实验 3-1 Hive 加载电商源数据

建议课时：60 分钟

一、实验目的

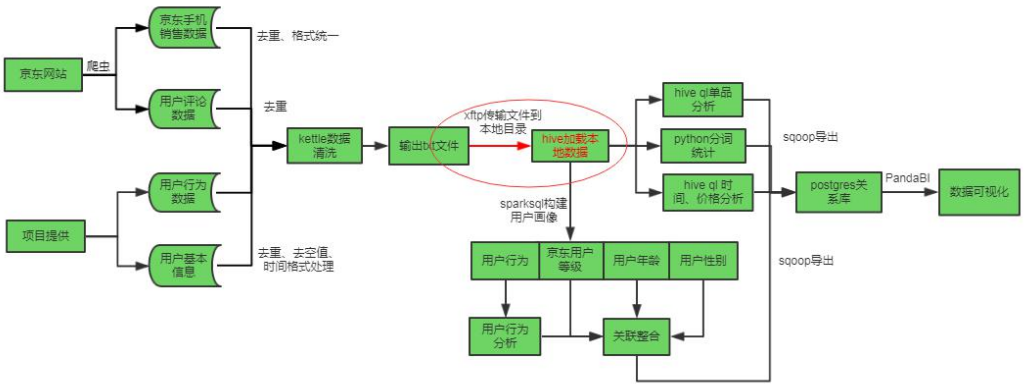
- 掌握 hive 建库、建表语句；
- 掌握 hive 加载数据的几种方式；
- 熟练 hive 表数据的常用查询语句；

二、实验环境

Dsight 智慧实验室中的 hadoop 环境

三、实验步骤

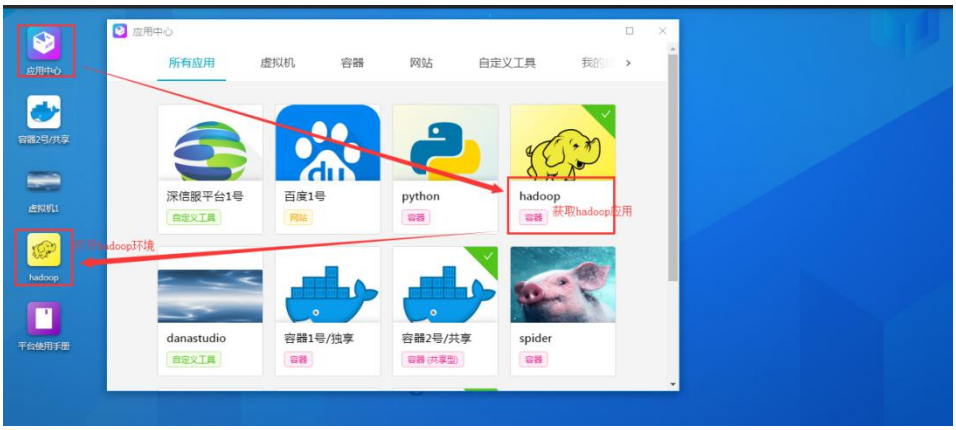
本节实验所做内容如下红色标注：



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：

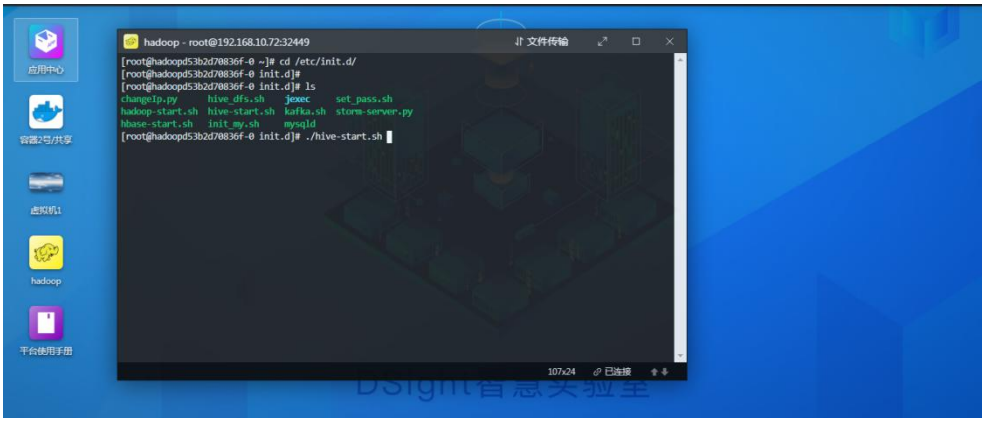
1. 进入实验室，打开 hadoop 环境



2. 启动 hive 进程

第一次进入实验环境时启动方式：cd /etc/init.d 运 行./hive-start.sh

之后进入实验室环境时 输入 hive，回车即可。



3. 创建数据库

create database databasename;

建议创建自己的数据库，默认使用 default 数据也可。

4. 使用数据库

use databasename;

创建数据库后，需要使用数据库才能在该数据库下进行建表、加载数据等后续操作。如果没有使用数据库，默认使用 default 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。

6. 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。

7. 查询加载的数据

通过已学的查询语句查询加载后的数据，保证加载数据能正常显示，若出现乱码及 时修改数据编码格式。

8. 处理手机销售信息表数据

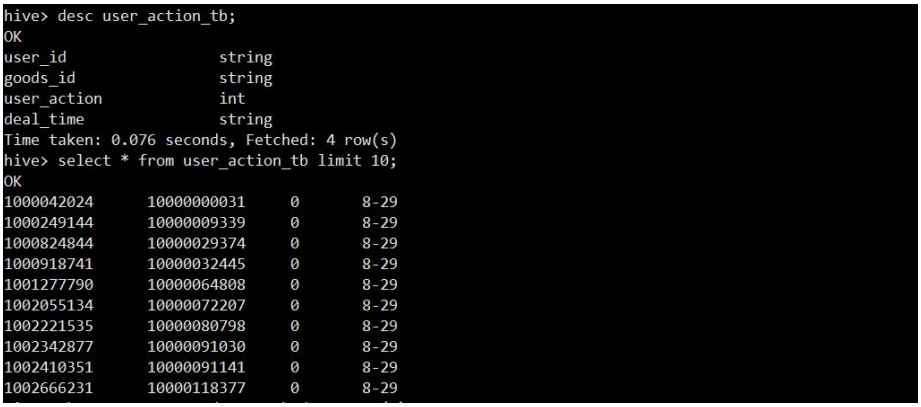
说明：由于爬虫过程是通过关键词搜索获取的，所获取数据不仅包含手机的销售数 据，此外，关于手机的部分外设（充电宝、数据线、手机膜、保护壳、耳机等）销 售数据，需要将这部分数据清洗掉。可根据手机销售信息表中的操作系统字段筛选。

要求：筛选手机的销售数据存储到新表。

```
hive> desc goods_sail_info;
OK
goods_id                string
goods_name              string
goods_property          string
store_name              string
store_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
Time taken: 0.068 seconds, Fetched: 15 row(s)
hive> select * from goods_sail_info limit 5;
OK
1001521853      【Gamevice手柄礼盒套装】ROG电竞游戏手机 8GB+128GB 黑色 骁龙845 冷凝散热 酷炫灯效 全面屏 全网通4G双卡双待        6.0英
寸^8GB^128GB^1200万+800万像素^骁龙845(SDM845)^800万像素^2160*1080^8.65                ROG京东自营官方旗舰店          1000107961  h
ttps://item.jd.com/100004545822.html    5999.0  ROG             8515    96              ROG PHONE               黑色            2018
年9月           Android
```

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。



```
hive> desc user_action_tb;
OK
user_id                 string
goods_id                string
user_action             int
deal_time               string
Time taken: 0.076 seconds, Fetched: 4 row(s)
hive> select * from user_action_tb limit 10;
OK
1000042024      10000000031     0       8-29
1000249144      10000009339     0       8-29
1000824844      10000029374     0       8-29
1000918741      10000032445     0       8-29
1001277790      10000064808     0       8-29
1002055134      10000072207     0       8-29
1002221535      10000080798     0       8-29
1002342877      10000091030     0       8-29
1002410351      10000091141     0       8-29
1002666231      10000118377     0       8-29
```

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

新表数据字段说明：

```
hive> desc user_info_new_tb;
OK
user_id                 string
user_name               string
addr                    string
gender                  string
age_region              int
age_region_alias        string
user_grade              string
Time taken: 0.054 seconds, Fetched: 7 row(s)
hive> select * from user_info_new_tb limit 5;
OK
1002719737      廉***巧         新疆 吐鲁番     女      1       18岁以下        铜牌会员
1002857045      i***T           北京 朝阳区     女      1       18岁以下        银牌会员
1003523670      c***5           海外 土耳其     女      1       18岁以下        金牌会员
1004745095      o***Y           云南 普洱       女      1       18岁以下        铜牌会员
1004853823      韩***9          其他            男      1       18岁以下        钻石会员
Time taken: 0.159 seconds, Fetched: 5 row(s)
```

四、实验成果

本次实验完成后，需要得到以下结果：

- 创建自己的数据库；
- 创建用户基本信息表并加载数据；
- 创建京东手机销售信息表并加载数据；
- 创建用户评价信息表并加载数据；
- 创建用户基本行为信息表并加载数据；
- 按要求处理手机销售表、用户信息表、用户行为表的数据；