ZZ052-大数据应用与服务赛项试题 01

一、背景描述

随着中国数字化转型战略的推进，传统通信行业正面临着数字化转型的挑战和机遇；用户对通信服务的需求已经发生了根本性的变化，通信运营商正在通过技术创新和服务升级来满足这些需求；数字化转型涉及到网络建设、数据管理、服务创新等方面，大数据技术成为关键驱动力之一。

为了应对这一转型，我们要求参赛者搭建通信行业大数据分析平台，并利用 Hive 数仓技术和 Spark 计算引擎对通信用户行为数据进行操作和分析；通过这样的平台，可以快速处理和挖掘海量数据，得出有价值的洞察和分析结果。

同时，在展示数据分析结果方面，我们要求参赛者结合前端可视化框架 ECharts 和 Python 可视化库 pyecharts，创建交互式的数据可视化图表；这些图表能够直观地展示数据分析结果，帮助管理者更好地决策企业的发展战略，并对销售、营销、客服和技术等部门的目标策略进行全面部署；通过数据可视化，销售部门可以了解产品销售趋势和市场份额；营销部门可以优化营销活动和广告投放策略；客服部门可以提供更好的客户服务；技术部门可以进行网络优化和故障排查。

二、模块一：平台搭建与运维

（一）任务一：大数据平台搭建

本模块需要使用 root 用户完成相关配置；所有组件均在/root/software 目录下。

1. 子任务一：基础环境准备

master、slave1、slave2三台节点都需要安装JDK

（1） 将JDK安装包解压到/root/software目录下；

（2） 在“/etc/profile”文件中配置JDK环境变量JAVA_HOME和PATH的值，并让配置文件立即生效；

（3） 查看JDK版本，检测JDK是否安装成功。

在master节点操作

（1） 在master上生成SSH密钥对；

（2） 将master上的公钥拷贝到slave1和slave2上；

在 master 上通过 SSH 连接 slave1 和 slave2 来验证。

2. 子任务二：Hadoop 完全分布式安装配置

master、slave1、slave2三台节点都需要安装Hadoop

（1） 在主节点将Hadoop安装包解压到/root/software目录下；

（2） 依次配置hadoop-env.sh、core-site.xml、hdfs-site.xml、mapred-site.xml、yarn-site.xml和workers配置文件；Hadoop集群部署规划如下表；

表1 Hadoop集群部署规划

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

（3） 在master节点的Hadoop安装目录下依次创建hadoopDatas/tempDatas 、 hadoopDatas/namenodeDatas 、hadoopDatas/datanodeDatas、hadoopDatas/dfs/nn/edits、hadoopDatas/dfs/snn/name 和hadoopDatas/dfs/nn/snn/edits目录;

（4） 在master节点上使用scp命令将配置完的Hadoop安装目录直接拷贝至slave1和slave2;

（5） 三台节点的“/etc/profile”文件中配置Hadoop环境变量HADOOP_HOME和PATH的值，并让配置文件立即生效;

（6） 在主节点格式化集群;

（7） 在主节点依次启动HDFS、YARN集群和历史服务

3. 子任务三：MySQL 安装配置

只在master节点操作

（1） 将MySQL 5.7.25安装包解压到/root/software目录下;

（2） 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） 安装好MySQL后，使用mysql用户初始化和启动数据库;

（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;

更改“mysql”数据库里的 user 表里的 host 项，从localhost 改成%即可实现用户远程登录;设置完成刷新配置信息，让其生效。

4. 子任务四：Hive 安装配置

只在master节点操作。

（1） 将Hive 3.1.2的安装包解压到/root/software目录下;

（2） 在“/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;

（3） 查看Hive版本，检测Hive环境变量是否设置成功;

（4） 切换到$HIVE_HOME/conf目录下，将hive-env.sh.template文件复制一份并重命名为hive-env.sh；然后，使用vim编辑器进行编辑，在文件中配置HADOOP_HOME、HIVE_CONF_DIR以及HIVE_AUX_JARS_PATH参数的值，将原有值删除并将前面的注释符#去掉；配置完成，保存退出；

（5） 将/root/software目录下的MySQL驱动包mysql-connector-java-5.1.47-bin.jar 拷贝到$HIVE_HOME/lib目录下；

（6） 在$HIVE_HOME/conf目录下创建一个名为hive-site.xml的文件，并使用vim编辑器进行编辑；

配置如下内容:

表2 配置内容

配置参数	描述	参数值
javax.jdo.option.ConnectionURL	连接元数据库的链接信息	jdbc:mysql://master:3306/hivedb?createDatabaseIfNotExist=true&useSSL=false&useUnicode=true&characterEncoding=UTF-8
javax.jdo.option.ConnectionDriverName	连接数据库驱动	com.mysql.jdbc.Driver
javax.jdo.option.ConnectionUse	连接数据库用户名称	root

rName		
javax.jdo.option.ConnectionPassword	连接数据库用户密码	123456

（7） 使用schematool命令，通过指定元数据库类型为“mysql”，来初始化源数据库的元数据;

（8） 使用CLI启动Hive，进入Hive客户端；在Hive默认数据库下创建一个名为student的管理表;

表3 数据表

字段	数据类型
id	int
name	string

（9） 通过insert语句往student表中插入一条测试数据。

5. 子任务五：Flume 安装配置

只在 master 节点操作。

（1） 将 Flume 1.11.0 的安装包解压到/root/software目录下;

（2） 在“/etc/profile”文件中配置Flume环境变量FLUME_HOME和PATH的值，并让配置文件立即生效;

（3） 使用 cd 命令进入 /root/software/apache-flume-1.11.0-bin/conf 目录下，使用 cp 命令将 flume-env.sh.template文件复制一份，并重命名为flume-env.sh;

使用vim命令打开 “flume-env.sh”配置文件，找到JAVA_HOME参数位置，将前面的“#”去掉，将值修改为本机JDK的实际位置；修改完成，保存退出；

（4） 查看Flume版本，检测Flume是否安装成功。

（二）任务二：数据库配置维护

1. 子任务一：数据库配置

在Hive中创建一个名为comm的数据库，如果数据库已经存在，则不进行创建。

2. 子任务二：创建相关表

（1） 在comm数据库下创建一个名为ods_behavior_log的外部表，如果表已存在，则先删除；分区字段为dt，即根据日期进行分区；同时，使用location关键字将表的存储路径设置为HDFS的/behavior/ods/ods_behavior_log目录；字段类型如下表所示；

表4 字段类型

字段	数据类型	说明
line	string	一整行JSON数据
dt	string	日期，分区字段

（2）使用load data子句将本地/root/eduhq/data/app_log/behavior目录下的每个数据文

件依次加载到外部表ods_behavior_log的对应分区中，按照日志文件对应日期定义静态分区(例如: dt='2023-01-01')

（3） 查看ods_behavior_log表的所有现有分区、前3行数据，并统计外部表ods_behavior_log数据总行数;

（4） 在comm数据库下创建一个名为dwd_behavior_log的外部表，如果表已存在，则先删除; 分区字段为dt，即根据日期进行分区; 另外，要求指定表的存储路径为HDFS的/behavior/dwd/dwd_behavior_log目录，存储文件类型为“orc”，文件的压缩类型为“snappy”；字段类型如下表所示;

表5 字段类型

字段	数据类型	说明
client_ip	string	客户端请求的IP地址
device_type	string	请求的设备类型，手机mobile或者电脑pc
type	string	上网的模式，4G、5G或WiFi
device	string	设备ID
url	string	访问的资源路径
province	string	省份
city	string	城市
ts	bigint	时间戳
dt	string	日期，分区字段

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

（1） 启动Hadoop集群，使用HDFS Shell指令，在HDFS根目录下级联创建一个名为/behavior/origin_log的目录，用于存储采集到的用户行为日志;

（2） 目录创建完成，使用HDFS Shell指令，将本地/root/eduhq/data/app_log/behavior目录下的所有用户行为日志文件采集至HDFS的/behavior/origin_log目录下;

（3） 采集完成，在本机打开浏览器，访问http://本机主机名:9870或http://本机IP地址:9870进入HDFS Web UI界面，查看是否成功将数据采集到HDFS上。

2. 子任务二：数据清洗

（1） 使用Windows操作系统上的Excel软件，打开名为"behavior2023-01-01.csv"的文件;

（2） 对数据进行清洗，专注处理名为"behavior2023-01-01.csv"的文件中的"time"列。将时间日期格式进行分列，分别处理为日期和时间两列。

（二）任务二：数据标注

开发一个简单的 Java 类 IpToLocUdf，继承 org.apache.hadoop.hive.ql.udf.generic.GenericUDF，

重载 initialize()、evaluate()和 getDisplayString()方法；该类需要实现通过 IP 从/root/eduhq/data/area.json 文件中随机获取“省份”和“城市”信息，完成数据的分类标注。

（三）任务三：数据统计

1. 子任务一：HDFS 文件上传下载

（1）将包“com.hive.udf”导出为名为 hive-udf-behavior-1.0.0.jar 的 JAR 文件，并保存在本地的/root/eduhq/udf_jars 目录中；

（2)将打包文件 hive-udf-behavior-1.0.0.jar 上传到 HDFS 的/hive/udf_jars 目录下；

（3）在 Hive 客户端，创建永久函数 url_trans 和 get_city_by_ip，并将它们与开发好的 class 相关联；

（4）在 Hive 客户端，使用 select 语句测试 url_trans 和 get_city_by_ip 函数；

（5)启动 Hive 的动态分区功能，并将 Hive 设置为非严格模式；

（6）使用 insert overwrite ... select ...子句将 ods_behavior_log 表中数据插入分区表 dwd_behavior_log 中，并实现根据 dt 进行动态分区。

2. 子任务二：数据统计

（1） 查看dwd_behavior_log表的所有现有分区；

（2） 查看外部表dwd_behavior_log的前3行数据，并验证URL协议是否被统一为“http”，以及通过IP是否能够获取到“省份”和“城市”信息;

（3） 统计外部表dwd_behavior_log数据总行数。

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

（1） 在comm数据库下创建一个名为dws_behavior_log的外部表，如果表已存在，则先删除；分区字段为dt，即根据日期进行分区；另外，要求指定表的存储路径为HDFS的/behavior/dws/dws_behavior_log目录，存储文件类型为“orc”，文件的压缩类型为“snappy”；字段类型如下表所示;

表6 字段类型表

字段	数据类型	说明
client_ip	string	客户端请求的IP地址
device_type	string	请求的设备类型，手机mobile或者电脑pc
type	string	上网的模式，4G、5G或WiFi
device	string	设备ID
url	string	访问的资源路径
province	string	省份

city	string	城市

（2） 启动Hive的动态分区功能，并将Hive设置为非严格模式;

（3） 使用insert overwrite ... select ...子句将dwd_behavior_log表中数据插入分区表dws_behavior_log中，并实现根据dt进行动态分区;

（4） 查看dws_behavior_log表的所有现有分区、前3行数据，并统计统计表数据总行数;

（5） 在comm数据库下创建一个名为dim_date的外部表，如果表已存在，则先删除；另外，要求指定表的存储路径为HDFS的/behavior/dim/dim_date目录，字段分隔符为“\t”，建表时添加TBLPROPERTIES('skip.header.line.count'='1')语句让Hive读取外表数据时跳过文件行首（表头）；字段类型如下表所示;

表7 字段类型表

字段	数据类型	说明
date_id	string	日期
week_id	string	周
week_day	string	星期
day	string	一个月的第几天
month	string	月份
quarter	string	季度
year	string	年份

is_workday	string	是否是工作日
holiday_id	string	国家法定假日标识

（6） 在comm数据库下创建一个名为dim_area的外部表，如果表已存在，则先删除；另外，要求指定表的存储路径为HDFS的/behavior/dim/dim_area目录，字段分隔符为“\t”；字段类型如下表所示;

表8 字段类型表

字段	数据类型	说明
city	string	城市/区/县
province	string	省份
area	string	地区

（7） 使用load data子句将本地/root/eduhq/data目录下的“dim_date_2023.txt”和“dim_area.txt”文件分别加载到外部表dim_date和dim_area中;

（8） 分别查看外部表dim_date和dim_area的前3行数据;

（9） 分别统计外部表dim_date和dim_area数据总行数;

（10）统计不同省份用户访问量;将统计结果导出到本地文件系统的/root/eduhq/result/ads_user_pro目录下，并指定列的分隔符为逗号（特别注意：因为省份是随机获取的，所以结果会有所差异）;

（11）统计不同时间段的网页浏览量将统计结果导出到本地文件系统的/root/eduhq/result/ads_user_hour目录下，并指定列的分隔符为逗号;

（12）不同网站访客的设备类型统计;将统计结果导出到本地文件系统的/root/eduhq/result/ads_visit_mode目录下，并指定列的分隔符为逗号;

（13）不同网站的上网模式统计;将统计结果导出到本地文件系统的/root/eduhq/result/ads_online_type目录下，并指定列的分隔符为逗号;

2. 子任务二：数据可视化

（1） 使用Pyecharts库绘制中国地图，以直观展示不同省份用户访问量分布情况;

- 文件名：ads_user_pro.py
- 文件存放地址：/root/eduhq/python/
- 数据目录:/root/eduhq/result/ads_user_pro目录
- 背景地址：/root/eduhq/images/img_1.png
- 图表名称：不同省份用户访问量分布图.html
- 图表存放地址：/root/eduhq/html/

（2） 使用Pyecharts库绘制一个带时间轴的柱形图，以直观展示不同经济大区用户的访问量统计情况;

- 文件名：ads_user_region.py
- 文件存放地址：/root/eduhq/python/

- 数据目录:

/root/eduhq/result/ads_user_region目录

- 背景地址: /root/eduhq/images/img_2.png
- 图表名称: 不同经济大区用户访问量统计柱形图.html
- 图表存放地址: /root/eduhq/html/

（3） 使用Pyecharts绘制网页浏览量统计折线图，直观展示不同时间段内的访问量变化趋势;

- 文件名: ads_user_hour.py
- 文件存放地址: /root/eduhq/python/
- 数据目录: /root/eduhq/result/ads_user_hour目录
- 背景地址: /root/eduhq/images/img_3.png
- 图表名称:不同时间段网页浏览量统计曲线图.html
- 图表存放地址: /root/eduhq/html/

（4） 使用Pyecharts绘制网页浏览量统计折线图，直观展示节假日和工作日不同时间段内的访问量变化趋势;

- 文件名: ads_hol_work_user.py
- 文件存放地址: /root/eduhq/python/
- 数据目录:

/root/eduhq/result/ads_hol_work_user目录

- 背景地址: /root/eduhq/images/img_3.png

- 图表名称：节假日和工作日各时间段网页浏览量统计曲线图.html
- 图表存放地址：/root/eduhq/html/

（5） 使用Pyecharts绘制堆积柱形图，直观地展示访客在不同设备类型上的访问次数情况;

- 文件名：ads_visit_mode.py
- 文件存放地址：/root/eduhq/python/
- 数据目录：/root/eduhq/result/ads_visit_mode目录
- 背景地址：/root/eduhq/images/img_2.png
- 图表名称：网站访客设备类型统计堆积柱形图.html
- 图表存放地址：/root/eduhq/html/

（6） 使用Pyecharts绘制堆积柱形图，直观地展示访客在不同上网模式下的访问次数情况;

- 文件名：ads_online_type.py
- 文件存放地址：/root/eduhq/python/
- 数据目录：/root/eduhq/result/ads_online_type目录
- 背景地址：/root/eduhq/images/img_2.png
- 图表名称：网站访客上网模式统计堆积柱形图.html
- 图表存放地址：/root/eduhq/html/

（7） 使用Pyecharts绘制词云图，直观地展示不同域名用户访问情况;

- 文件名：ads_user_domain.py
- 文件存放地址：/root/eduhq/python/
- 数据目录：/root/eduhq/result/ads_user_domain目录
- 背景地址：/root/eduhq/images/img_2.png
- 图表名称：不同域名用户访问统计词云.html
- 图表存放地址：/root/eduhq/html/

（二）任务二：业务分析

（1） 统计每天不同经济大区用户访问量；将统计结果导出到本地文件系统的/root/eduhq/result/ads_user_region目录下,并指定列的分隔符为逗号;

（2） 统计节假日和工作日的浏览量差异；将统计结果导出到本地文件系统的/root/eduhq/result/ads_hol_work_user目录下,并指定列的分隔符为逗号;

（3） 统计不同域名的用户访问量；将统计结果导出到本地文件系统的/root/eduhq/result/ads_user_domain目录下,并指定列的分隔符为逗号;