ZZ052–大数据应用与服务赛项试题 02

一、背景描述

随着互联网、大数据等技术的高速发展，通用设备制造业在“中国制造 2025”计划的推动下正向定制化服务转型。传统的设备销售模式正在向销售服务模式转变，这为企业带来了新的机遇和挑战。在这个转型过程中，商业模式的创新变得至关重要。信息化与现代服务的结合成为制造企业转型和管理升级的重要手段。

从管理角度来看，企业需要全局掌握已售出设备的整体运行状况，以提高服务效率、满意度和及时率。同时，企业还需要提升决策效率，降低服务成本。这些挑战可以通过大数据综合开发来解决，通过对设备数据进行采集、存储和分析，企业可以实现对设备运行状况的全面监控和管理。利用大数据分析与应用服务，可以优化服务调度和资源分配，提高服务效率和满意度。同时，通过数据分析和决策支持系统，可以提升企业的决策效率，并降低服务成本。

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：基础环境准备

（1)对三台环境更新主机名,配置hosts文件,以node01作为时钟源并进行时间同步;

（2）执行命令生成公钥、私钥，实现三台机器间的免秘登陆;

（3）从宿主机/root 目录下将文件 jdk-8u212-linux-x64.tar.gz 复制到容器node01 中的/root/software 路径中（若路径不存在，则需新建），将node01 节点JDK 安装包解压到/root/software 路径中(若路径不存在，则需新建);

（4）修改容器中/etc/profile 文件，设置 JDK 环境变量并使其生效，配置完毕后在node01 节点分别执行“java -version”和“javac”命令

2. 子任务二：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）在node01 将Hadoop 解压到/root/software(若路径不存在，则需新建)目录下，并将解压包分发至 node02、node03 中，其中三个节点节点均作为datanode，配置好相关环境，初始化Hadoop 环境namenode;

（2）开启集群，查看各节点进程。

3. 子任务三：Hive 安装配置

（1）从宿主机/root 目录下将文件 apache-hive-3.1.2-bin.tar.gz、mysql-connector-java-5.1.37.jar 复制到容器 node03 中的/root/software 路径中（若路径不存在，则需新建），将 node03 节点 Hive 安装包解压到/root/software 目录下；

（2）设置 Hive 环境变量，并使环境变量生效，执行命令 hive --version 查看版本信息；

（3）修改相关配置，添加依赖包，将 MySQL 数据库作为 Hive 元数据库，初始化 Hive 元数据。

4. 子任务四：Flume 安装配置

（1）从宿主机/root 目录下将文件 apache-flume-1.11.0-bin.tar.gz 复制到容器 node03 中的/root/software 路径中（若路径不存在，则需新建），将 node03 节点 Flume 安装包解压到/root/software 目录下；

（2）完善相关配置，配置 Flume 环境变量，并使环境变量生效，执行命令 flume-ng version。

（二）任务二：数据库配置维护

1. 子任务一：数据库配置

（1）在主机 node3 上安装 mysql-community-server，启动 mySQL 服务，根据临时密码进入数据库，并修改本地密码为“123456”；

（2）开启 MySQL 远程连接权限，所有 root 用户都可以使用 123456 进行登录连接。

2. 子任务二：导入相关表

（1）将本地/root/eduhq/equipment/目录下的数据文件 root_sl_src.sql 导入 MySQL 对应数据库 root_sl_src;

（2）将本地/root/eduhq/equipment/目录下的数据文件 root_sl_ugoogds_src.sql 导入 MySQL 对应数据库 root_sl_ugoogds_src。

3. 子任务三：维护数据表

结合已导入的两份 sql 数据，对其中的数据进行如下查询和操作。

（1）对‘root_sl_src’数据库中的‘province’数据表进行修改，修改字段 province_id 为 24 的记录的 province_name，修改为‘内蒙古自治区’；

（2）对‘root_sl_src’数据库中的‘city’数据表进行删除，删除字段 city_id 为 142 的记录。

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

（1） 使用load命令将提供的数据导入到Hive，全部数据表如下所示，结合要求对指定数据进行获取:

表1 竞赛内容数据

all_merchant.csv	province.csv
city.csv	sms_installation_jobs.csv
equipment_category.csv	sms_installation_plan_details.csv
kms_categories.csv	sms_installation_plans.csv
kms_causes.csv	sms_sis.csv
kms_measures.csv	sms_sos.csv
prod_equipment_temp.csv	

在获取数据时，对应要求如下:

* 数据存储位置为Hive数据库equipment_dashboard

* 创建省份表ods_province，将province.csv数据导入ods_province，自行定义表结构

* 创建城市表ods_city，将city.csv数据导入ods_city，自行定义表结构

* 其他数据已存入Hive对应数据库中，可直接进行操作

（2） 使用put命令将工单故障记录数据上传至HDFS;

* 工单故障记录文件为sms_so_failure_logs_shell.txt

* 写入位置为hdfs上/source/logs/sms_so_failure_logs/

（3） 使用put命令将设备数据上传至HDFS;

* 设备数据文件为province_iso_shell.txt

* 写入位置为HDFS上/source/logs/province_iso/

2. 子任务二：数据清洗

（1）对/root/eduhq/equipment/目录下工单故障记录表sms_so_failure_logs.txt进行文本清洗，删除数据中第一行标题，避免在 Hive 导入时报错，同时删除前两列脏数据，结果另存为sms_so_failure_logs_shell.txt;

（2）对/root/eduhq/equipment/目录下设备表province_iso.txt 进行文本清洗，删除数据中第一行标题，避免在Hive导入时报错，同时删除前两列脏数据，结果另保存为province_iso_shell.txt。

（二）任务二：数据标注

使用MapReduce编写任务，对工单故障记录表sms_so_failure_logs进行操作，其中针对空字段进行分类，统

一处理，添加设备状态标签“未获取”；

添加标签后的数据保存至 HDFS，具体路径为 /source/mr/sms_so_failure_logs/;

* 判断每行字段的长度，保证字段一致

* 针对时间字段，进行时间格式化，统一时间

* 针对空字段，统一清洗，如设置为未获取，根据实际需求来定义。

（三）任务三：数据统计

1. 子任务一：文件上传下载

（1）Hive 中创建库 equipment_dashboard，作为 Hive 数据仓库公用的数据，并切换到此数据库下;

（2）将标注后/source/mr/sms_so_failure_logs 数据，上传至Hive表ods_sms_so_failure_log,自行创建数据表;

（3）将/source/mr/province_iso/数据，上传至 Hive 表 ods_province_iso，自行创建数据表。

2. 子任务二：数据统计

（1）统计设备数量;

（2）统计用户数量。

四、模块三：业务分析与可视化

（一）任务一：数据可视化

1. 子任务一：数据分析

（1）分析故障类型分布，进行正序排序展示前五名;

（2）对交付状态分析，进行正序排序展示前五名;

（3）对设备状态分析，查看各状态分布。

2. 子任务二：数据可视化

使用离线数仓结合关键信息，将结果可视化展出，提高数据可读性。

（1）制作设备类型 TOP5 饼状图;

（2）制作设备状态饼状图;

（3）制作交付状态条形图;

（4）制作设备数量数字卡片;

（5）制作用户数量数字卡片;

（6）制作设备省份分布 TOP5 饼状图;

（7）制作设备维保分析折线图;

（8）制作故障类型分布 TOP5 柱状图。

（二）任务二：业务分析

1. 子任务一：业务分析

（1）对设备类型进行分析，进行正序排序展示前五名；

（2）对设备维保进行分析，了解设备维保时间变化趋势；

（3）对设备分布省份进行分析，了解设备在不同地域的市场情况。

2. 子任务二：报表分析

根据设备表 province_iso.txt 中数据，通过 Excel 生成报表对 region name 区域数据进行透视分析，及时把握市场信息。