上海高职院校学生技能大赛

大数据应用开发师生同赛样题

模块A：大数据平台及组件搭建

（一）任务1：大数据平台及组件搭建

子任务一：基础环境搭建

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） 修改三台节点的主机名称分别为 master、slave1 和 slave2，并编辑 hosts 文件添加IP主机名映射;

（2） 配置免密登录，在 master 节点上生成 RSA 非对称加密类型的 SSH 密钥对;

（3） 将 master 节点中的公钥拷贝到 slave1 和 slave2 节点上;

（4） 在 master 节点通过 SSH 连接 slave1 和 slave2 验证免密登录配置是否成功;

（5） 配置 java 环境，解压 master 中的 /opt/software 目录下 jdk-8u212-linux-x64.tar.gz 安装包到 /opt/module(若路径不存在，则需新建) 路径下;

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

16

子任务二：Zookeeper与Hadoop搭建配置

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） 在master节点将/opt/software目录下的apache-zookeeper-3.5.7-bin.tar.gz包解压到/opt/module(若路径不存在，则需新建)路径下；

（2） 在master节点上面将配置的Zookeeper环境变量文件及Zookeeper解压包拷贝到slave1、slave2节点；

（3） 将slave1节点上面/opt/zookeeper-3.5.7/data目录下的myid文件内容修改为2，将slave2节点上面/opt/zookeeper-3.5.7/data目录下的myid文件内容修改为3。

（4） 在master将/opt/software目录下的Hadoop解压到/opt/module(若路径不存在，则需新建)目录下，配置相关配置文件，并将解压包分发至slave1、slave2中，其中master、slave1、slave2节点均作为datanode，配置好相关环境，初始化Hadoop环境namenode。

子任务三：MySQL数据库安装配置

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） 将MySQL 5.7.25安装包解压到/opt/module目录下；

（2） 首先，通过yum命令移除可能与MySQL冲突的MariaDB库。接下来，使用rpm -ivh依次安装mysql-community-common、mysql-community-libs、mysql-community-libs-compat、mysql-community-client和mysql-community-server包；

（3） 在成功安装MySQL后，以mysql用户身份执行命令初始化数据库，并使用不安全模式（生成空密码）进行设置。随后启动MySQL服务；

16

（4） 使用root用户无密码登录MySQL，然后将root用户的密码修改为123456，修改完成退出MySQL，重新登录验证密码是否修改成功；

（5） 为了允许MySQL用户远程登录，需要在"mysql"数据库的"user"表中更改用户的host项，将所有用户的localhost权限修改为允许任意远程主机访问；设置完成刷新配置信息，让其生效。

子任务四：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） 将 master 节点的/opt/software目录下 apache-hive-3.1.2-bin.tar.gz 安装包解压到/opt/module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查看版本号；

（3） 完成相关配置并添加所依赖包，将MySQL数据库作为Hive元数据库。初始化Hive元数据，并通过schematool相关命令执行初始化。

子任务五： Spark安装部署

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） 将master中的/opt/software目录下 spark-3.1.1-bin-hadoop3.2.tgz 安装包解压到/opt/module路径中(若路径不存在，则需新建)；

（2） 修改容器中/etc/profile文件，设置Spark环境变量并使环境变量生效，运行命令spark-submit --version查看版本号。

16

模块B：大数据平台管理与运维

（一）任务1：大数据平台管理与运维

子任务一：启动大数据组件服务

本任务需要使用root 用户启动大数据组件服务，具体要求如下：

（1） 在master节点、slave1节点、slave2节点分别启动zookeeper；

（2） 启动Hadoop集群（包括hdfs和yarn），使用jps命令查看master节点与slave1节点的Java进程；

（3） 在master节点中查看MySQL Server服务状态是否为开机自启；

（4） 在master节点开启hive元数据服务；

（5） 启动Spark集群，使用jps命令查看master、slave1和slave2节点的Java进程。

子任务二：文件系统垃圾回收

开启垃圾回收站，可以将要删除的文件首先放置在回收站中，等待配置的时间结束，进行真正的数据删除。：

（1） 设置HDFS垃圾回收机制，要求保留垃圾回收站的文件或文件夹7天，超过就自动删除；

16

（二）任务2：数据装载与程序调用

子任务一：文件上传下载

本任务需要使用HDFS命令，已安装Hadoop及需要配置前置环境，具体要求如下：

（1） 在HDFS上创建 /user/hadoop/input目录；

（2） 在master节点将demo.csv文件上传到HDFS的/user/hadoop/input目录下；

（3） 修改权限，赋予目录/user/hadoop/input最高777权限；

子任务二：Hadoop程序运行

本任务需要使用 Hadoop 默认提供的 sudoku 示例来完成如下数独题目解题任务，将数独解题结果写入主机/temp/result.data：

```
8 5 ? 3 9 ? ? ? ?
? ? 2 ? ? ? ? ? ?
? ? 6 ? 1 ? ? ? 2
? ? 4 ? ? 3 ? 5 9
? ? 8 9 ? 1 4 ? ?
3 2 ? 4 ? ? 8 ? ?
9 ? ? ? 8 ? 5 ? ?
? ? ? ? ? ? 2 ? ?
? ? ? ? 4 5 ? 7 8
```

子任务三：Spark程序运行

本任务需要使用Spark编写程序并打jar包上传集群，使用spark-sumbit方式对HDFS上数据进行单词数统计任务，具体要求如下：

（1） 程序数据路径为集群节点测试数据对应路径/user/hadoop/input；

（2） 程序结果保存路径为/user/hadoop/output；

（3） 程序jar包名称为`sparkwordcount.jar`，保存位置为master主机/root/wordcount/；

（4） 查看 HDFS 中的/user/hadoop/output单词数统计结果。

（5） 下载结果文件保存至master节点指定目录/tmp下：

16

模块C：数据采集与处理

（一）任务1：数据采集与处理

子任务一：采集网站数据

现提供免费开源电商系统，包含PC、h5、微信小程序、支付宝小程序、百度小程序、头条&抖音小程序、QQ小程序、APP、多商户，遵循MIT开源协议发布、基于ThinkPHP5.1框架研发。根据比赛中提供的环境，进行数据爬取，具体要求如下：

根据给出的网站地址进行相关爬取操作，爬取形式不限，编写代码爬取网站的商品ID、名称、价格、浏览量、销量、库存，并将数据写入/opt/mall/goods.txt文件中。

子任务二：数据预处理

使用Python代码编写数据清洗的相关功能，所用数据为爬取的goods.txt数据，具体要求如下：

（1） 过滤掉销量大于浏览量的异常数据记录，保留其余数据，结果保存到/opt/mall_etl01/目录下，并将结果文件命名为“views_exception.csv”；

（2） 验证数据真实性，过滤掉销量大于库存的异常数据记录，保留其余数据，结果保存到/opt/mall_etl02/目录下，并将结果文件命名为“stock_exception.csv”。

16

模块D：大数据分析与挖掘

（一）任务1：数据分析

子任务一：数据存储

使用Hive对数据进行上传存储操作，所用数据为爬取的goods.txt数据，具体要求如下：

（1） 依次创建数据库data，数据表goods，结合数据特征自定义对应表字段；

（2） 将爬取后的数据上传至对应数据表；

（3） 查看上传后的表数据，按照ID升序查询前三条；

子任务二：数据统计

使用Hive对上传后的数据进行数据操作，具体要求如下：

（1） 将价格为空（null）的数据结果写入目录/root/goods01；

（2） 将名称中带有“连衣裙”、“女士”的数据结果写入目录/root/goods02；

（3） 对以上两类数据进行剔除，将筛选后的数据保存至中间表goods1；

（4） 查询中间表，按照价格降序查找前三条商品信息，格式为：（名称 价格）；

（5） 按照空格对“名称”字段数据进行分割，要求第一个元素title[0]作为对应商品品牌，其他元素作为对应商品特征，对各品牌进行信息统计；

（6） 对排名第一的品牌进行分析，对其商品特征进行数据统计；

16

（二）任务2：数据挖掘

子任务一：分类算法挖掘

对于企业来说，获得新客户的成本是非常高的，那么用户的留存分析就成了关键。每个客户的信息都是不一样的，很难找到其中的规律，但在客户数量足够多的情况下，可以使用数据挖掘算法进行分类预测，建立流失判别模型，对可能流失的用户进行重点关注。

（1） 读取数据，对数据中缺失值进行删除；

（2） 查看数据特征，对数据进行特征类型转换；

（3） 查看数据特征及特征个数；

（4） 对流失信息进行分组统计，计算流失比例；

（5） 对数值型特征进行对比，并进行数值特征可视化展示；

（6） 对数据进行分类特征查看；

（7） 使用随机森林分类算法对数据进行训练预测，查看不同特征对应重要性；

模块E：数据可视化

（一）任务1：数据可视化

子任务一：数据可视化

使用 pyecharts 库来创建直观、互动的图表。这些图表将帮助揭示数据中的关键模式和趋势。 具体要求如下：

（1） 使用饼状图展示不同品牌数据占比，其中标签为不同品牌，占比为出现频次；

16

（2） 使用柱状图展示某品牌的商品特征，柱状图中的每个柱子代表一个特征，高度代表该特征频次；

模块F：大数据综合分析及报告

（一）任务1：综合分析报告

子任务一：综合分析报告

根据任务具体描述以及数据仓库的数据撰写报告，通过数据洞察数据背后的业务含义并给出业务改进的建议。