**分析师笔试题-李乔乔**

**必做题：SQL题**

时间30分钟

**数据集：**

**T1：订单表，数据量每天2千万左右**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **下单时间order\_date** | **用户id user\_id** | **订单id order\_id** | **商品id goods\_id** | **购买数量buy\_num** | **购买价格buy\_price** |
| 20190101 | a000001 | ORDER01111 | G010101 | 3 | 28 |
| …… | …… | …… | …… | …… | …… |

**T2：商品表：数据量1千万左右**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **主键**  **key\_id** | **商品ID goods\_id** | **商品名称goods\_name** | **商品分类catgroy** | **商品单价price** | **入库时间p\_date** |
|  | G010101-20190101 | G010101 | 鼠标 | 数码产品 | 30 | 20190101 |
|  | G010102-20190101 | G010102 | 面粉 | 食品 | 20 | 20190101 |
|  | …… | …… | …… | …… | …… | …… |

**T3：用户登陆日志：数据量十亿级左右**

|  |  |  |
| --- | --- | --- |
| 用户ID  user\_id | 登陆时间  login\_datetime | 登陆IP  login\_ip |
| u12345 | 2020-01-01 12:09:34 | 192.168.1.1 |
| …… | …… | …… |

**题目：**

**1、统计近30天、60天每一个用户的订单量**

|  |
| --- |
| SQL  select user\_id count(order\_id) from 订单表  where order\_date between 20211022 and 20211122  group by user\_id ;  select user\_id count(order\_id) from 订单表  where order\_date between 20210922 and 20211122  group by user\_id ; |

**2、统计数码产品近半年的总的折扣比例**

|  |
| --- |
| SQL  select **g**oods\_id, goods\_name,(buy\_price/price) 折扣比例 from 订单表  left join 商品表 on 订单表.goords\_id=商品表.goods\_id  where order\_date between 20210522 and 20211122; |

**3、针对数据集T3，计算生成如下表数据**

|  |  |  |  |
| --- | --- | --- | --- |
| user\_id | 最近一次登陆ip | 最近一次登陆时间 | 近7日登陆ip数 |

|  |
| --- |
| SQL  select uesr\_id , min(login\_datetime),  (select login\_ip from T3 where date(login\_datetime )=min(date(login\_datetime))),  (  select count(distinct login\_ip)  where (date(login\_datetime) between 20211115 and 20211122 )  and (a. uesr\_id=b. uesr\_id )  from T3 b group by user\_id  )  from T3 a  group by user\_id; |

**选做题：python题**

时间20分钟，

**数据集：**

**data.csv**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Id | MSSubClass | MSZoning | LotFrontage | LotArea |
| 1 | 60 | RL | 65 | 8450 |
| 2 | 20 | RL | 80 | 9600 |
| 3 | 60 | RL | 68 | 11250 |
| 4 | 70 | RL | 60 | 9550 |
| 5 | 60 | RL | 84 | 14260 |

**题目：**

**1、统计上表中各个字段的缺失值数量在本列的占比**

|  |
| --- |
| Python  import pandas as pd  a=pd.read\_csv(‘data.csv’)  def missing\_percent(df):  nan\_percent = 100\*(df.isnull().sum()/len(df))  #nan\_percent = nan\_percent[nan\_percent > 0].sort\_values() 排序可以  return nan\_percent  print(missing\_percent(a)) |

**2、以MSZoning字段做分组计数统计**

|  |
| --- |
| Python  grouped =a.groupby('MSZoning').count()  grouped |