
3. 抽取ods库中表table3最新分区的数据，并结合dim\_table3最新分区现有的数据，根据id合并数据到dwd库中dim\_table3的分区表，分区字段为etl\_date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”。若该条记录第一次进入数仓dwd层则dwd\_insert\_time、dwd\_modify\_time均存当前操作时间，并进行数据类型转换。若该数据在进入dwd层时发生了合并修改，则dwd\_insert\_time时间不变，dwd\_modify\_time存当前操作时间，其余列存最新的值。使用hive cli执行show partitions dwd.dim\_table3命令，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
4. 抽取ods库中表table4最新分区的数据，并结合HBase中table4\_offline表中的数据合并抽取到dwd库中fact\_table4的分区表，分区字段为etl\_date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（年月日必须是今天，时分秒只需在比赛时间范围内即可），抽取HBase中的数据时，只抽取XXXX年XX月XX日的数据（以rowkey为准），并进行数据类型转换。使用hive cli查询modified\_time为XXXX年XX月XX号当天的数据，并按照id进行升序排序，将结果截图复制粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
5. 抽取ods库中表table5表最新分区的数据，并结合HBase中table5\_offline表中的数据合并抽取到dwd库中fact\_table5的分区表，分区字段为etl\_date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（年月日必须是今天，时分秒只需在比赛时间范围内即可），抽取HBase中的数据时，只抽取XXXX年XX月XX日的数据（以rowkey为准），并进行数据类型转换。使用hive cli查询modified\_time为XXXX年XX月XX号当天的数据，并按照id进行升序排序，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
6. 抽取ods库中表table6最新分区的数据到dwd库中fact\_table6的分区表，分区字段为etl\_date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（年月日必须是今天，时分秒只需在比赛时间范围内即可），并进行数据类型转换。使用hive cli执行show partitions dwd.fact\_table6命令，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
7. 抽取ods库中表table7最新分区的数据到dwd库中table7的分区表，分区字段为etl\_date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（年月日必须是今天，时分秒只需在比赛时间范围内即可），并进行数据类型转换。使用hive cli执行show partitions dwd.table7命令，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
8. 抽取ods库中表table8最新分区的数据到dwd库中fact\_table8的分区表，分区字段为etl\_date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（年月日必须是今天，时分秒只需在比赛时间范围内即可），并进行数据类型转换。使用hive cli执行show partitions dwd.fact\_table8命令，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
9. 抽取ods库中表table9最新分区的数据，并结合HBase中table9\_offline表中的数据合并抽取到dwd库中table9的分区表，分区字段为etl\_date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（年月日必须是今天，时分秒只需在比赛时间范围内即可），抽取HBase中的数据时，只抽取XXXX年XX月XX日的数据（以rowkey为准），并进行数据类型转换。使用hive cli查询modified\_time为XXXX年XX月XX号当天的数据，并按照id进行升序排序，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
10. 抽取ods库中表table10最新分区的数据到dwd库中dim\_table10的分区表，分区字段为etl\_date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（年月日必须是今天，时分秒只需在比赛时间范围内即可），并进行数据类型转换。使用hive cli执行show partitions dwd.dim\_table10命令，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
11. 抽取ods库中表table11最新分区的数据到dwd库中dim\_table11的分区表，分区字段为etl\_date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（年月日必须是今天，时分秒只需在比赛时间范围内即可），并进行数据类型转换。使用hive cli执行show partitions dwd.dim\_table11命令，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
12. 将dwd库中表的数据关联到dws库中table12的分区表，分区字段为etl\_date，类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd），并添加dws\_insert\_user、dws\_insert\_time、dws\_modify\_user、dws\_modify\_time四列，其中dws\_insert\_user、dws\_modify\_user均填写“user1”，dws\_insert\_time、dws\_modify\_time均填写当前操作时间（年月日必须是今天，时分秒只需在比赛时间范围内即可），并进行数据类型转换。使用hive cli统计最新分区中得数据总量，将结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下。

### 任务二：离线指标计算

1. 编写Scala工程代码，根据dwd的订单表，求各省份下单时间为XXXX年的支付转化率，并将计算结果写入clickhouse的ds\_result库的表。在Linux的clickhouse命令行中根据ranking字段查询出转化率前三的省份，将SQL语句与执行结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
2. 编写Scala工程代码，根据dwd的表，计算所有订单中各商品所有订单总销售金额排名，并将计算结果写入clickhouse的ds\_result库的表。然后在Linux的clickhouse命令行中根据sales\_rank升序查询前5行，将SQL语句与执行结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下；
3. 编写Scala工程代码，根据dwd的登录日志表求最近连续三周登录的用户数，并将计算结果写入clickhouse的ds\_result库的表。然后在Linux的clickhouse命令行中根据[active](javascript:;)\_total降序查询，将SQL语句与执行结果截图粘贴至客户端桌面【Release\模块D提交结果.docx】中对应的任务序号下。

## 模块E：数据可视化（15分）

**环境说明：**

|  |
| --- |
| **数据接口地址及接口描述详见各模块服务端说明。** |

### 任务一：用柱状图展示消费额最高的省份

编写Vue工程代码，根据接口，用柱状图展示XXXX年消费额最高的5个省份，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客户端桌面【Release\模块E提交结果.docx】中对应的任务序号下。

### 任务二：用饼状图展示各地区消费能力

编写Vue工程代码，根据接口，用饼状图展示XXXX年各地区的消费总额占比，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客户端桌面【Release\模块E提交结果.docx】中对应的任务序号下。

### 任务三：用折柱混合图展示省份平均消费额和地区平均消费额

编写Vue工程代码，根据接口，用折柱混合图展示XXXX年各省份平均消费额（四舍五入保留两位小数）和地区平均消费额（四舍五入保留两位小数）的对比情况，柱状图展示平均消费额最高的5个省份，折线图展示这5个省所在的地区的平均消费额变化，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客户端桌面【Release\任务E提交结果.docx】中对应的任务序号下。

## 模块F：综合分析（10分）

### 任务一：XXXXXXXXXXXXXXXXXX，原因是什么？

将内容编写至客户端桌面【Release\模块F提交结果.docx】中对应的任务序号下。

### 任务二：XXXXXXX有哪些XXXXXXXX？列举并简要描述。

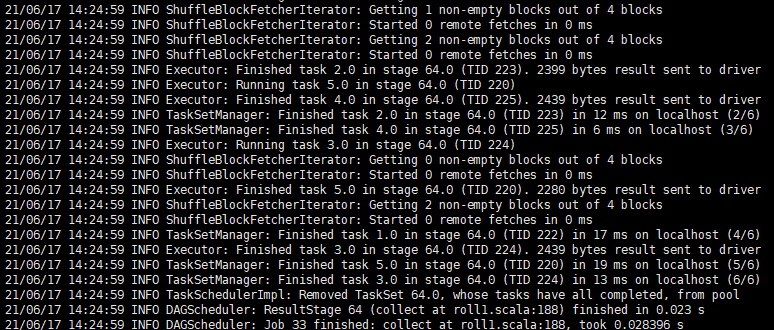
将内容编写至客户端桌面【Release\模块F提交结果.docx】中对应的任务序号下。

### 任务三：简要描述竞赛过程中的问题并进行总结。

将内容编写至客户端桌面【Release\模块F提交结果.docx】中对应的任务序号下。

## 附录：补充说明

### 命令行截图样例（保证清晰）



### 表结构说明

MySQL数据库中已自带注释，自行连接使用工具查看。

### 若IDEA运行代码时候出现

scalac: No ‘scala-library\*.jar’ in Scala compiler classpath in Scala SDK Maven: org.scala-lang:scala-library:2.12.0

则需要在Flie->Project Structure->Project Settings->Libraries->添加scala的包（2.12大版本一致即可）。