### HQL 第1题

表结构：uid,subject\_id,score

求：找出所有科目成绩都大于某一学科平均成绩的学生

数据集如下

1001 01 90

1001 02 90

1001 03 90

1002 01 85

1002 02 85

1002 03 70

1003 01 70

1003 02 70

1003 03 85

1）建表语句

create table score(

uid string,

subject\_id string,

score int)

row format delimited fields terminated by '\t';

2）求出每个学科平均成绩

select

uid,

score,

avg(score) over(partition by subject\_id) avg\_score

from

score;t1

3）根据是否大于平均成绩记录flag，大于则记为0否则记为1

select

uid,

if(score>avg\_score,0,1) flag

from

t1;t2

4）根据学生id进行分组统计flag的和，和为0则是所有学科都大于平均成绩

select

uid

from

t2

group by

uid

having

sum(flag)=0;

5）最终SQL

select

uid

from

(select

uid,

if(score>avg\_score,0,1) flag

from

(select

uid,

score,

avg(score) over(partition by subject\_id) avg\_score

from

score)t1)t2

group by

uid

having

sum(flag)=0;

### HQL 第2题

我们有如下的用户访问数据

|  |  |  |
| --- | --- | --- |
| userId | visitDate | visitCount |
| u01 | 2017/1/21 | 5 |
| u02 | 2017/1/23 | 6 |
| u03 | 2017/1/22 | 8 |
| u04 | 2017/1/20 | 3 |
| u01 | 2017/1/23 | 6 |
| u01 | 2017/2/21 | 8 |
| U02 | 2017/1/23 | 6 |
| U01 | 2017/2/22 | 4 |

要求使用SQL统计出每个用户的累积访问次数，如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 用户id | 月份 | 小计 | 累积 |
| u01 | 2017-01 | 11 | 11 |
| u01 | 2017-02 | 12 | 23 |
| u02 | 2017-01 | 12 | 12 |
| u03 | 2017-01 | 8 | 8 |
| u04 | 2017-01 | 3 | 3 |

数据集

u01 2017/1/21 5

u02 2017/1/23 6

u03 2017/1/22 8

u04 2017/1/20 3

u01 2017/1/23 6

u01 2017/2/21 8

u02 2017/1/23 6

u01 2017/2/22 4

1）创建表

create table action

(userId string,

visitDate string,

visitCount int)

row format delimited fields terminated by "\t";

2）修改数据格式

select

userId,

date\_format(regexp\_replace(visitDate,'/','-'),'yyyy-MM') mn,

visitCount

from

action;t1

3）计算每人单月访问量

select

userId,

mn,

sum(visitCount) mn\_count

from

t1

group by

userId,mn;t2

4）按月累计访问量

select

userId,

mn,

mn\_count,

sum(mn\_count) over(partition by userId order by mn)

from t2;

5）最终SQL

select

userId,

mn,

mn\_count,

sum(mn\_count) over(partition by userId order by mn)

from

( select

userId,

mn,

sum(visitCount) mn\_count

from

(select

userId,

date\_format(regexp\_replace(visitDate,'/','-'),'yyyy-MM') mn,

visitCount

from

action)t1

group by userId,mn)t2;

### HQL 第3题

有50W个京东店铺，每个顾客访客访问任何一个店铺的任何一个商品时都会产生一条访问日志，访问日志存储的表名为Visit，访客的用户id为user\_id，被访问的店铺名称为shop，请统计：

1）每个店铺的UV（访客数）

2）每个店铺访问次数top3的访客信息。输出店铺名称、访客id、访问次数

数据集

u1 a

u2 b

u1 b

u1 a

u3 c

u4 b

u1 a

u2 c

u5 b

u4 b

u6 c

u2 c

u1 b

u2 a

u2 a

u3 a

u5 a

u5 a

u5 a

1）建表

create table visit(user\_id string,shop string) row format delimited fields terminated by '\t';

2）每个店铺的UV（访客数）

select shop,count(distinct user\_id) from visit group by shop;

3）每个店铺访问次数top3的访客信息。输出店铺名称、访客id、访问次数

（1）查询每个店铺被每个用户访问次数

select shop,user\_id,count(\*) ct

from visit

group by shop,user\_id;t1

（2）计算每个店铺被用户访问次数排名

select shop,user\_id,ct,rank() over(partition by shop order by ct) rk

from t1;t2

（3）取每个店铺排名前3的

select shop,user\_id,ct

from t2

where rk<=3;

（4）最终SQL

select

shop,

user\_id,

ct

from

(select

shop,

user\_id,

ct,

rank() over(partition by shop order by ct) rk

from

(select

shop,

user\_id,

count(\*) ct

from visit

group by

shop,

user\_id)t1

)t2

where rk<=3;

### HQL 第4题

已知一个表STG.ORDER，有如下字段:Date，Order\_id，User\_id，amount。请给出sql进行统计:数据样例:2017-01-01,10029028,1000003251,33.57。

1）给出 2017年每个月的订单数、用户数、总成交金额。

2）给出2017年11月的新客数(指在11月才有第一笔订单)

建表

create table order\_tab(dt string,order\_id string,user\_id string,amount decimal(10,2)) row format delimited fields terminated by '\t';

1）给出 2017年每个月的订单数、用户数、总成交金额。

select

date\_format(dt,'yyyy-MM'),

count(order\_id),

count(distinct user\_id),

sum(amount)

from

order\_tab

where

date\_format(dt,'yyyy')='2017'

group by

date\_format(dt,'yyyy-MM');

2）给出2017年11月的新客数(指在11月才有第一笔订单)

select

count(user\_id)

from

order\_tab

group by

user\_id

having

date\_format(min(dt),'yyyy-MM')='2017-11';

### HQL 第5题

有日志如下，请写出代码求得所有用户和活跃用户的总数及平均年龄。（活跃用户指连续两天都有访问记录的用户）日期 用户 年龄

数据集

2019-02-11,test\_1,23

2019-02-11,test\_2,19

2019-02-11,test\_3,39

2019-02-11,test\_1,23

2019-02-11,test\_3,39

2019-02-11,test\_1,23

2019-02-12,test\_2,19

2019-02-13,test\_1,23

2019-02-15,test\_2,19

2019-02-16,test\_2,19

1）建表

create table user\_age(dt string,user\_id string,age int)row format delimited fields terminated by ',';

2）按照日期以及用户分组，按照日期排序并给出排名

select

dt,

user\_id,

min(age) age,

rank() over(partition by user\_id order by dt) rk

from

user\_age

group by

dt,user\_id;t1

3）计算日期及排名的差值

select

user\_id,

age,

date\_sub(dt,rk) flag

from

t1;t2

4）过滤出差值大于等于2的，即为连续两天活跃的用户

select

user\_id,

min(age) age

from

t2

group by

user\_id,flag

having

count(\*)>=2;t3

5）对数据进行去重处理（一个用户可以在两个不同的时间点连续登录），例如：a用户在1月10号1月11号以及1月20号和1月21号4天登录。

select

user\_id,

min(age) age

from

t3

group by

user\_id;t4

6）计算活跃用户（两天连续有访问）的人数以及平均年龄

select

count(\*) ct,

cast(sum(age)/count(\*) as decimal(10,2))

from t4;

7）对全量数据集进行按照用户去重

select

user\_id,

min(age) age

from

user\_age

group by

user\_id;t5

8）计算所有用户的数量以及平均年龄

select

count(\*) user\_count,

cast((sum(age)/count(\*)) as decimal(10,1))

from

t5;

9）将第5步以及第7步两个数据集进行union all操作

select

0 user\_total\_count,

0 user\_total\_avg\_age,

count(\*) twice\_count,

cast(sum(age)/count(\*) as decimal(10,2)) twice\_count\_avg\_age

from

(

select

user\_id,

min(age) age

from

(select

user\_id,

min(age) age

from

(

select

user\_id,

age,

date\_sub(dt,rk) flag

from

(

select

dt,

user\_id,

min(age) age,

rank() over(partition by user\_id order by dt) rk

from

user\_age

group by

dt,user\_id

)t1

)t2

group by

user\_id,flag

having

count(\*)>=2)t3

group by

user\_id

)t4

union all

select

count(\*) user\_total\_count,

cast((sum(age)/count(\*)) as decimal(10,1)),

0 twice\_count,

0 twice\_count\_avg\_age

from

(

select

user\_id,

min(age) age

from

user\_age

group by

user\_id

)t5;t6

10）求和并拼接为最终SQL

select

sum(user\_total\_count),

sum(user\_total\_avg\_age),

sum(twice\_count),

sum(twice\_count\_avg\_age)

from

(select

0 user\_total\_count,

0 user\_total\_avg\_age,

count(\*) twice\_count,

cast(sum(age)/count(\*) as decimal(10,2)) twice\_count\_avg\_age

from

(

select

user\_id,

min(age) age

from

(select

user\_id,

min(age) age

from

(

select

user\_id,

age,

date\_sub(dt,rk) flag

from

(

select

dt,

user\_id,

min(age) age,

rank() over(partition by user\_id order by dt) rk

from

user\_age

group by

dt,user\_id

)t1

)t2

group by

user\_id,flag

having

count(\*)>=2)t3

group by

user\_id

)t4

union all

select

count(\*) user\_total\_count,

cast((sum(age)/count(\*)) as decimal(10,1)),

0 twice\_count,

0 twice\_count\_avg\_age

from

(

select

user\_id,

min(age) age

from

user\_age

group by

user\_id

)t5)t6;

### HQL 第6题

请用sql写出所有用户中在今年10月份第一次购买商品的金额，表ordertable字段（购买用户：userid，金额：money，购买时间：paymenttime(格式：2017-10-01)，订单id：orderid）

1）建表

create table ordertable(

userid string,

money int,

paymenttime string,

orderid string)

row format delimited fields terminated by '\t';

2）查询出

select

userid,

min(paymenttime) paymenttime

from

ordertable

where

date\_format(paymenttime,'yyyy-MM')='2017-10'

group by

userid;t1

select

t1.userid,

t1.paymenttime,

od.money

from

t1

join

ordertable od

on

t1.userid=od.userid

and

t1.paymenttime=od.paymenttime;

select

t1.userid,

t1.paymenttime,

od.money

from

(select

userid,

min(paymenttime) paymenttime

from

ordertable

where

date\_format(paymenttime,'yyyy-MM')='2017-10'

group by

userid)t1

join

ordertable od

on

t1.userid=od.userid

and

t1.paymenttime=od.paymenttime;