Turing AI 万维图灵 | 大数据系列课程

大数据技术-第十章：购物网站用户行为分析综合案例

本地数据集上传到数据仓库Hive

CONTENTS

01

实验数据集的下载

实验数据集的下载

本案例采用的数据集为user.zip，包含了一个大规模数据集raw_user.csv（包含2000万条记录），和一个小数据集small_user.csv（只包含30万条记录）。小数据集small_user.csv是从大规模数据集raw_user.csv中抽取的一小部分数据。之所以抽取出一少部分记录单独构成一个小数据集，是因为在第一遍跑通整个实验流程时，会遇到各种错误、各种问题，先用小数据集测试，可以大量节约程序运行时间。等到第一次完整实验流程都顺利跑通以后，可以最后用大规模数据集进行最后的测试。把数据集user.zip文件下载到Linux系统的“/opt/software/”目录下面。

实验数据集的下载

下面需要把user.zip进行解压缩，需要首先建立一个用于运行本案例的目录bigdatacase，

请执行以下命令：

```
## 创建一个bigdatacase目录
[root@master ~]$  mkdir /usr/local/src/bigdatacase
## 在bigdatacase目录下面创建一个dataset目录，用于保存数据集
[hadoop@master ~]$  mkdir /usr/local/src/bigdatacase/dataset

#解压缩user.zip文件到/usr/local/bigdatacase/dataset目录
[hadoop@master ~]$  unzip /opt/software/user.zip -d /usr/local/src/bigdatacase/dataset
#查看/usr/local/bigdatacase/dataset目录下文件
[hadoop@master ~]$  ls /usr/local/src/bigdatacase/dataset
#在dataset目录下有两个文件：raw_user.csv和small_user.csv。取出前面5条记录看一下
[hadoop@master ~]$ head -5 raw_user.csv

#删除small_user中的第1行
[hadoop@master ~]$ sed -i '1d' /usr/local/src/bigdatacase/dataset/small_user.csv
#上面的1d表示删除第1行，同理，3d表示删除第3行，nd表示删除第n行

#下面再用head命令去查看文件的前5行记录，就看不到字段名称这一行了
[hadoop@master ~]$ head -5 /usr/local/src/bigdatacase/dataset/small_user.csv
```

02

数据集的预处理

数据集的预处理

1.下面要建一个脚本文件pre_deal.sh，请把这个脚本文件放在dataset目录下，和数据集small_user.csv放在同一个目录下：

```
[hadoop@master ~]$  vi /usr/local/src/bigdatacase/dataset
/pre_deal.sh

#!/bin/bash
#下面设置输入文件，把用户执行pre_deal.sh命令时提供的第一个参数作为输入文件名称
infile=$1
#下面设置输出文件，把用户执行pre_deal.sh命令时提供的第二个参数作为输出文件名称
outfile=$2
#注意，最后的$infile> $outfile必须跟在}’ 这两个字符的后面
awk -F "," 'BEGIN{
srand();
        id=0;
        Province[0]="山东";Province[1]="山西";Province[2]="河南";Province[3]="河北";Province[4]="陕西";Province[5]="内蒙古";Province[6]="上海市";
        Province[7]="北京市";Province[8]="重庆市";Province[9]="天津市";Province[10]="福建";Province[11]="广东";Province[12]="广西";Province[13]="云南";
        Province[14]="浙江";Province[15]="贵州";Province[16]="新疆";Province[17]="西藏";Province[18]="江西";Province[19]="湖南";Province[20]="湖北";
        Province[21]="黑龙江";Province[22]="吉林";Province[23]="辽宁"; Province[24]="江苏";Province[25]="甘肃";Province[26]="青海";Province[27]="四川";
        Province[28]="安徽"; Province[29]="宁夏";Province[30]="海南";Province[31]="香港";Province[32]="澳门";Province[33]="台湾";
    }
    {
        id=id+1;
        value=int(rand()*34);
        print id"\t"$1"\t"$2"\t"$3"\t"$5"\t"substr($6,1,10)"\t"Province[value]
    }' $infile> $outfile
```

数据集的预处理

2.下面就可以执行pre_deal.sh脚本文件，来对small_user.csv进行数据预处理，命令如下：

```
## 进入目录/usr/local/src/bigdatacase/dataset/
[hadoop@master ~]$ cd /usr/local/src/bigdatacase/dataset/
## 执行pre_deal.sh程序，对small_user.csv进行数据预处理
[hadoop@master dataset]$ bash ./pre_deal.sh small_user.csv user_table.txt

## 可以使用head命令查看前10行数据：
[hadoop@master dataset]$ head -10 user_table.txt
```

03

导入数据库

导入数据库

1. 启动Hadoop集群相关组件。

```
##在master节点启动Hadoop
[hadoop@master dataset]$ start-all.sh

## 启动master的zookeeper
[hadoop@master dataset]$ zkServer.sh start
## 启动slave1的zookeeper
[hadoop@slave1 ~]$ zkServer.sh start
## 启动slave2的zookeeper
[hadoop@slave2 ~]$ zkServer.sh start

## 启动HBase
[hadoop@master dataset]$ start-hbase.sh
```

导入数据库

2. 把user_table.txt上传到HDFS中。

```
## 在HDFS的根目录下面创建一个新的目录bigdatacase，并在该目录下创建子目录dataset
[hadoop@master dataset]$ hdfs dfs -mkdir -p /bigdatacase/dataset

## Linux本地文件user_table.txt上传到分布式文件系统HDFS的/bigdatacase/dataset目录下
[hadoop@master dataset]$ hdfs dfs -put
/usr/local/src/bigdatacase/dataset/user_table.txt /bigdatacase/dataset

## 查看一下HDFS中的user_table.txt的前10条记录
[hadoop@master dataset]$ hdfs dfs -cat /bigdatacase/dataset/user_table.txt | head -
10
```

导入数据库

3. 在Hive上创建数据库。

首先要确MySQL数据库已经启动。在这个新的终端中执行下面命令进入Hive。

```
## 启动Hive
[root@master ~]$ hive

## 需要在Hive中创建一个数据库dblab，命令如下：
hive> create database dblab;
## 使用dblab数据库
hive> use dblab;
```

导入数据库

4. 创建外部表。

在hive命令提示符下输入如下命令。

```
hive>  CREATE EXTERNAL TABLE dblab.bigdata_user(id INT,uid
STRING,item_id STRING,behavior_type INT,item_category
STRING,visit_date DATE,province STRING) COMMENT 'Welcome to
xmudblab!' ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE LOCATION '/bigdatacase/dataset';
```

导入数据库

5. 查询数据。

在hive命令提示符下执行下面命令查看表的信息。

```
## 显示数据库中所有表
hive> show tables;
## 查看bigdata_user表的各种属性;
hive> show create table bigdata_user;
## 查看表的简单结构
hive> desc bigdata_user;

## 查看10条记录
hive> select * from bigdata_user limit 10;
## 查看10条记录中behavior_type字段的信息
hive> select behavior_type from bigdata_user limit 10;
```

Turing AI 万维图灵 | 大数据系列课程

大数据

BIG DATA

智 / 能 / 科 / 技　　放 / 眼 / 未 / 来