

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、Hive 安装配置和 Flink 集群的搭建，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 slvae1、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 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	元数据库的连接信息	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。

1.3 Flink 集群搭建:

(1) 在 master 节点将/data 目录下的 flink-1.14.0-bin-scala_2.12.tgz 安装包解压到 /opt/software 目录下（需自行创建/opt/software 目录）。

(2) 进入 Flink 安装目录的 conf 目录，修改 workers 文件，注释原文件内容并添加 slave1

和 slave2。

(3) 进入 Flink 安装目录的 conf 目录，修改 masters 文件，注释原文件内容并添加 master:8081。

(4) 进入 Flink 安装目录的 conf 目录，按照下表修改或添加 flink-conf.yaml 文件中参数（参数值中 master 需替换为/etc/hosts 文件中的具体 IP）。

表 3 flink-conf.yaml 文件参数

参数名称	参数值
jobmanager.rpc.address	master
jobmanager.heap.size	512m
taskmanager.memory.flink.size	1024m
taskmanager.numberOfTaskSlots	3
parallelism.default	1
rest.address	master
rest.bind-address	master
io.tmp.dirs	/opt/logs/flink/tmp

(5) 在每个节点上创建日志文件目录/opt/logs/flink/tmp。

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

(7) 在 master 节点修改/etc/profile 文件，设置 Flink 环境变量 FLINK_HOME 和 PATH 的值，并使配置文件立即生效。

(8) 启动 Flink 集群，并查看各节点的进程。

【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) 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-connector-java-8.0.30.jar”

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

“wget -P /data/ http://house.tipdm.com/SZ-Competition/software/flink-1.14.0-bin-scala_2.12.tgz”

主观题 2：大数据技术应用（40 分）

2、根据下列任务，完成相应 Python 代码。

（1）读取 result1.csv，将“总氮百分比”列数据字段为“复混肥料”、“有机-无机复混肥料”、“床土调酸剂”的值设置为 None，然后计算该列的平均值以填充 None 值；计算 result1.csv 中各肥料产品的氮、磷、钾养分百分比之和，称为总无机养分百分比。总无机养分百分比的公式为：

总无机养分百分比 = 总氮百分比 + P2O5 百分比+K2O 百分比

将总氮百分比、P2O5 百分比、K2O 百分比相加得到总无机养分百分比，并将计算结果保留三位小数，并将数据保存为“result2_1.csv”。

（2）从附件 2 中筛选出复混肥料的产品，将所有复混肥料按照总无机养分百分比的取值等距分为 10 组。根据每个产品所在的分组，为其打上分组标签（标签用 1~10 表示），将完整的结果保存到文件“result2_2.xlsx”中。

（3）从附件 2 中筛选出有机肥料的产品，将产品按照总无机养分百分比和有机质百分比分别等距分为 10 组，并为每个产品打上分组标签 (1,1), (1,2), ..., (10,10)，将完整的结果保存到文件“result2_3.xlsx”中。

【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/result1.csv”

“/data/附件 2.csv”

3、基于附件.xlsx 的招聘数据，根据题目要求运用 Python 完成下列任务。

（1）读取附件.xlsx 数据，构造各个职业主题的岗位招聘量透视表（列为职业主题，行为岗位），并找出每个职业主题最多招聘量的岗位。

（2）薪资计算方式有日、月、年三种计薪方式，将薪资计算方式统一成月薪（假设一个月有 30 天），相应的“最低薪资”“最高薪资”进行值修改，删除“薪资计算方式”列。根据“最低薪资”“最高薪资”计算平均薪资，结果作为新列保存到数据中，列名记为“平均薪资”

(3) 使用 Seaborn 库绘制不同学历的平均薪资箱线图：设定 x 轴标签为“学历”；y 轴标签为“平均薪资”；图形标题（title）为“不同学历的平均薪资分布”（中文字体设置：WenQuanYi Zen Hei）。

(4) 统计各个岗位的招聘需求量，根据招聘需求量进行降序排序。绘制岗位需求量前十的柱状图：设定 x 轴标签为“岗位”；y 轴标签为“招聘需求量”；图形标题（title）为“招聘需求量 Top10 的岗位”（中文字体设置：WenQuanYi Zen Hei）。

(5) 统计“职位关键词”列的关键词并绘制词云图。

【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/附件.xlsx”

主观题 3：人工智能技术应用（30 分）

4、请编写相应的 Python 代码，完成下列任务。

(1) 从 result2_2.csv 中筛选出复混肥料的产品，按照氮、磷、钾养分的百分比，使用聚类算法将这些产品分为 4 类。

(2) 根据聚类结果为返回每个簇的中心，并将完整的预测结果保存到文件“y_pred.npy”中。

(3) 使用 matrix.csv 数据，并完成下列任务：

①读取 matrix.csv 数据，并赋值给变量 data。

②将读取的数据去除“user_id”、“merchant_id”、“label”三列数据再进行数据的标准化，并将数据划分为训练集和验证集，其中标签数据为“label”列。

③构建 LightGBM 回归模型，并设置模型的训练次数为 100，树的深度为 10，树的最大叶节点为 25，学习率为 0.1，特征采样率为 0.5 以及设置模型评估指标。

④根据划分的数据，对构建好的 LightGBM 回归模型进行训练；同时添输出模型训练集和验证集的准确率，并设置模型精度在 100 次没有更新的时候停止训练。

【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/result2_2.csv”

“/data/matrix.csv”

河北省职业院校技能大赛数字技术应用