友乾营

专注数据技术的社群

大批量大客群交集运算性能优化

18 期

www.raqsoft.com.cn/yqy

问题描述

用户画像应用广泛

用户数趋势
年龄段分布
工作行业分布
...

客群是客户的分类

客户可属于多个客群

客群交集是指计算多个客群的共同客户，再进行维度过滤，计算出符合条件的客户数。维度包括：性别、年龄段等。

问题描述

友乾营

数据量大

客户数量
千万、上亿

客群众多

客群总数
几千个

客群很大

一个客群包含千万、上亿客户

交集过滤

几个客群求交集之后按维度过滤

客群交集面临的难题：客群众多，包含的客户量巨大。任意几个求交集，无法预先算好，必须实时计算。现有解决方案太慢，无法实现秒级响应。

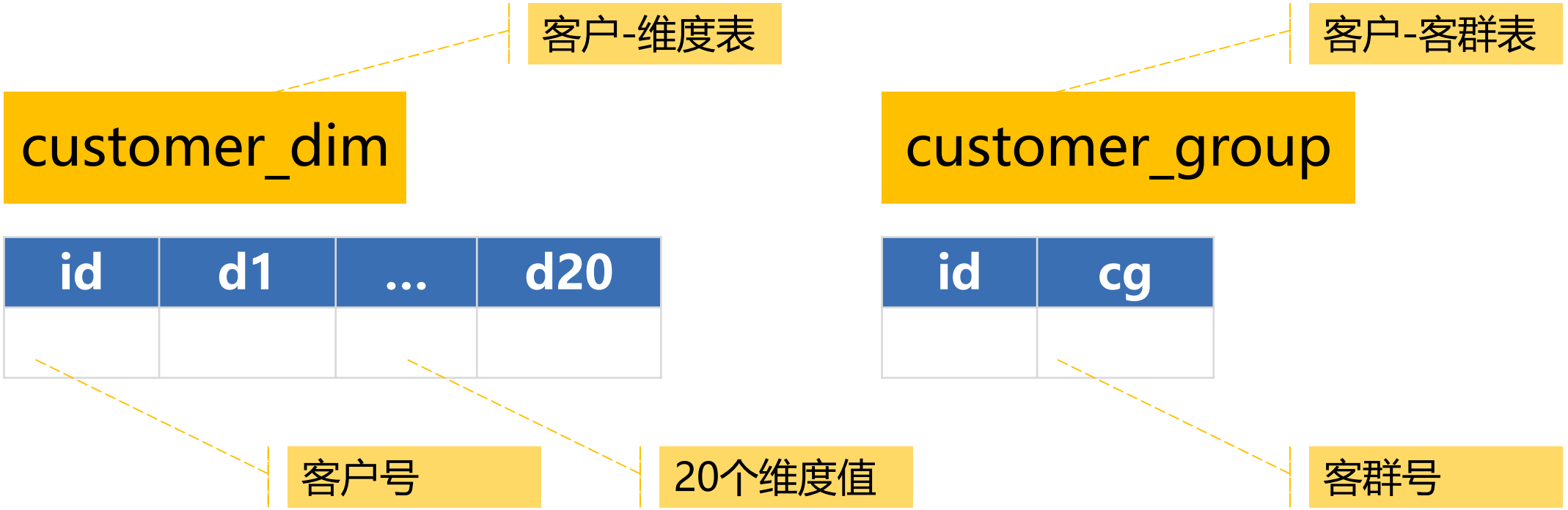
问题描述-维度特征

友乾营

教育程度	年龄分布	行龄分布	性别分布
本科、大专、 硕士、博士 …	25-30岁 30-35岁 35-40岁 …	1-2年 2-3年 3-4年 …	女 男 其他

维度没有层次关系，维度的属性一般是几个到几十个。比如：性别的属性是两三个，年龄段的属性是十几个。参与过滤的维度总数是十到二十个。

问题原型一：两表关联



id是客户编号，cg是客群编号，d1到d20是二十个维度的取值。计算客群交集并过滤的时候，需要做两个大表的join，性能不理想。

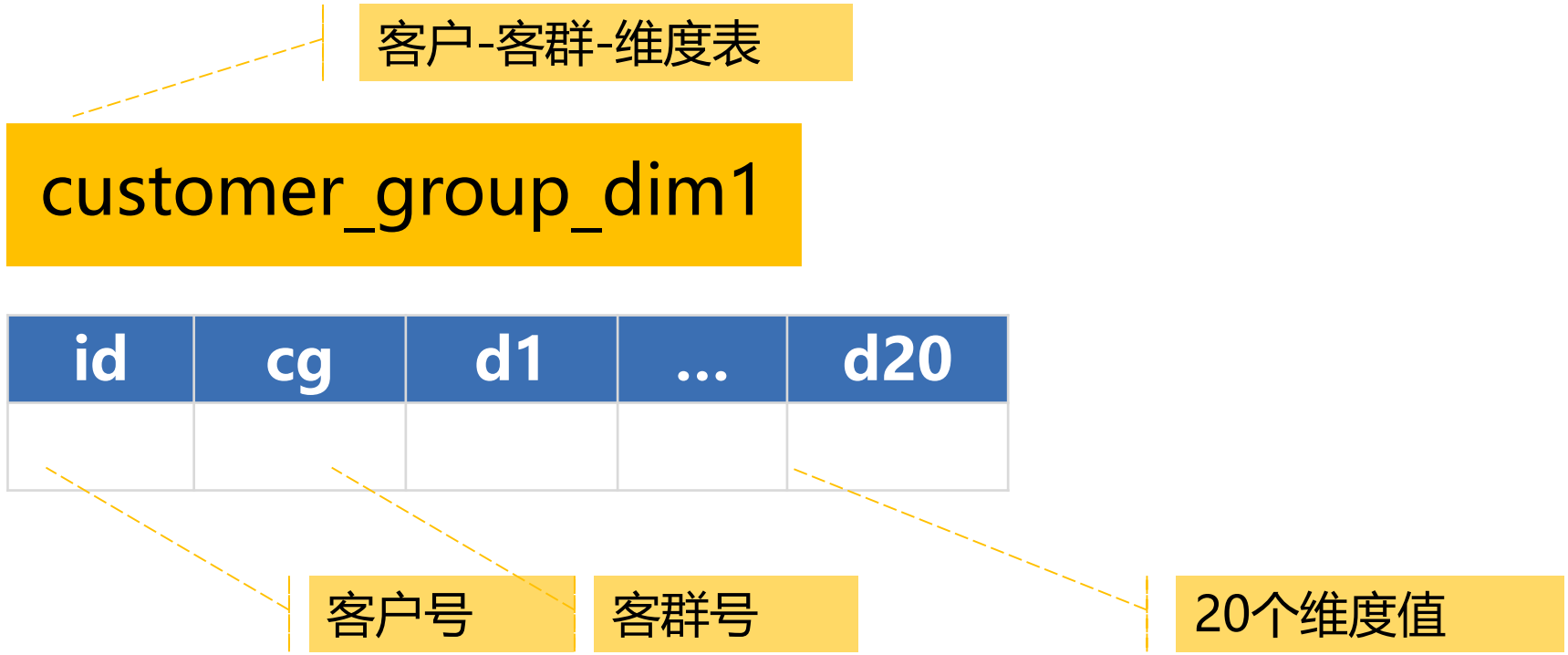
问题原型一：两表关联SQL

友乾营

```
select count(*) from (
        select count(g.cg) from customer_group g
                                left join customer_dim d on g.id=d.id
                where g.cg in ('18','25')
                and d.d2 in ('2','4')
                and d.d4 in ('8','10','11')
                group by g.id
                having count(g.cg)=2
)
```

客户-维度表有几千万到上亿条记录，客户-客群表有几亿到十几亿条记录，两个大表按照客户id做join关联计算，无法达到秒级响应要求。

问题原型二：数据冗余



把两个表合成一个客户-客群-维度表，可以避免大表关联join计算。但是一个客户属于几个到十几个客群，每个客户相同的维度值会大量冗余。

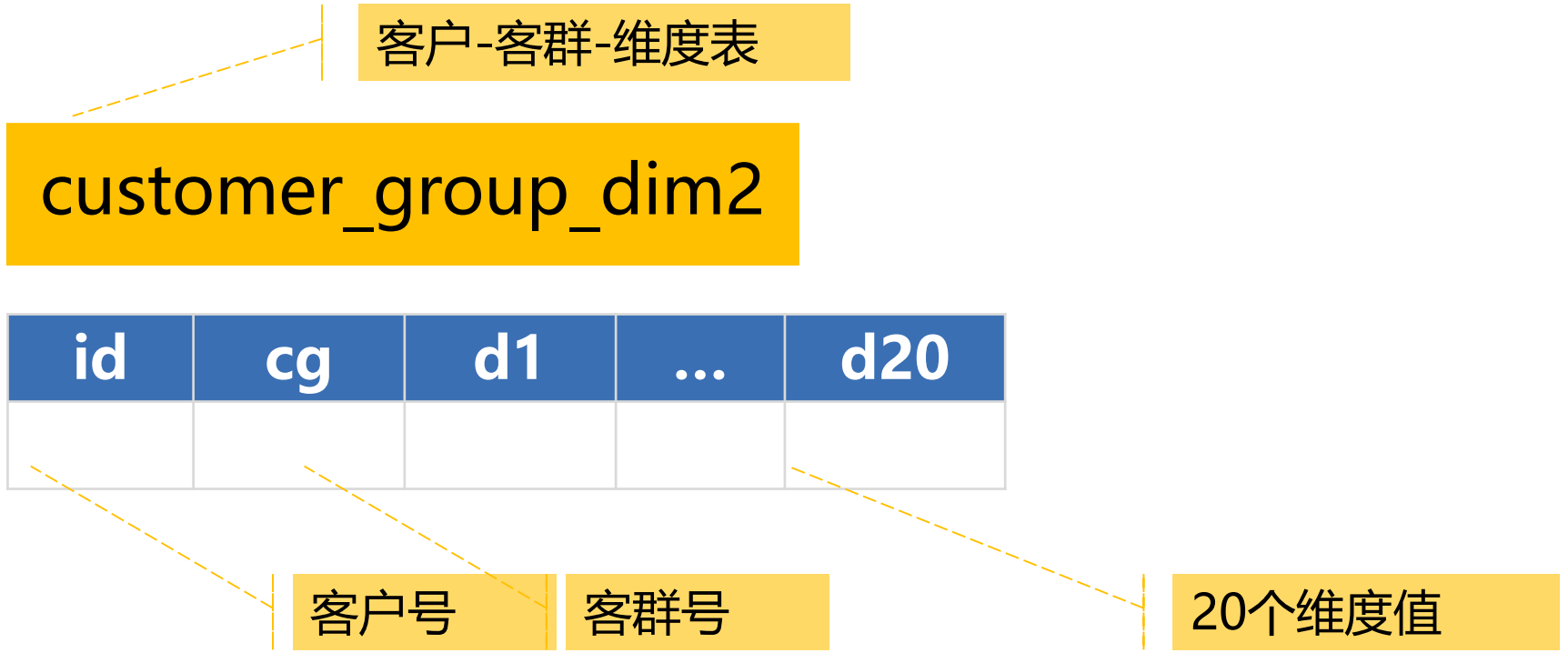
问题原型二：数据冗余SQL

友乾营

```
select count(*) from (
        select count(cg) from customer_group_dim1
                where cg in ('18','25')
                and d2 in ('2','4')
                and d4 in ('8','10','11')
                group by id having count(cg)=2
)
```

一个客户属于几个到十几个客群，相同的维度值会出现十倍以上的冗余。对于行存数据库来说，查询速度会慢很多。

问题原型三：逗号分隔



将多个客群号用逗号分隔字符串存入一个字段cg中，可以避免维度值冗余。例：cg为“18,25,157”表示这个客户属于三个客群。查询速度依然很慢。

问题原型三：逗号分隔SQL

友乾营

```
select count(*) from customer_group_dim2
                where (cg like '18,%' or cg like ',18,' or cg like '%,18’)
                and (cg like '25,%' or cg like ',25,' or cg like '%,25')
                and d2 in ('2','4')
                and d4 in ('8','10','11')
```

按照逗号分隔的字符串不符合数据范式，算交集时要用字符子串比较，性能很差。

解决办法

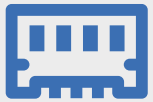
友乾营

专注数据技术的社群

集算器加速

解题思路-内存外存

友乾营



内存

随机存取快
容量有限

硬盘

读写较慢
容量大

客户画像数据量几千万甚至上亿，而且每个月都有一份完整的数据。主数据放到内存中成本太高，要放到硬盘存储。维度表等放到内存存储。

解题思路-改变存储方案

友乾营

主文件

d1	...	d20	c1	...	c300

20个维度值

9000个客群号

新存储方案：字段d1-20，存储维度属性值序号；字段c1-300，是所属客群位图。每个客户一条记录，因为是统计条数，所以无需存储客户号。

解题思路

友乾营

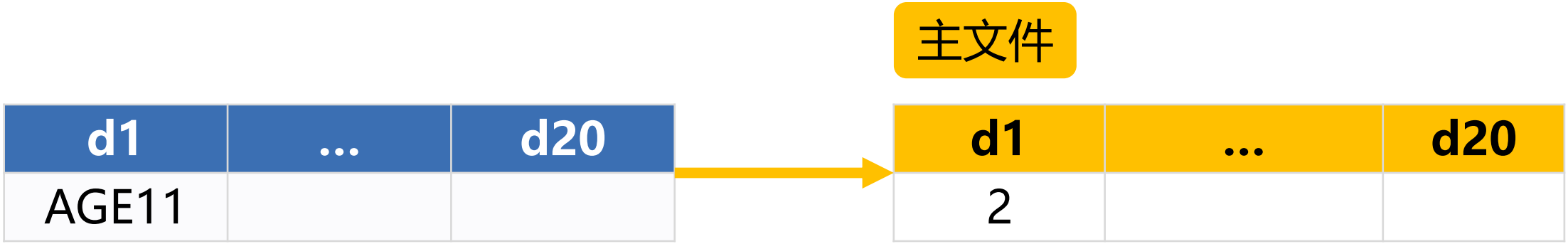
专注数据技术的社群

1

解题关键

维度存储

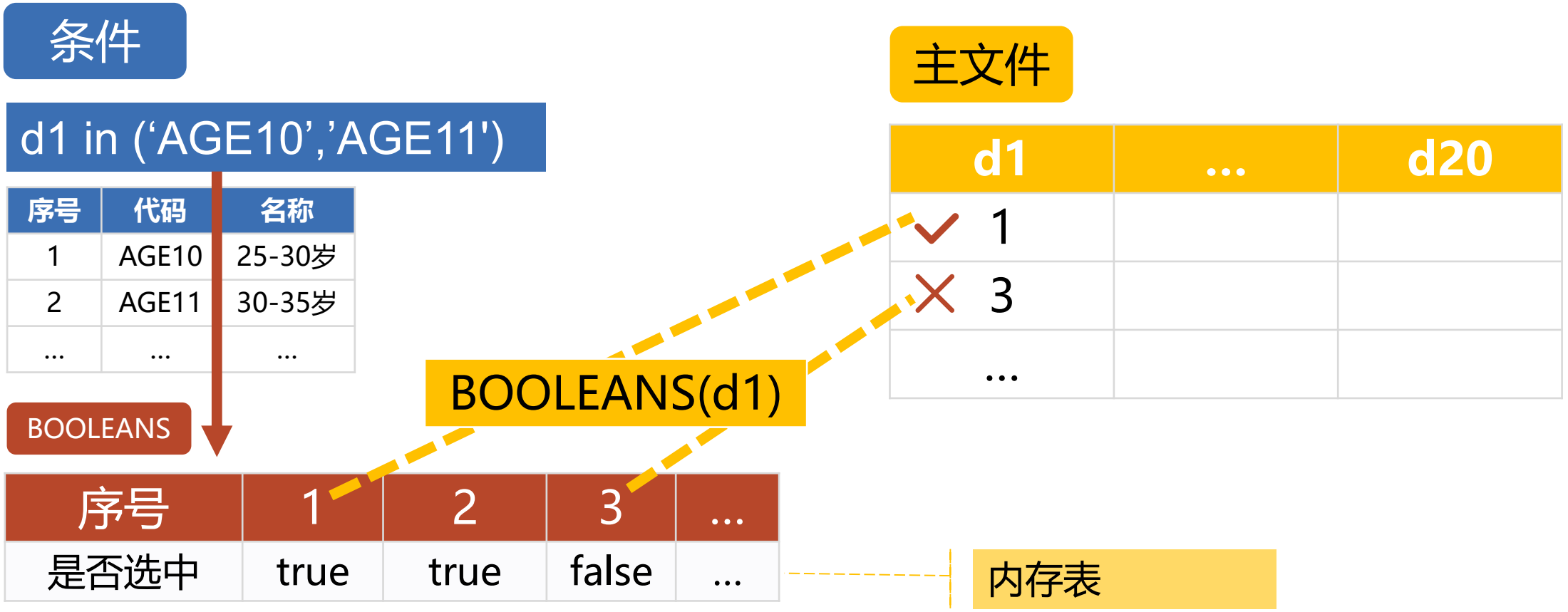
解题关键1-改变维度存储方式



序号	代码	名称
1	AGE10	25-30岁
2	AGE11	30-35岁
...	...	...

年龄段维度d1原存储的是代码‘AGE11’，在代码表中的顺序号是2，因此在主文件中改为存储2。一个客户的维度值用20个整数即可存储。

解题关键1-改变维度过滤算法



根据输入条件和代码表生成内存表BOOLEANS。在主文件查询年龄段1或者2的时候，用取得的d1值1，找到内存表中1对应的是true，因此第1条记录满足条件。

解题思路

友乾营

专注数据技术的社群

2

解题关键

客群存储

解题关键2-改变客群存储方式

主文件

c1	…	c300
0		0

序号	1	2	…	30
是否选中	0	0	…	0

2个客群求交集，最多取2列即可，利用集算器列存特性，大幅度减少需要读取的数据量

每个int32位，用30位存储所属客群。例：c6的第7位为1，表示该客户属于（6-1）×30+7=157号客群。300个整数，共存储9000个客群。

解题关键2-改变客群存储方式

友乾营

原型

id	cg
001	18
001	25
001	157
…	…

主文件

c1	…	c6	…	c300
16908288	0	64	0	0

序号	…	18	…	25	…
是否选中	0	1	0	1	0

…	7	…
0	1	0

客户001属于18、25、157三个客群。因此要把c1第18、25位，c6的第7位设置为1。
计算对应十进制数是16908288和64，存入c1和c6。

解题关键2-改变客群交集算法

友乾营

客群

cg in (‘18’,’25')

序号	…	18	…	25	…
是否选中	0	1	0	1	0

and(c1,16908288)==16908288

主文件

c1	…	c5	…	c300
✓ 16908288	0	64	0	0
✗ 130	0	0	0	0
…	…	…	…	…

求18、25客群的交集，先计算第18、25位为1的int是16908288。依次取主文件c1，和16908288“按位与”，结果还是16908288的符合条件。

解决办法

友乾营

专注数据技术的社群

SPL代码解析

SPL解析：核心代码非常简单

友乾营

```
=file("data/ctx/databitC.ctx").create().cursor
(c1,d2,d8;
and(c1,16908288)==16908288
&&queryBOOLEANS(1).BOOLEANS(d2)
&&queryBOOLEANS(2).BOOLEANS(d8)
).skip()
```

新建文件对象，建立游标

需要读取的字段名

按位与，18、25号客群交集

过滤符合条件的维度

遍历游标，得到结果

在最核心的代码之前，要把传入参数计算为：1、需要的列名；2、按位与，计算客群交集表达式；3、过滤符合条件的维度表达式。

SPL解析：输入参数、常数

友乾营

	A	B
1	/输入参数、常数	
2	=cust_cls="18,25"	
3	=where="d2:2,4;d8:8,10,11"	
4	[2,4,5,5,7,7,9,15,35,40]	/10个维度的取值长度

A2：输入参数，要求交集的两个客群编号。

A3：输入参数，过滤维度和条件，格式是“维度id:性值顺序号,属性值顺序号;...”。

A4：10个维度的取值长度，依次是2、4、5、6...。

SPL解析：准备客群查询条件和列名

	A	B
5	/准备客群查询条件、列名	
6	=to(1,29).new(tmp=to(30).(0),tmp(~)=1,bits(tmp):b).new(b)	=cust_cls.split(",").(int(~))
7	=B6.new(ys=int(~%30),zs=int(~\30),if(ys==0,zs,int(zs+1)):position,if(ys==0,536870912,A6(ys).b):value).new(~.position,~.value)	

A6：计算二十九位的二进制数，每一位为1其他为0时对应的整数，存入序列。例如：0...001对应的是1；0...010对应的是2；1...000对应的是268435456。

B6：将输入参数需要交叉的客群号用逗号分成序列，并转换成整数。

A7：计算需要求交集的客群序列中，每个客群在字段c1到c300中的位置和十进制数值。这里用A6的结果，要比在循环中计算bits要快很多。

SPL解析：准备客群查询条件和列名

	A	B
8	=A7.groups(~.position;iterate(or(~.value,~~),0):value)	
9	=A8.("c"/position)	=A9.concat(",")
10	=A8.("and(c"/position/","/value/")=="/value)	=A10.concat("&&")

A8：合并在同一个位置的十进制数值。

A9：在每个位置前面加上字符c，变成字段名称；B9：把所有的字段名称合并成一个字符串，逗号分隔。

A10：把客群计算成过滤条件表达式，例如：and(c1,3)==3；B10：多个条件用“与”连接。

SPL解析：准备维度查询条件、列名

友乾营

	A	B
11	/准备维度查询条件、列名	
12	=queryBOOLEANS=create(DIMID,BOOLEANS)	=A4.([false]*~)
13	=where.split(";").(~.split(":")).new(~(1):DIMID,~(2).split(",):VALUES)	
14	for A13	=A14.VALUES.(int(~))
15		=int(mid(A14.DIMID,2))
16		=B12(B15)(B14)=true
17		=A12.insert(0,A14.DIMID,B12(B15))

B12：根据10个维度的长度准备一个长度为10的序列，每个成员是对应长度的false序列。

A13：把where条件转换成序列，DIMID是维度的字段名，VALUES是对应的维度值顺序号，如右图：

序号	DIMID	VALUES
1	d2	[2,4]
2	d8	[8,10,11]

A14-B17：把枚举值转成对应位置上的布尔值存入A12准备的内存表，d2、d8的长度是分别是4、15，所以对应的BOOLEANS分别有4、15个成员，和上面的VALUES对应的位置为true，其他为false，如右图：

序号	DIMID	BOOLEANS
1	d2	[false,true,false, ...]
2	d8	[false,false,false, ...]

SPL解析：准备维度查询条件、列名

友乾营

	A	B
18	=A12.("queryBOOLEANS("/#/").BOOLEANS("/DIMID/")")	
19	=A18.concat("&&")	=A12.(DIMID).concat(",")

A18-A19：计算过滤表达式，例如：queryBOOLEANS(1).BOOLEANS(d2)&&queryBOOLEANS(2).BOOLEANS(d8)

B19：计算列名，例如：计算需要的字段名称，例如：d2,d8。

SPL解析：开始查询

友乾营

	A	B
20	/合并查询条件和列名、开始查询	
21	=if(len(B9)==0,B19,B9/","/B19)	=if(len(A19)==0,B10,B10/"&&"/A19)
22	=file("data/ctx/databitC.ctx").create()	
23	=A22.cursor(${A21};${B21})	
24	return A23.skip()	

A21：合并列名，例如：c1,d2,d8

B21：合并过滤条件，例如：

and(c1,16908288)==16908288&&queryBOOLEANS(1).BOOLEANS(d2)&&queryBOOLEANS(2).BOOLEANS(d8)

A22：建立主文件对象对应组表。A23：用准备好的列名、过滤条件新建游标。

A24：计算符合条件的记录数，并返回结果。

效果实测

友乾营

专注数据技术的社群

性能测试

模拟真实环境，大数据量实测

友乾营

1亿客户

每客群包含上千万客户

…

7000客群

任意两个求交集

…

20维度

对应几百个维度属性

…

12个月

并发计算12个月数据

…

同样的服务器，对比测试

VS

ORACLE
DATABASE

集算器比Oracle快8倍!

Oracle对比测试，选择最通用的问题原型一，两表关联方式。用户表、客群表均以ID字段建好了索引。

润乾线上直销系统

好多乾 – 润乾互联网营销

http://www.raqsoft.com.cn/wx/hdq-strategy.html