2022年全国职业院校技能大赛

高职组

“大数据技术与应用”

赛项赛卷（10卷）

任

务

书

参赛队编号：

**背景描述**

工业互联网是工业全要素、全产业链、全价值链的全面连接，是人、机、物、工厂互联互通的新型工业生产制造服务体系，是互联网从消费领域向生产领域、从虚拟经济向实体经济拓展的核心载体，是建设现代化经济体系、实现高质量发展和塑造全球产业竞争力的关键支撑。党中央、国务院高度重视工业互联网发展，习近平总书记连续四年对推动工业互联网发展做出重要指示。加快发展工业互联网产业，不仅是各国顺应产业发展大势，抢占产业未来制高点的战略选择，也是我国推动制造业质量变革、效率变革和动力变革，实现高质量发展的客观要求。

为完成工业大数据分析工作，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 Hive、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）（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：**宿主机可通过Asbru工具或SSH客户端进行SSH访问；  相关软件安装包在宿主机的/opt目录下，请选择对应的安装包进行安装，用不到的可忽略；  所有模块中应用命令必须采用绝对路径；  从本地仓库中拉取镜像，并启动3个容器  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  同时将/opt目录下的所有安装包移动到3个容器节点中。 |

### 任务一：Hadoop 伪分布式安装配置

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/root/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，将节点命名为master，并做免密登录，将全部命令复制并粘贴至对应报告中;
4. 修改Hadoop环境变量，并将（/etc/profile）配置文件内容复制粘贴至对应报告中；
5. 在主机Master上将Hadoop解压到/opt目录下，配置好相关文件（并要求namenode的web端访问端口配置为15887，配置指定MapReduce程序运行在Yarn上，指定ResourceManager的地址为伪分布式部署所在主机），将全部命令以及配置文件内容复制并粘贴至对应报告中；
6. 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；
7. 启动Hadoop，查看节点jps进程，将查看结果复制粘贴至对应报告中。

### 任务二：Flume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. 设置Flume环境变量，并使环境变量只对当前root用户生效，将变量内容复制并粘贴至对应报告中；
2. 修改并配置flume-env.sh文件，将修改内容复制并粘贴至对应报告中；
3. 复制Flume启动的配置文件flume-conf.properties.template并命名为conf-file，修改其中内容，监控Hadoop的namenode日志，传输到HDFS的是/tmp/flume目录，将conf-file中的修改内容复制并粘贴至对应报告中；
4. 启动Flume传输Hadoop日志，查看HDFS中/tmp/flume目录下生成的文件，将查看命令及结果（至少10条结果）复制并粘贴至对应报告中。

### 任务三：Hive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Master节点Hive安装包解压到/opt目录下，将完整命令复制粘贴至对应报告中；
2. 把解压后的apache-hive-1.2.2-bin文件夹更名为hive；进入hive文件夹使用ls进行查看，将完整命令复制粘贴至对应报告中；
3. 设置Hive环境变量，并使环境变量只对当前root用户生效；并将环境变量配置内容复制并粘贴至对应报告中；
4. 将Hive安装目录里hive-default.xml.template文件更名为hive-site.xml；将完整命令复制粘贴至对应报告中；
5. 修改hive-site.xml配置文件，将MySQL数据库作为Hive元数据库。将配置文件中配置Hive元存储的相关内容复制并粘贴至对应报告中；
6. 初始化Hive元数据，将MySQL数据库JDBC驱动拷贝到Hive安装目录的lib文件夹下；并通过schematool相关命令执行初始化，将初始化结果复制粘贴至对应报告中；
7. 完善其他配置并启动Hive，将命令行输出结果复制粘贴至对应报告中。

## 模块B：离线数据处理（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：**各主机可通过Asbru工具或SSH客户端进行SSH访问；  Master节点MySQL数据库用户名/密码：root/123456（已配置远程连接）；  Hive的元数据启动命令为:  nohup hive --service metastore &  Hive的配置文件位于/opt/apache-hive-2.3.4-bin/conf/  Spark任务在Yarn上用Client运行，方便观察日志。 |

### 任务一：数据抽取

编写Sqoop脚本，将MySQL库中表EnvironmentData，ChangeRecord，Basemachine，MachineData， ProduceRecord，BaseEmployee全量抽取到Hive的ods库中对应表EnvironmentData，ChangeRecord，Basemachine， MachineData， ProduceRecord，BaseEmployee中。

1. 抽取MySQL的shtd\_store库中EnvironmentData表的全量数据进入Hive的ods库中表EnvironmentData，字段排序、类型不变，同时添加静态分区，分区字段类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.EnvironmentData命令，然后查询展示5行数据，将Sqoop提交命令及hive cli的执行结果分别截图复制粘贴至对应报告中；
2. 抽取MySQL的shtd\_store库中ChangeRecord表的全量数据进入Hive的ods库中表ChangeRecord，字段排序、类型不变，同时添加静态分区，分区字段类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.ChangeRecord命令，然后查询展示5行数据，将Sqoop提交命令及hive cli的执行结果分别截图复制粘贴至对应报告中；
3. 抽取MySQL的shtd\_store库中Basemachine表的全量数据进入Hive的ods库中表Basemachine，字段排序、类型不变，同时添加静态分区，分区字段类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.Basemachine命令，然后查询展示5行数据，将Sqoop提交命令及hive cli的执行结果分别截图复制粘贴至对应报告中；
4. 抽取MySQL的shtd\_store库中ProduceRecord表的全量数据进入Hive的ods库中表ProduceRecord，字段排序、类型不变，同时添加静态分区，分区字段类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.ProduceRecord命令，然后查询展示5行数据，将Sqoop提交命令及hive cli的执行结果分别截图复制粘贴至对应报告中；
5. 抽取MySQL的shtd\_store库中MachineData表的全量数据进入Hive的ods库中表MachineData，字段排序、类型不变，同时添加静态分区，分区字段类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.ProduceRecord命令，然后查询展示5行数据，将Sqoop提交命令及hive cli的执行结果分别截图复制粘贴至对应报告中；
6. 抽取MySQL的shtd\_store库中BaseEmployee表的全量数据进入Hive的ods库中表BaseEmployee，字段排序、类型不变，同时添加静态分区，分区字段类型为String，且值为当前比赛日的前一天日期（分区字段格式为yyyyMMdd）。并在hive cli执行show partitions ods.ProduceRecord命令，然后查询展示5行数据，将Sqoop提交命令及hive cli的执行结果分别截图复制粘贴至对应报告中。

### 任务二：数据清洗

编写Scala工程代码，将ods库中相应表数据全量抽取到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. 抽取ods库中EnvironmentData的全量数据进入Hive的dwd库中表fact\_environment\_data，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均填写操作时间，并进行数据类型转换。并在hive cli执行desc dwd.fact\_environment\_data命令，将结果内容复制粘贴至对应报告中；
2. 抽取ods库中ChangeRecord的全量数据进入Hive的dwd库中表fact\_change\_record，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均填写操作时间，并进行数据类型转换。并在hive cli执行desc dwd.fact\_change\_record命令，将结果内容复制粘贴至对应报告中；
3. 抽取ods库中Basemachine的全量数据进入Hive的dwd库中表dim\_machine,抽取数据之前需要对数据根据BaseMachineID进行去重处理。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均填写操作时间，并进行数据类型转换。在hive cli中按照Base\_machine顺序排序，查询dim\_machine前2条数据，将结果内容复制粘贴至对应报告中；
4. 抽取ods库中ProduceRecord的全量数据进入Hive的dwd库中表fact\_produce\_record,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均填写操作时间，并进行数据类型转换。在hive cli中按照produce\_machine\_id顺序排序，查询fact\_produce\_record前2条数据，将结果内容复制粘贴至对应报告中；
5. （该表要求使用Java编写MR代码实现）抽取ods库中MachineData的全量数据进入Hive的dwd库中表fact\_machine\_data。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均填写操作时间，并进行数据类型转换。并在hive cli执行show partitions dwd. fact\_machine\_data命令，将结果内容和MR的Diver类代码复制粘贴至对应报告中；
6. （该表要求使用Java编写MR代码实现）抽取ods库中BaseEmployee的全量数据进入Hive的dwd库中表dim\_employee。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均填写操作时间，并进行数据类型转换。并在hive cli执行show partitions dwd. fact\_machine\_data命令，将结果内容和MR的Diver类代码复制粘贴至对应报告中。

### 任务三：指标计算

1. 编写Hive SQL代码，根据dwd层dwd.fact\_machine\_data表联合dwd.dim\_machine表统计出每个车间、每台设备、每天状态为“运行”的时长，将统计结果最长的两台设备信息写入DWS层的dws.machine\_data\_top2\_day\_aggr表（表结构如下）中，然后在Linux的MySQL命令行中根据车间号倒序查询5条数据，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 |
| day | string | 日期 | 如：2021-10-01 |
| firstid | int | 第一 | 当日运行总时长第一 |
| secondid | int | 第二 | 当日运行总时长第二 |
| firsttime | int | 第一的时长 | 秒 |
| secondtime | int | 第二的时长 | 秒 |
| machinefactory | int | 车间号 | 统计所在车间号 |

1. 编写Hive SQL代码，根据DWS层的dws.machine\_data\_top2\_day\_aggr表联合dwd.dim\_employee表、dwd.dim\_machine表，统计每个月，各设备负责员（MachineResponsEmpID）所负责的设备中运行时长最长的设备，将计算结果存入MySQL数据库shtd\_store的emplongestmachine表（表结构如下）中，然后在Linux的MySQL命令行中根据月份、设备负责人员、设备id倒序排序，查询出前5条数据，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 |
| machineresponsempid | int | 设备负责人员 |  |
| machineid | int | 设备id | 负责人员下运行最长设备 |
| month | int | 月 | 数据统计的月 |
| year | int | 年 | 数据统计的年 |

1. 编写Scala工程代码，根据dwd层dwd.fact\_environment\_data表，统计检测设备（BaseID）每日的PM10的检测平均浓度，然后将每个设备的平均浓度与厂内所有检测设备每日检测结果的平均浓度做比较（结果值为：高/低/相同）存入MySQL数据库shtd\_store的表MachineRunningCompare（表结构如下）中，然后在Linux的MySQL命令行中根据检测设备ID倒序排序，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 |
| BaseID | int | 检测设备ID |  |
| MachineAvg | decimal | 单设备检测平均值 |  |
| FactoryAvg | decimal | 厂内所有设备平均值 |  |
| Comparison | string | 比较结果 | 高/低/相同 |
| EnvDate | string | 检测日期 | 如：2021-12-12 |

## 模块C：数据挖掘（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：**各主机可通过Asbru工具或SSH客户端进行SSH访问；  Master节点MySQL数据库用户名/密码：root/123456（已配置远程连接）；  Hive的元数据启动命令为:  nohup hive --service metastore &  Hive的配置文件位于/opt/apache-hive-2.3.4-bin/conf/  Spark任务在Yarn上用Client运行，方便观察日志。  该模块均使用Scala编写，利用Spark相关库完成。 |

### 任务一：特征工程

1. 根据dwd库中fact\_machine\_data表，根据以下要求转换： 获取最大分区的数据后，首先解析列get\_xmldata中的数据（数据格式为xml，采用dom4j解析，会给出解析demo），并获取 主轴转速，主轴倍率，主轴负载，进给倍率，进给速度，PMC程序号，循环时间，运行时间，有效轴数，总加工个数，已使用内存，未使用内存，可用程序量，注册程序量等相关的值，同时转换machine\_record\_stat字段的值，若值为报警，则填写1，否则填写0，以下为表结构，将数据保存在dwd.fact\_machine\_learning\_data，在hive cli中按照machine\_record\_id顺序排序，查询dwd.fact\_machine\_learning\_data前1条数据，将结果内容复制粘贴至对应报告中。

dwd.fact\_machine\_learning\_data表结构：

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 中文含义 | 备注 |
| machine\_record\_id | int | 自增长id |  |
| machine\_id | double | 机器id |  |
| machine\_record\_mainshaft\_speed | double | 主轴转速 | 默认0.0 |
| machine\_record\_mainshaft\_multiplerate | double | 主轴倍率 | 默认0.0 |
| machine\_record\_mainshaft\_load | double | 主轴负载 | 默认0.0 |
| machine\_record\_feed\_speed | double | 进给倍率 | 默认0.0 |
| machine\_record\_feed\_multiplerate | double | 进给速度 | 默认0.0 |
| machine\_record\_pmc\_code | double | PMC程序号 | 默认0.0 |
| machine\_record\_cricle\_time | double | 循环时间 | 默认0.0 |
| machine\_record\_run\_time | double | 运行时间 | 默认0.0 |
| machine\_record\_efficive\_shaft | double | 有效轴数 | 默认0.0 |
| machine\_record\_amount\_process | double | 总加工个数 | 默认0.0 |
| machine\_record\_use\_memory | double | 已使用内存 | 默认0.0 |
| machine\_record\_free\_memory | double | 未使用内存 | 默认0.0 |
| machine\_record\_amount\_use\_code | double | 可用程序量 | 默认0.0 |
| machine\_record\_amount\_free\_code | double | 注册程序量 | 默认0.0 |
| machine\_record\_state | double | 机器状态 | 默认0.0 |

### 任务二：报警预测

1. 根据任务一的结果，建立随机森林（随机森林相关参数考生可自定义，不做限制），使用任务一的结果训练随机森林模型，然后再将hive中dwd.fact\_machine\_learning\_data\_test（表结构与dwd.fact\_machine\_learning\_data一致，但machine\_record\_state列值为空）转成向量，预测其是否报警将结果输出到mysql库shtd\_industry中的ml\_result表中。在Linux的MySQL命令行中查询出所有数据并按照machine\_record\_id顺序排序，将SQL语句与执行结果截图粘贴至对应报告中。

ml\_result表结构：

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 中文含义 | 备注 |
| machine\_record\_id | int | 自增长id |  |
| machine\_record\_state | double | 机器状态 | 报警为1，其他状态则为0 |

## 模块D：数据采集与实时计算（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：**各主机可通过Asbru工具或SSH客户端进行SSH访问；  请先检查ZooKeeper、Kafka、Redis端口看是否已启动，若未启动则各启动命令如下：  ZK启动（netstat -ntlp查看2181端口是否打开）  /usr/zk/zookeeper-3.4.6/bin/zkServer.sh start  Redis启动（netstat -ntlp查看6379端口是否打开）  /usr/redis/bin/redis-server /usr/redis/bin/redis.conf  Kafka启动（netstat -ntlp查看9092端口是否打开）  /opt/kafka/kafka\_2.11-2.0.0/bin/kafka-server-start.sh -daemon（空格连接下一行）/opt/kafka/kafka\_2.11-2.0.0/config/server.properties  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。 |

### 任务一：实时数据采集

1. 在Master节点使用Flume采集/data\_log目录下实时日志文件中的数据，将数据存入到Kafka的Topic中（topic名称分别为ChangeRecord和EnvironmentData，分区数为4），将Flume的配置截图粘贴至对应报告中；
2. Flume接收数据注入kafka 的同时，将数据备份到HDFS目录/user/test/flumebackup下，将备份结果截图粘贴至对应报告中。

### 任务二：使用Flink处理Kafka中的数据

编写Scala工程代码，使用Flink消费Kafka中Topic为ChangeRecord及EnvironmentData的数据并进行相应的数据统计计算。

1. 使用Flink消费Kafka中EnvironmentData主题的数据,监控各环境检测设备数据，当温度（Temperature字段）持续10分钟高于39度时记录为预警数据（设备id，温度，预警时间），将该环境检测设备的预警数据写入Mysql表envtemperaturealarm中，然后在Linux的MySQL命令行中根据MachineID逆序排序，查询出前3条，将SQL语句与执行结果截图粘贴至对应报告中;
2. 使用Flink消费Kafka中ChangeRecord主题的数据，每隔10分钟输出最近2小时内各车间运行时间最长的前两台设备，数据输出到MySQL表runlongestmachine表（表结构如下）中，然后在Linux的MySQL命令行中根据车间号正序查询前5条数据，此查询操作共做两次，间隔大于10分钟，将SQL语句与两次的执行结果截图粘贴至对应报告中；

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 中文含义 | 备注 |
| MachineFactory | int | 车间号 |  |
| Outtime | varchar | 毫秒数时间戳 | 计算数据输出时的时间 |
| Longestmachine | int | 最长设备 | 运行时间最长的设备号 |
| Secondmachine | int | 次长设备 |  |

1. 使用Flink消费Kafka中ChangeRecord主题的数据，统计每个设备状态信息，当某设备的状态在24小时内连续三小时为“预警”状态没有改变，则将该设备的状态数据写入MySql表recordstatealarm中，然后在Linux的MySQL命令行中根据MachineID逆序排序，查询出前3条，将SQL语句与执行结果截图粘贴至对应报告中。

## 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：用饼状图展示某设备当天各状态时间

编写Vue工程代码，根据接口，用饼状图展示某日该设备的各个状态的运行时长占比，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。

### 任务二：用柱状图展示各车间设备每月平均运行时长

编写Vue工程代码，根据接口，用柱状图展示各车间设备每月平均运行时长，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：用折线图展示机器运行时长

编写Vue工程代码，根据接口，用折线图展示某机器在某一周的运行时长，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。

### 任务四：用散点图展示PM2.5浓度变化

编写Vue工程代码，根据接口，用散点图展示设备所处环境的PM2.5浓度的变化，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对应报告中。

### 任务五：用条形图展示各机器每日产量

编写Vue工程代码，根据接口，用柱状图展示某机器在某天机器的产量，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。

### 任务六：用折柱混合图展示机器日均产量和车间平均产量

编写Vue工程代码，根据接口，用折柱状图展示某机器在某月的日均产量和车间的月均产量，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：HDFS存储小文件问题

在HDFS数据存储是常常遇到小文件问题，小文件带来的危害有哪些？你知道的解决小文件问题的方法有哪些？将内容编写至对应报告中。

### 任务二：如何解决数据倾斜问题

当使用Spark产生了数据倾斜的情况时,请问有哪些方法可以解决这个问题呢？将内容编写至对应报告中。

### 任务三：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。