# ZZ052–大数据应用与服务赛项试题 09

## 一、背景描述

随着互联网、大数据等技术的高速发展，信息技术的迅猛发展和数字中国战略的推进，传统电力企业面临着许多挑战和机遇。这个行业涉及到发电、输电、配电和销售等多个环节，人工操作和决策是其中的核心驱动力。然而，传统电力行业也面临着一些问题，例如生产效率低下、能源浪费、环境污染等。为了适应市场需求的变化、提高生产效率和经营管理水平，传统电力企业亟需进行数字化转型。

通过数字化转型，传统电力企业可以实现以下目标：首先，提高生产效率。通过引入大数据技术和智能化设备，传统电力企业可以优化能源生产过程，提高发电效率，减少资源浪费。其次，提升经营管理水平。数字化转型可以帮助电力企业建立起全面的数据监控和分析系统，实现对供应链、设备运行状态、能源消耗等方面的实时监控和分析，从而优化运营管理决策，提高经营效益。最后，推动可持续发展。

数字化转型可以帮助传统电力企业实现清洁能源的有效利用和管理，减少环境污染，推动可持续发展。，传统电力行业中各个部门面临的一些问题。首先，在发电部门，通过应用大数据分析技术，可以实现对发电设备的运行状态进行实时监测和优化，提高发电效率和可靠性。其次，在输电

部门，利用大数据分析输电线路的负载情况和故障预测，可以优化输电网络的运行，提高输电效率和可靠性。在配电部门，通过应用大数据分析和挖掘用户用电数据，可以实现对用户需求的精准预测和调整，提高配电效率和用户满意度。

## 二、模块一：平台搭建与运维

宿主机中 3 个台虚拟机，主机名分别修改为 node01、node02、node03，按照要求进行集群搭建。

### （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录；

（3）从宿主机/root 目录下将文件 jdk-8u212-linux-x64.tar.gz 复制到 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

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 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 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。

### 5. 子任务五：Sqoop 安装配置

（1）从宿主机/root 目录下将文件 sqoop-1.4.7.bin__hadoop-2.6.0.tar 复制到 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，添加 java-json 和 mysql-connector 到 Sqoop 指定路径。

## （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：创建相关表

（1） 结合数据特征，创建数据表fraud.fraud;

（2） 将本地/root/hqedu/目录下的数据文件fraud.txt导入MySQL对应数据库表。

### 3. 子任务三：维护数据表

结合数据进行如下查询和操作。

（1） 使用EXPLAIN语句获取查询语句的执行计划;

（2） 为fraud表字段class(案件副类别)创建索引;

（3） 对比索引查询情况。

## 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

编写agent文件power.conf，使用Flume采集无人机巡检数据power.txt，数据文件参考数据清洗部分;

目标数据源类型为HDFS

写入位置为hdfs上/source/logs/power/

#### 2. 子任务二：数据清洗

（1） 对/root/eduhq/目录下无人机巡检表power.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new_power.txt;

（2） 对/root/eduhq/目录下巡查人员表power_people.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new_power_people.txt。

**（二）任务二：数据标注**

（1）根据国家统计局2022年收集到的电力用电量舆情文本数据，抽取出部分数据，使用开源标注工具根据不同任务需求对数据进行分类标注，并将标注结果数据导出；

（2）标注文本中的行业分类。

### （三）任务三：数据统计

#### 1. 子任务一：文件上传

参考之前采集输出文件路径，然后进入某个分区查看日志文件，并将查看的日志文件下载至/root/eduhq。

#### 2. 子任务二：数据统计

统计各个电压等级对应的线路名称。

## 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

完成无人机巡检杆塔总数，根据提供的/root/eduhq 路径下 power.txt 数据，针对无人机巡检杆塔总数，计算不同型号无人机巡检总杆塔所占的比重。（比重=各个机型对应总的巡检杆塔总数/所有机型总的巡检杆塔数），将实现结果写出到 HDFS 文件系统/root/power_opt2/目录。

#### 2. 子任务二：数据可视化

使用离线数仓分析出企业重点指标后，可以结合关键信息，将结果可视化展出，提高数据可读性。

（1） 无人机与巡检员工作量统计词云图;

（2） 月度无人机巡检统计柱状图;

（3） 月度巡检人员统计柱状图。

## （二）任务二：业务分析

### 1．子任务一：业务分析

完成月报表数据分析，根据提供的/root/eduhq 路径下 power.txt 数据分析这个月巡检人员实际完成的巡检数量，将实现结果写出到 HDFS 文件系统/root/power_opt3/目录。

### 2．子任务二：报表分析

根据电力信息表中数据，通过 Excel 生成报表对 label 区域数据进行透视分析，及时把握行业信息。