实验 3-1 Hive 加载电商源数据

建议课时： 60 分钟

一、 实验目的  
λ 掌握 hive 建库、 建表语句；

λ 掌握 hive 加载数据的几种方式；

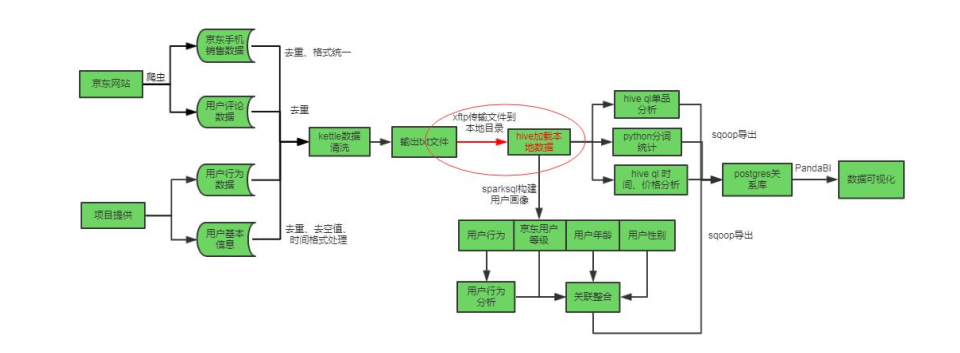
λ 熟练 hive 表数据的常用查询语句；

二、 实验环境

Dsight 智慧实验室中的 hadoop 环境

三、 实验步骤

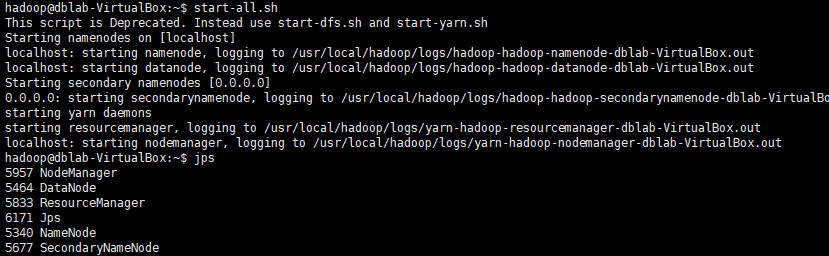
本节实验所做内容如下红色标注：



本节实验主要是通过Hive 加载本地源数据到 hdfs 分布式文件系统。

具体实验步骤如下：

1. 进入实验室，



2. 启动hive 进程



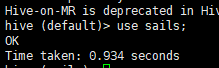
3. 创建数据库

创建自己的数据库sails，

create database zhaoteng;



4. 使用数据库



5. 创建表

创建用户基本信息表、手机销售信息表、用户评价表、用户行为表。

①用户基本信息表：

create table user\_action\_tb(

user\_id string,

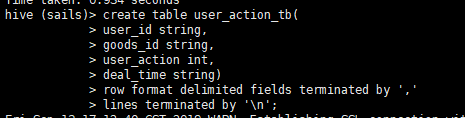
goods\_id string,

user\_action int,

deal\_time string)

row format delimited fields terminated by ','

lines terminated by '\n';



②手机销售信息表：

create table goods\_sail\_tb(

goods\_id string,

goods\_name string,

goods\_property string,

store\_name string,

stroe\_id string,

goods\_url string,

goods\_price float,

keyword string,

sail\_count int,

good\_rate int,

brand string,

model string,

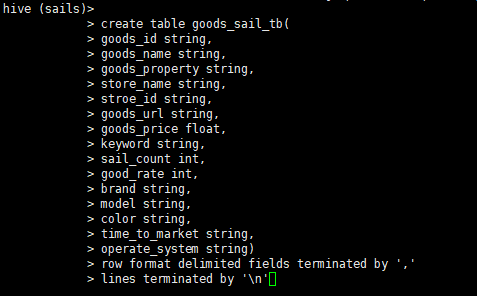
color string,

time\_to\_market string,

operate\_system string)

row format delimited fields terminated by ','

lines terminated by '\n';



③用户评价表

create table user\_comment\_tb(

goods\_id string,

comment string,

releasetime string ,

user\_id string,

user\_name string,

user\_level string,

goods\_color string,

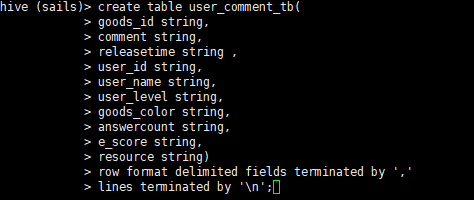
answercount string,

e\_score string,

resource string)

row format delimited fields terminated by ','

lines terminated by '\n';



④用户行为表

建表语句：

create table user\_action(

user\_id string,

goods\_id string,

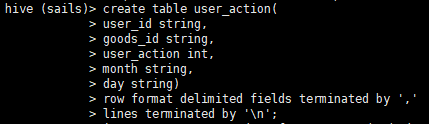
user\_action int,

month string,

day string)

row format delimited fields terminated by ','

lines terminated by '\n';



1. 加载数据

通过hive 加载本地数据的方式依次加载上述表中的源数据。

①加载用户信息数据：

load data local inpath '/home/hadoop/data/data1/userinfo.txt' into table user\_info\_tb;

②加载用户行为数据：

load data local inpath '/home/hadoop/data/data1/action.txt' into table user\_action;

③加载用户评价表：

load data local inpath '/home/hadoop/data/data1/comment.txt' into table user\_comment\_tb;

④加载手机销售信息表：

load data local inpath '/home/hadoop/data/data1/iphone.txt' into table goods\_sail\_tb;

7. 查询加载的数据

出现了中文乱码的情况，数据库建的时候默认为utf-8，在windows系统中中创建的文件，默认编码是GBK，所以将文件重新保存为utf-8，即可成功导入正确的数据。

8. 处理手机销售信息表数据

由于爬虫过程是通过关键词搜索获取的，所获取数据不仅包含手机的销售数据，此外，关于手机的部分外设（充电宝、数据线、手机膜、保护壳、耳机等）销售数据，需要将这部分数据清洗掉。观察表中数据可看出，所有有关手机的操作系统的列字段又有信息，所以可根据手机销售信息表中的操作系统字段进行筛选即，使用操作系统非空和商品name的内容包括‘手机’字段合并查询。

筛选手机的销售数据存储到新表goods\_sail\_info 中。

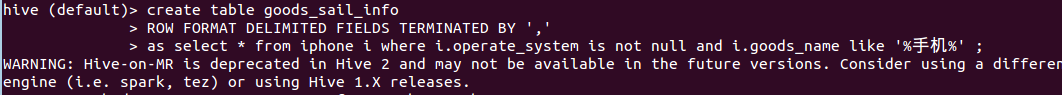
语句：

create table goods\_sail\_info

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

as select \* from iphone i where i.operate\_system is not null and i.goods\_name like '%手机%' ;

截图：



9. 处理用户行为表数据

要求：将用户行为源表中交易月份与交易日拼接为新字段存储到新表。

思想：使用concat函数以"-"拼接交易月份和交易日这两个字段，生成新的字段deal\_time

新的用户行为表

create table user\_action\_tb(

user\_id string,

goods\_id string,

user\_action int,

deal\_time string)

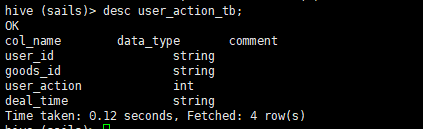
row format delimited fields terminated by ','

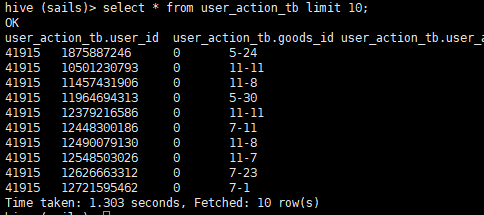
lines terminated by '\n';

导入数据

insert into table user\_action\_tb select user\_id, goods\_id, user\_action, concat(month,'-',day) from user\_action;

截图：





10.处理用户信息表数据

要求：关联用户评论表和用户信息表将用户等级、用户年龄段划分后存

入新表。

1表示年龄<18

2表示年龄在[18,24]

3表示年龄在[25,29]

4表示年龄在[30,34]

5表示年龄在[35,39]

6表示年龄在[40,49]

7表示年龄>=50

语句实现：

创建中间表

create table user\_info\_mid as

select i.user\_id,i.user\_name,i.addr,i.gender,

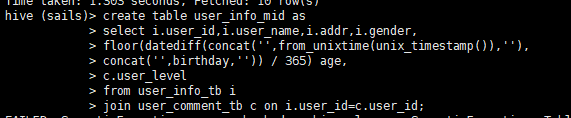
floor(datediff(concat('',from\_unixtime(unix\_timestamp()),''),

concat('',birthday,'')) / 365) age,

c.user\_level

from user\_info\_tb i

join user\_comment\_tb c on i.user\_id=c.user\_id;



分年龄段导入user\_info\_new\_tb表

insert into user\_info\_new\_tb select user\_id,user\_name,addr,gender,

case when age>=0 and age<=17 then 1

when age >= 18 and age <=24 then 2

when age >= 25 and age <= 29 then 3

when age>= 30 and age <= 34 then 4

when age>= 35 and age <= 39 then 5

when age>=40 and age<=49 then 6

when age>=50 then 7 end as age\_region,

case when age>=0 and age<=17 then "18岁以下"

when age>=18 and age<=24 then "[18,24]"

when age>=25 and age<=29 then "[25,29]"

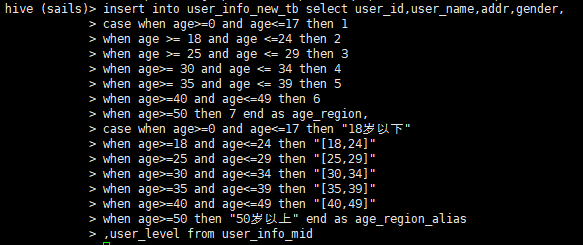
when age>=30 and age<=34 then "[30,34]"

when age>=35 and age<=39 then "[35,39]"

when age>=40 and age<=49 then "[40,49]"

when age>=50 then "50岁以上" end as age\_region\_alias

,user\_level from user\_info\_mid



查询数据

