# 数据预处理、存储、查询

**【实验名称】**

数据预处理、存储、查询

**【实验目的】**

掌握HDFS的基本操作、、数据仓库Hive的安装与基本操作、基本的数据预处理方法

**【实验要求】**

 本地数据集上传到数据仓库Hive

**【实验设备】**

基于第10课在hadoop1主节点上进行操作

**【实验步骤】**

## 本地数据集上传到数据仓库Hive

### 1.1实验数据集解压

实验数据集存在于/opt/hadoop-package中的user.zip数据集



#### 1.1.1创建一个存放数据集的文件夹

[root@hadoop1 ~]# mkdir -p /usr/local/bigdatacase/dataset

#### 1.1.2进入/opt/hadoop-package目录，将user.zip数据集解压到刚刚创建的文件夹中

[root@hadoop1 ~]# cd /opt/hadoop-package/

[root@hadoop1 hadoop-package]# unzip user.zip -d /usr/local/bigdatacase/dataset/



#### 1.1.3查看dataset文件夹下面的文件

[root@hadoop1 hadoop-package]# cd /usr/local/bigdatacase/dataset/

[root@hadoop1 dataset]# ll

我们可以看到dataset文件夹中有raw\_user.scv和small\_user.csv两个数据集



#### 1.1.4查看raw\_user.csv数据集的前10条记录

在dataset文件夹下我们来查看raw\_user.csv数据集的前10条记录

[root@hadoop1 dataset]# head -10 raw\_user.csv



## 2.数据集的预处理

### 2.1删除raw\_user.csv和small\_user.csv数据集中的第1行记录

进入dataset文件夹中

[root@hadoop1 ~]# cd /usr/local/bigdatacase/dataset/

删除raw\_user.csv中的第1行记录

[root@hadoop1 dataset]# sed -i '1d' raw\_user.csv

删除small\_user.csv中的第1行记录

[root@hadoop1 dataset]# sed -i '1d' small\_user.csv



### 2.2对字段进行预处理

进入dataset文件夹中

[root@hadoop1 ~]# cd /usr/local/bigdatacase/dataset/

使用vi编辑器新建一个pre\_deal.sh脚本文件

[root@hadoop1 dataset]# vi pre\_deal.sh

在这个脚本文件下加入下面代码：

#!/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

加入代码如图：



执行pre\_deal.sh脚本文件，对small\_user.csv进行数据预处理并将预处理后生成的结果保存为user\_table.txt

[root@hadoop1 dataset]# bash ./pre\_deal.sh small\_user.csv user\_table.txt

查看生成的user\_table.txt的前10行数据

[root@hadoop1 dataset]# head -10 user\_table.txt



## 3.上传到HDFS文件系统

### 3.1先查看HDFS文件系统是否启动

执行jps命令查看HDFS文件系统是否启动，如无启动请先启动HDFS文件系统

[root@hadoop1 ~]# jps



### 3.2.在HDFS上创建/bigdatacase/dataset文件夹并将本地系统dataset文件夹下的user\_table.txt上传到刚创建的dataset中

在HDFS上创建/bigdatacase/dataset文件夹

[root@hadoop1 ~]# hdfs dfs -mkdir -p /bigdatacase/dataset

将本地系统中dataset文件夹下的user\_table.txt到HDFS文件系统中

[root@hadoop1 ~]# hdfs dfs -put /usr/local/bigdatacase/dataset/user\_table.txt /bigdatacase/dataset



**1.3.3在HDFS文件系统中查看user\_table.txt的前10条记录**

[root@hadoop1 ~]# hdfs dfs -cat /bigdatacase/dataset/user\_table.txt | head -10



至此，本节实验结束！