

主观题 1：数字技术平台部署（30 分）

本任务需要在 Linux 中完成 Flume 的安裝配置、ZooKeeper 集群和 Kafka 集群的搭建，请使用 root 用户完成相关配置，具体要求如下。

1. Flume 安裝配置：

- （1） 将 master 节点的/data 目录下的 Flume 安装包解压到/opt/software 目录下(需自行创建 /opt/software 目录)。
- （2） 进入 Flume 安装目录的 conf 目录，将 flume-env.sh.template 重命名为 flume-env.sh，并将/etc/profile 文件中的 Java 安装目录（JAVA_HOME）添加至 flume-env.sh 文件末尾。
- （3） 删除 Flume 安装目录的 lib 目录下的 guava-11.0.2.jar 包，之后查看/etc/profile 文件中的 Hadoop 安装目录（HADOOP_HOME），将 Hadoop 安装目录的/share/hadoop/common/lib/ 目录中的 guava-27.0-jre.jar 复制至 Flume 安装目录的 lib 目录。
- （4） 在 master 节点的/etc/profile 文件中配置 Flume 环境变量 FLUME_HOME 和 PATH 的值，并使配置文件立即生效，之后查看 Flume 版本，检测 Flume 是否安装成功。

2. ZooKeeper 集群搭建：

- （1） 将 master 节点的/data 目录下的 ZooKeeper 安装包解压到/opt/software 目录下(需自行创建/opt/software 目录)。
- （2） 在 master 节点切换至 ZooKeeper 安装目录的 conf 目录下，将 zoo_sample.cfg 重命名为 zoo.cfg，并按照下表修改或添加 zoo.cfg 文件中参数。

表 1 zoo.cfg 文件参数

参数名称	参数值
dataDir	/usr/lib/zookeeper
dataLogDir	/var/log/zookeeper
clientPort	2181
tickTime	2000
initLimit	5
syncLimit	2
server.1	master:2888:3888
server.2	slave1:2888:3888
server.3	slave2:2888:3888

- （3） 在各节点新建 zoo.cfg 文件中的“dataDir”和“dataLogDir”对应目录。
- （4） 在 master、slave1、slave2 节点的“dataDir”目录下新建“myid”文件，三个节点的

文件内容依次为 1、2、3。

(5) 将 master 节点配置好的 ZooKeeper 文件远程发送至 slave1、slave2 节点相同目录下。

(6) 在 master 节点的/etc/profile 文件中配置 ZooKeeper 环境变量 ZK_HOME 和 PATH 的值，并使配置文件立即生效。

(7) 将 master 节点配置好的/etc/profile 文件远程发送至 slave1、slave2 节点，并使配置文件立即生效。

(8) 分别在三个节点启动 ZooKeeper 集群，并在所有节点启动后查看各节点的状态。

(9) 分别在三个节点使用 jps 命令查看启动 ZooKeeper 后的进程号，并使用 kill 命令强制终止进程后再次查看当前启动的进程。

3. Kafka 集群搭建:

(1) 将 master 节点的/data 目录下的 Kafka 安装包解压到/opt/software 目录下(需自行创建 /opt/software 目录)。

(2) 进入 Kafka 安装目录的 config 目录修改 server.properties 配置文件，将“broker.id”改为“0”，“log.dirs”改为“/opt/logs/kafka-logs”，“zookeeper.connect”改为“master:2181,slave1:2181,slave2:2181”。

(3) 将 master 节点配置好的 Kafka 文件远程发送至 slave1、slave2 节点相同目录下，并将 slave1、slave2 节点的 server.properties 配置文件中的 broker.id 分别修改为 1、2。

(4) 在 master 节点的/etc/profile 文件中配置 Kafka 环境变量 KAFKA_HOME 和 PATH 的值，并使配置文件立即生效。再将 master 节点配置好的/etc/profile 文件远程发送至 slave1、slave2 节点，同样使配置文件立即生效。

(5) 分别在各节点启动 ZooKeeper 集群，确保 ZooKeeper 集群启动后再在各节点启动 Kafka 集群，并查看各节点进程。

【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) 若想切换至 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) 安装包获取需要在 Linux 终端使用 wget 命令获取：

```
“          wget          -P          /data/
http://house.tipdm.com/SZ-Competition/software/apache-flume-1.11.0-bin.tar.gz”
```

```
“  
wget -P /data/  
http://house.tipdm.com/SZ-Competition/software/apache-zookeeper-3.6.3-bin.tar.gz”  
“wget -P /data/ http://house.tipdm.com/SZ-Competition/software/kafka_2.12-2.4.1.tgz”
```

主观题 2：大数据技术应用（40 分）

2、城市游客接纳能力是城市规划建设中的重要指标，其中城市的酒店房间数量是城市游客接纳能力的关键要素。在企业消费平台上，各地区的酒店信息能够反映一个地区商业活动的密集程度。例如酒店总量多的城市大都具有强烈的吸纳外来人员的能力，订单数量能够反映该地区的有较多的商业往来。请编写程序或脚本根据数据文件 `hotel.csv` 统计以下的相关信息，具体要求如下：

- （1）导入相关库，读取数据并筛选并展示评分最高的数据信息。
- （2）分别统计各个商圈的酒店总数，进行倒序排序展示前五名。
- （3）统计各个商圈酒店的平均房间数，进行正序排序展示前五名。
- （4）统计所有五星级酒店的平均评分。

【数据获取】

下载题目附件中的数据，并上传到实训平台中。

【文件读取路径】

```
“/data/hotel.csv”
```

3、基于故宫评论数据，根据题目要求运用 Python 完成下列任务。

- （1）导入相关库，并读取“故宫评论情感.csv”数据。
- （2）根据数据中的“emotion”列数据绘制情感分直方图，并将 X 轴标签设置为“情感分”，Y 轴标签设置为“数量”，直方图的标题设置为“情感分直方图”。
- （3）设置停用词表，并往停用词表中添加“的”、“了”等停用词；将“评论内容”列的句子数据进行分词，然后绘制相应的词云图，并将词云图保存到“词云图.png”。
- （4）假设情感分大于等于 0.5 的评论为积极评论，小于 0.5 的评论为消极评论，分别统计数据中积极评论数量和消极评论数量，并绘制对应的积极与消极评论的占比扇形图。

【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/故宫评论情感.csv”

主观题 3：人工智能技术应用（30 分）

4、基于广州各区二手房价的数据，按照题目要求使用 Python 完成下列任务。

（1）导入相关库，读取“广州二手房价.csv”数据，合并“region”与“positionInfo”列数据，并新建“region_positionInfo”列用于保存；去除“unitPrice”列数据存在的特殊符号“，”。

（2）将新建的“region_positionInfo”列数据进行数值型编码，并删除原来的“region”与“positionInfo”列数据。

（3）将“totalPrice”列数据作为标签，其余数据作为特征数据，并将训练数据进行归一化处理。

（4）按照训练集：测试集为 8：2 的比例进行数据集的划分，并将随机种子设置为 4。

（5）查看特征数据之间的特征重要性。

（6）构建随机森林回归模型并训练，输出模型在训练集上的得分，并使用模型对测试集数据进行预测。

（7）计算模型的评价指标 r^2 系数、MAE（平均绝对误差）、MSE（均方误差）与 RMSE（均方跟误差）。

【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/广州二手房价.csv”