

# 主观题 1：数字技术平台部署（30 分）

本任务需要在 Linux 下中完成完全分布式 Hadoop 集群搭建、MySQL 和 Hive 安装及配置，并验证组件的可用性，请使用 root 用户完成相关配置，具体要求如下。

## 1.1 完全分布式 Hadoop 集群搭建：

（1）在 master 主节点将/data 目录下的 JAVA 安装包和 Hadoop 安装包解压到/opt/software 目录下（需自行创建/opt/software 目录），并将解压后的 JAVA 远程拷贝至 slave1、slave2 相同目录下，最后查看各节点的/opt/software 目录结构并将查看结果截图。

（2）在 3 个节点的/etc/profile 文件中配置 JDK 环境变量 JAVA\_HOME、Hadoop 环境变量 HADOOP\_HOME 和 PATH 的值，并让配置文件立即生效，之后在 master 节点使用“java -version”查看 JAVA 版本，检测 JAVA 是否安装成功，将查看 JAVA 版本结果截图。

（3）根据下表修改 core-site.xml、hdfs-site.xml、mapred-site.xml、yarn-site.xml、workers、hadoop-env.sh、yarn-env.sh 配置文件，以及 HDFS 与 YARN 的启动和关闭脚本。

表 1 Hadoop 集群部署规划

服务器	master	slave1	slave2
HDFS	NameNode		
HDFS		SecondaryNameNod	
HDFS		DataNode	DataNode
YARN	ResourceManager		
YARN		NodeManager	NodeManager
历史日志服务器	JobHistoryServer		

（4）在 master 节点上使用 scp 命令将配置完的 Hadoop 安装目录远程拷贝至 slave1 和 slave2 相同目录下，之后查看 slave1 和 slave2 的/opt/software 目录结构并将查看结果截图。

（5）在主节点格式化集群，成功格式化之后在主节点依次启动 HDFS、YARN 服务、JobHistoryServer 服务，并查看其节点进程，将查看结果截图。

（6）在 HDFS 文件系统中创建/etc 目录，之后将本地/etc/profile 文件上传至 HDFS 的/etc 目录下，并查看该目录下的文件和目录。

（7）使用 hadoop-mapreduce-examples-3.1.4.jar 包中的“wordcount”类对 HDFS 上的/etc/profile 文件内容进行单词计数，设置输出路径为“/output/”，查看最终单词计数结果中出现次数最多的 5 个单词。

## 1.2 MySQL 配置：

- (1) 将 /data 目录下的 mysql-community.repo 文件复制到 YUM 仓库的配置文件 /etc/yum.repos.d/。
- (2) 加载 MySQL 镜像源并使用 yum 命令下载安装 mysql-community-server。
- (3) 在/var/log/mysqld.log 文件中查询 MySQL 初始密码，并使用初始密码登录 MySQL。
- (4) 使用初始密码登录 MySQL 后，先将密码修改为符合 MySQL8.x 密码规则的复杂密码，之后再修改密码规则并将密码改为简单密码 123456。
- (5) 修改密码后退出 MySQL，使用修改之后的密码重新登录 MySQL，赋予 root 用户外部连接权限。

1.3 Hive 安装及配置：

- (1) 将/data 目录下的 Hive 安装包解压到/opt/software 目录下（需自行创建/opt/software 目录）。
- (2) 进入 Hive 安装目录的 conf 目录，将 hive-env.sh.template 重命名为 hive-env.sh，之后查看 /etc/profile 文件中的 Hadoop 安装目录（HADOOP\_HOME），并将查看到的 HADOOP\_HOME 添加至 hive-env.sh 文件末尾。
- (3) 在 Hive 安装目录的 conf 目录下新建 hive-site.xml 配置文件并添加内容。

表 2 hive-site.xml 部分参数

配置参数	描述	参数值
hive.metastore.warehouse.dir	元数据库位置	hdfs://master:8020/user/hive/warehouse
javax.jdo.option.ConnectionURL	元数据库的链接信息	jdbc:mysql://master:3306/hive?createDatabaseIfNotExist=true
javax.jdo.option.ConnectionDriverName	连接数据库驱动	com.mysql.cj.jdbc.Driver
javax.jdo.option.ConnectionUserName	连接数据库用户名称	root
javax.jdo.option.ConnectionPassword	连接数据库用户密码	123456

- (4) 将/data 目录下的 MySQL 驱动 mysql-connector-java-8.0.30.jar 复制到 Hive 安装目录的 lib 目录，同时将该 lib 目录下的 jline-2.12.jar 复制到各节点的 Hadoop 安装目录的 /share/hadoop/yarn/lib/目录中。
- (5) 删除 Hive 安装目录的 lib 目录下的 guava-19.0.0.jar 包，并将 Hadoop 安装目录的 /share/hadoop/common/lib/目录中的 guava-27.0-jre.jar 复制至 Hive 安装目录的 lib 目录下。
- (6) 在 master 主节点的/etc/profile 文件中配置 Hive 环境变量 HIVE\_HOME 和 PATH 的值，并让配置文件立即生效。
- (7) 初始化 Hive 元数据库，之后依次启动 Hadoop 集群、MySQL 服务和 Hive 元数据服务。

(8) 进入 Hive CLI 在 Hive 中创建一个名为 school 的数据库，并在该数据库下创建一个名为 student 的数据表，字段包括 “id、name、gender、age”，数据类型分别为 “int、string、string、int”。

(9) 先使用 insert 语句向表中插入三条测试数据，再使用 select 语句查看表数据。

#### 【说明】

(1) 进入环境后需先在 Linux 终端执行命令 “initnetwork”，或者双击桌面上名称为 “初始化网络” 的图标，初始化实训平台网络。

(2) 提供的环境中的 master、slave1、slave2 节点已设置 SSH 免密登录，且各节点时间已同步。若要切换至 slave1 或 slave2 节点，可以打开新的 Linux 终端窗口，然后输入 “ssh slave1” 或 “ssh slave2” 即可切换到对应的节点。

(3) 安装包获取需要先在 Linux 终端使用 wget 命令获取：

```
“ wget -P /data/ http://house.tipdm.com/SZ-Competition/software/jdk-8u281-linux-x64.tar.gz ”
```

```
“wget -P /data/ http://house.tipdm.com/SZ-Competition/software/hadoop-3.1.4.tar.gz”
```

```
“wget -P /data/ http://house.tipdm.com/SZ-Competition/software/mysql-community.repo”
```

```
“  
                                wget                                -P                                /data/  
http://house.tipdm.com/SZ-Competition/software/mysql-connector-java-8.0.30.jar”
```

```
“wget -P /data/ http://house.tipdm.com/SZ-Competition/software/apache-hive-3.1.2-bin.tar.gz”
```

## 主观题 2：大数据技术应用（40 分）

2、有一份二手房数据：市区、小区、户型、朝向、楼层、装修情况、电梯、面积(m<sup>2</sup>)、价格(万元)、年份，已存入到 `house_sales.csv` 文件中，请使用 `pandas` 读取 `house_sales.csv` 并完成下列任务。

（1）删除面积为空或为零的记录，并将结果存储为 `cleaned_data_c1_N.csv`，N 为删除的数据条数；

（2）删除“价格(万元)”为空或异常高（超过平均价格的 3 倍）的记录，并将结果存储为 `cleaned_data_c2_N.csv`，N 为删除的数据条数；

（3）对房型数据进行标准化，例如将不规范的房型描述（2 房间 2 卫）转换为标准格式（2 室 0 厅），并存储为 `cleaned_data_c3_N.csv`，N 为修改的数据条数；

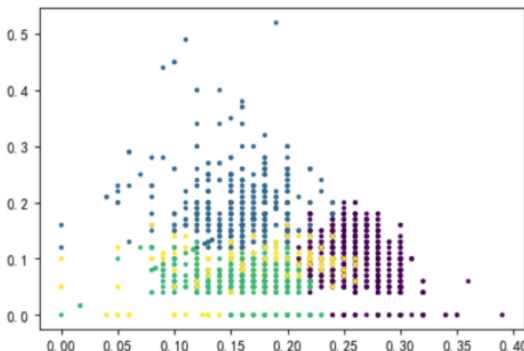
（4）删除电梯字段为空白的记录，并将结果存储为 `cleaned_data_c4_N.csv`，N 为删除的数据条数；

（5）删除面积(m<sup>2</sup>)小于 20 的记录，将结果存储为 `cleaned_data_c5_N.csv`，N 为删除的数据条数；

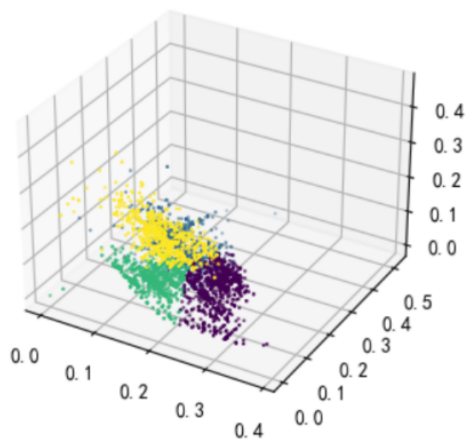
3、根据题目要求，编写 Python 代码。

（1）读取附件 `22.csv` 赋值给 `mix`，筛选“总氮百分比”，“P2O5 百分比”，“K2O 百分比”三列数据并将数据中的百分数转为小数；读取 `y_pred.npy` 数据并赋值给 `y_pred`。

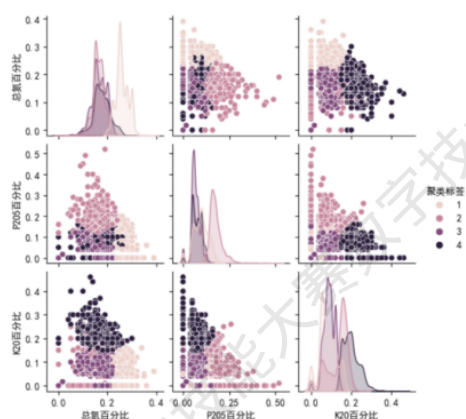
（2）将 x 轴设置为“总氮百分比”列，“P2O5 百分比”列设置为 y 轴，绘制散点图；其中 `y_pred` 数据设置为散点图的点颜色，点的大小设置为 5。



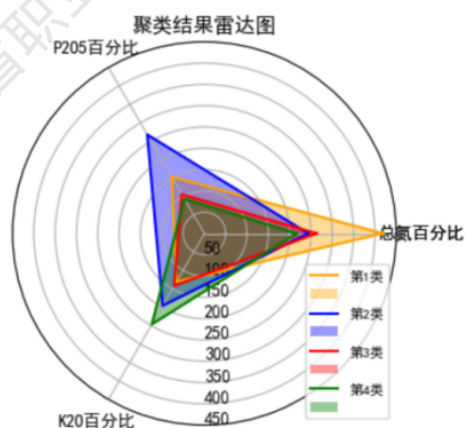
（3）根据筛选的“总氮百分比”，“P2O5 百分比”，“K2O 百分比”三列数据绘制 3D 散点图，其中 `y_pred` 数据设置为散点图的点颜色，点的大小设置为 0.5。



(4) 在 `mix` 数据中增加“聚类标签”列，值为 `y_pred+1`。使用 `seaborn` 库绘制散点图矩阵，并针对“聚类标签”列进行分类。



(5) 根据任务(4)中构建的数据，针对“聚类标签”列进行分组并对其他列数据进行求和操作。使用分组求和得到的数据绘制雷达图。



#### 【数据获取】

下载题目附件中的数据，上传到实训平台中

#### 【文件读取路径】

“/data/附件 22.csv”

“/data/y\_pred.npy”

### 主观题 3：人工智能技术应用（30 分）

4、为促进产品的销售，厂商经常会通过多个渠道投放广告。根据某公司在电视、广播和报纸上的广告投放数据预测广告收益，可以作为公司制定广告策略的重要考依据。请根据下列任务，编写相关 Python 代码。

（1）导入相关库，读取“广告收益数据.csv”数据，并查看数据的基本情况。

（2）查看数据中的缺失值，针对缺失值使用该列的平均值进行填充，并打印输出缺失值填充前后的情况。

（3）删除数据框中的第一列数据。

（4）查看数据的重复值并删除重复值所在的行。

（5）划分特征变量和目标变量，其中“收益”为目标变量，并根据特征变量和目标变量划分对应的训练集和测试集，测试集的比值为 0.2，并设置随机种子为 123。

（6）搭建 LightGBM 回归模型，使用训练集数据对模型进行训练，使用测试集数据对模型进行预测，并打印预测内容的前 10 个结果数据，将预测值和真实值放到一个 DataFrame 里进行查看比对。

（7）计算获取模型预测 RMSE、MSE 和 R2 指标值。

（8）使用预测值和真实值绘制折线图，展示预测值与真实值的变化趋势，并将折线图标题设置为“LightGBM 模型预测广告收益”。

#### 【数据获取】

下载题目附件中的数据，上传到实训平台中

#### 【文件读取路径】

“/data/广告收益数据.csv”