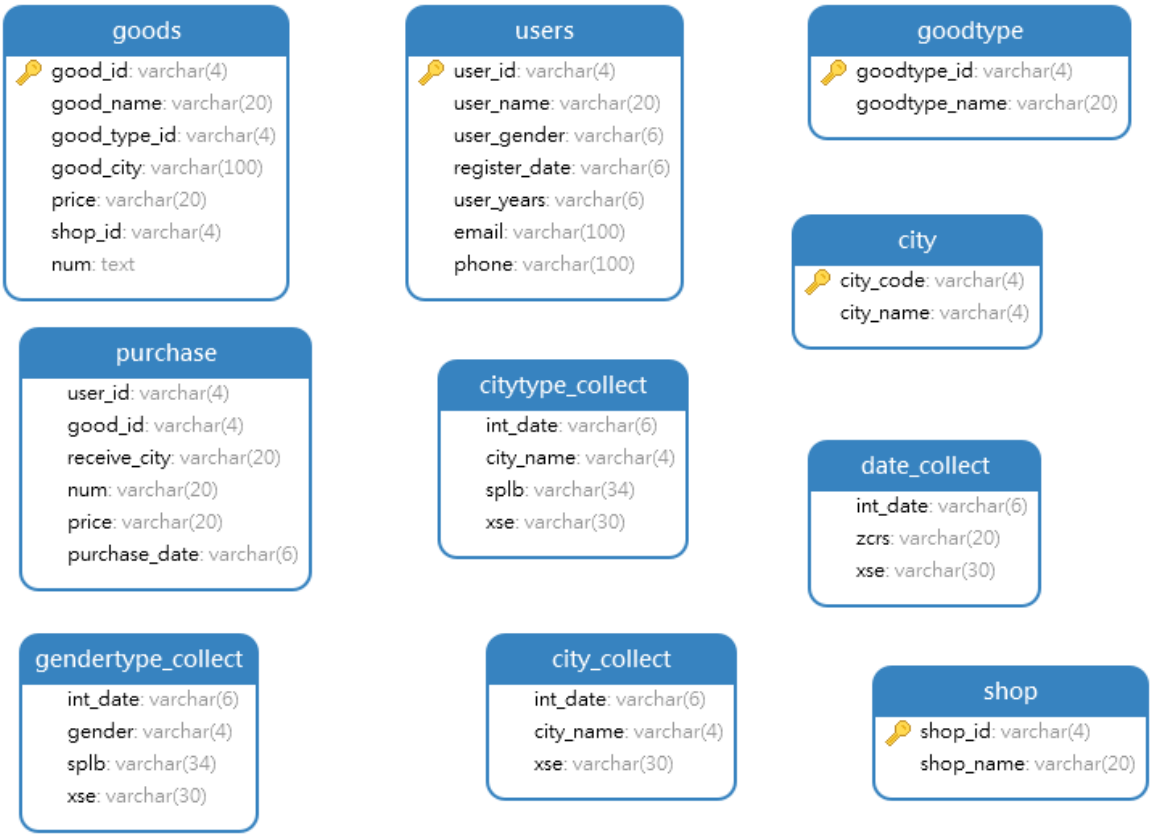
实训第五周-数据治理

初始化数据库

ODS建表及DW+DM建表

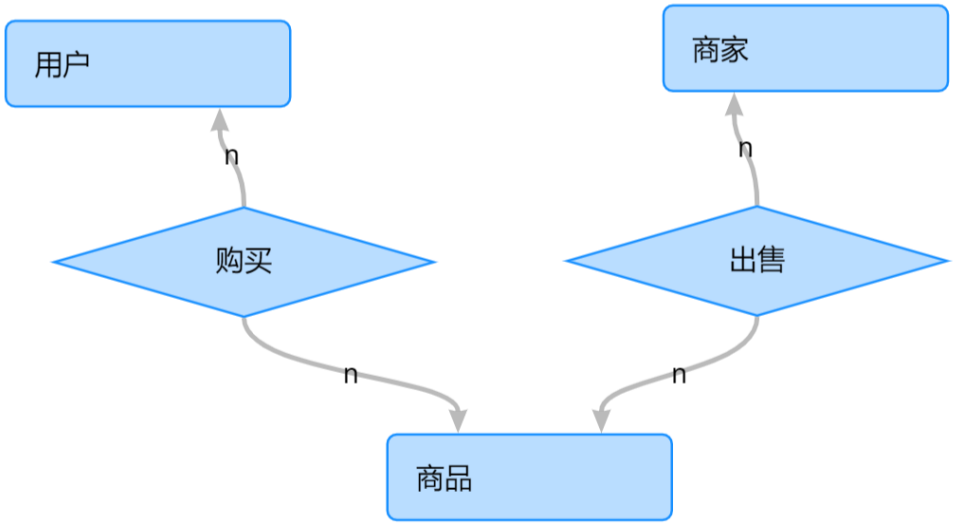


插入数据

```
insert into "week5"."city"(city_code,city_name) values('001','广州');
insert into "week5"."city"(city_code,city_name) values('002','北京');
insert into "week5"."city"(city_code,city_name) values('003','深圳');
insert into "week5"."goodtype"(goodtype_id,goodtype_name) values('001','食品');
insert into "week5"."goodtype"(goodtype_id,goodtype_name) values('002','文具');
insert into "week5"."goodtype"(goodtype_id,goodtype_name) values('003','运动器材');
insert into "week5"."shop"(shop_id,shop_name) values('001','6.6杂货店');
insert into "week5"."shop"(shop_id,shop_name) values('002','辉瑞美食铺');
insert into "week5"."shop"(shop_id,shop_name) values('003','瑞兴超市');
insert into "week5"."goods"(good_id,good_name,good_type_id,good_city,price,shop_id,num) values('001','喜之郎果冻','001','001','13.52','002',600);
insert into "week5"."goods"(good_id,good_name,good_type_id,good_city,price,shop_id,num) values('002','瑞幸咖啡','001','002','43.59','002',500);
insert into "week5"."goods"(good_id,good_name,good_type_id,good_city,price,shop_id,num) values('003','yonex羽毛球拍','003','002','532.99','003',300);
insert into "week5"."goods"(good_id,good_name,good_type_id,good_city,price,shop_id,num) values('004','双鱼乒乓球拍','003','003','57.66','003',400);
insert into "week5"."goods"(good_id,good_name,good_type_id,good_city,price,shop_id,num) values('005','晨光中性笔','002','003','2.50','001',200);
insert into "week5"."goods"(good_id,good_name,good_type_id,good_city,price,shop_id,num) values('006','百马钢笔','002','001','106.78','001',1000);
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('001','爱德华','male','202005','2000','123456@126.com','13522254456');
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('002','露易丝','female','202006','1995','12pp56@126.com','13977254456');
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('003','内马尔','male','202103','2003','1uu456@126.com','1344544536');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('001','001','002','30','400.36','202102');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('002','003','001','50','200.35','202102');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('001','002','003','60','280.23','202102');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('001','005','001','20','132.21','202103');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('003','006','002','30','3201.59','202103');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('002','002','002','10','442.66','202104');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('003','001','003','50','673.25','202104');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('002','003','001','40','21023.59','202103');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('001','004','003','20','1123.55','202103');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('003','005','002','35','108.76','202104');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('002','006','001','15','1653.89','202105');
insert into "week5"."purchase"(user_id,good_id,receive_city,num,price,purchase_date) values('001','002','002','23','896.38','202106');
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('004','巴克洛夫','male','202103','2003','1uu456@126.com','1344544536');
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('005','特洛伊','male','202103','2009','1uu4126.com','134454456');
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('006','何罗喜','male','202103','2004','1uu456@126.com','134454k4536');
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('007','德克力','male','202103','2003','1uu45126.com','134444536');
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('008','张华西','female','202103','1996','1u56@126.com','134544536');
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('009','诸葛曦','female','202103','2008','1uu456@126.55.com','1344574536');
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('010','梁何利','female','202103','2013','1uud456@126.com','1344544836');
insert into "week5"."users"(user_id,user_name,user_gender,register_date,user_years,email,phone) values('011','罗曼园','female','202103','2001','1us456@126.com','1344544596');
```

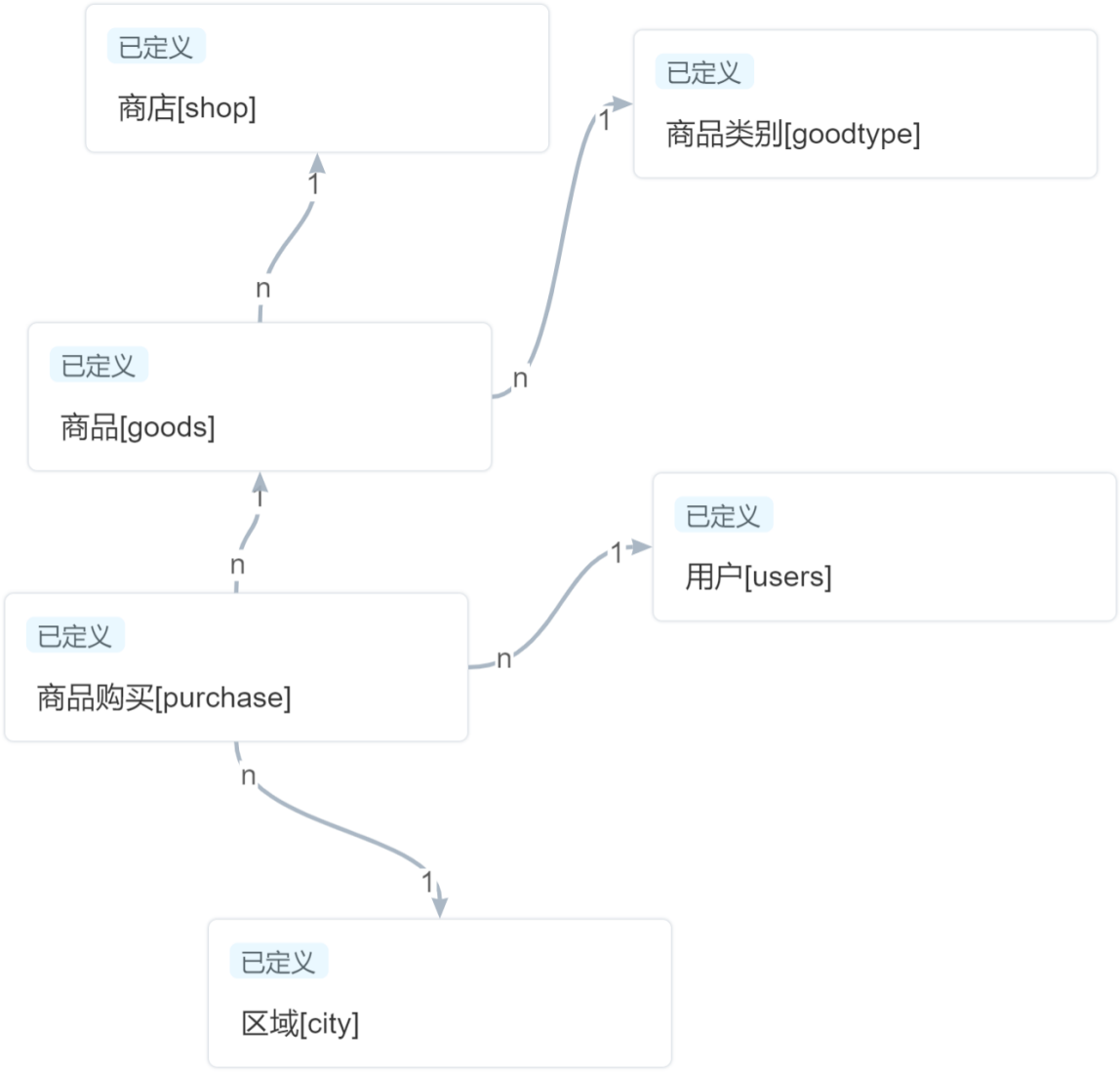
模型

业务系统物理模型

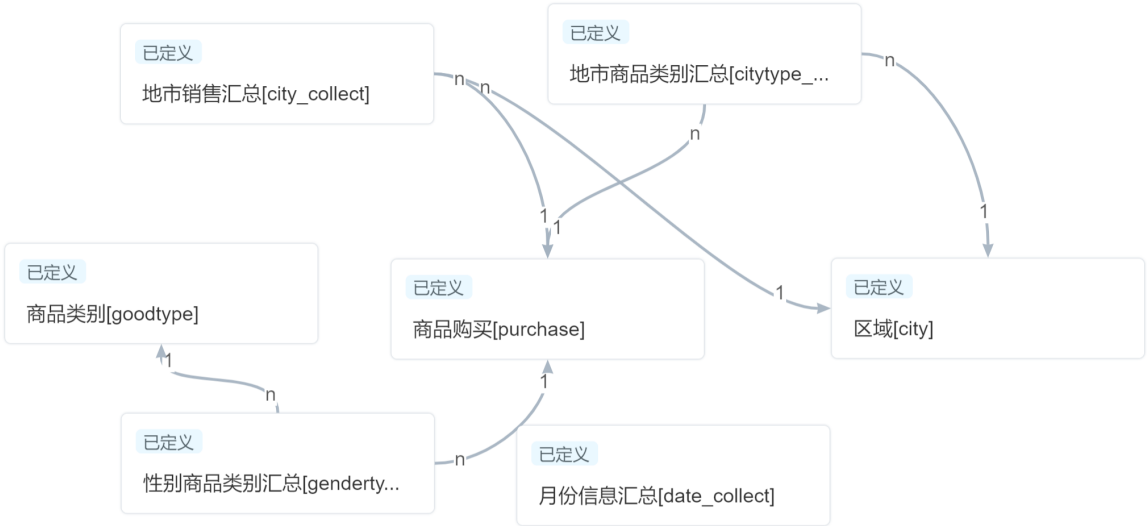


数据仓库物理模型

ODS


已定义
商店[shop]
已定义
商品类别[goodtype]
已定义
商品[goods]
已定义
用户[users]
已定义
商品购买[purchase]
已定义
区域[city]
1
n

DW+DM


已定义
地市销售汇总[city_collect]
已定义
地市商品类别汇总[citytype_...
已定义
商品类别[goodtype]
已定义
商品购买[purchase]
已定义
区域[city]
已定义
性别商品类别汇总[genderty...
已定义
月份信息汇总[date_collect]
1
n

sql需求

按月份查看销售额及新注册人数

```
select a.month as month, COALESCE(price,0) as turnover,COALESCE(nums,0) from
((select sum(cast(price as decimal))as price,purchase_date as month from
week5.purchase group by month)a
left join
(select count(*) as nums,register_date as month from week5.users group by
month)b on a.month=b.month)
union
select b.month as month,COALESCE(price,0) as turnover,COALESCE(nums,0) from
((select sum(cast(price as decimal))as price,purchase_date as month from
week5.purchase group by month)a
right join
(select count(*) as nums,register_date as month from week5.users group by
month)b on a.month=b.month)
```

month	turnover	nums
202102	880.94	0
202105	1653.89	0
202103	25480.94	1
202104	1224.67	0
202106	896.38	0
202005	0	1
202006	0	1

按月份查看收获城市销售额度

将月份设为一个维度聚合

```
select sum(cast(price as decimal))as turnover,city_name as city,purchase_date as
month from(
week5.purchase a left join week5.city b on a.receive_city=b.city_code
)group by month,b.city_name
```

turnover	city	month
1123.55	深圳	202103
673.25	深圳	202104
1653.89	广州	202105
21155.80	广州	202103
400.36	北京	202102
200.35	广州	202102
280.23	深圳	202102
551.42	北京	202104
896.38	北京	202106
3201.59	北京	202103

将指定月份设为查询条件

```
select sum(cast(price as decimal))as turnover,city_name as city from(
week5.purchase a left join week5.city b on a.receive_city=b.city_code
)where purchase_date='202103' group by b.city_name
```

turnover	city
1123.55	深圳
21155.80	广州
3201.59	北京

按月查看城市，商品类别销售额

将月份设为一个维度聚合

```
select sum(cast(a.price as decimal))as turnover,city_name as city,good_name as
goods,purchase_date as month from(
week5.purchase a left join week5.city b on a.receive_city=b.city_code left join
week5.goods c on a.good_id=c.good_id
)group by month,b.city_name,goods
```

turnover	city	goods	month
200.35	广州	yonex羽毛球拍	202102
280.23	深圳	瑞幸咖啡	202102
442.66	北京	瑞幸咖啡	202104
896.38	北京	瑞幸咖啡	202106
21023.59	广州	yonex羽毛球拍	202103
1123.55	深圳	双鱼乒乓球拍	202103
108.76	北京	晨光中性笔	202104
673.25	深圳	喜之郎果冻	202104
400.36	北京	喜之郎果冻	202102
3201.59	北京	百马钢笔	202103
132.21	广州	晨光中性笔	202103
1653.89	广州	百马钢笔	202105

将指定月份设为查询条件

```
select sum(cast(a.price as decimal))as turnover,city_name as city,good_name as
goods from(
week5.purchase a left join week5.city b on a.receive_city=b.city_code left join
week5.goods c on a.good_id=c.good_id
)where purchase_date='202103' group by b.city_name,goods
```

turnover	city	goods
132.21	广州	晨光中性笔
1123.55	深圳	双鱼乒乓球拍
3201.59	北京	百马钢笔
21023.59	广州	yonex羽毛球拍

按月查看城市，商品类别销售额

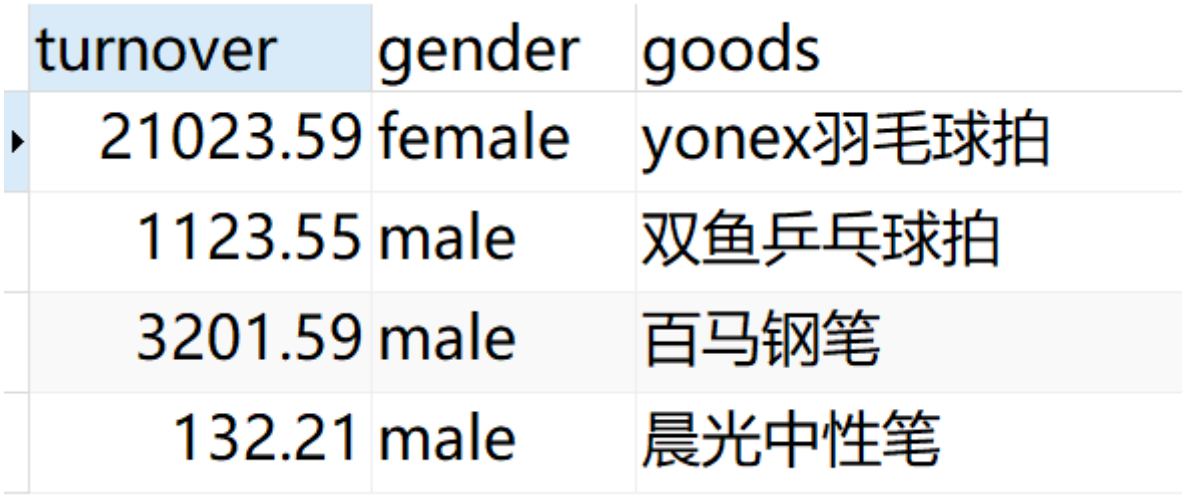
将月份设为一个维度聚合

```
select sum(cast(a.price as decimal))as turnover,user_gender as gender,good_name
as goods,purchase_date as month from(
week5.purchase a left join week5.users b on a.user_id=b.user_id left join
week5.goods c on a.good_id=c.good_id
)group by month,b.user_gender,goods
```

turnover	gender	goods	month
21023.59	female	yonex羽毛球拍	202103
1123.55	male	双鱼乒乓球拍	202103
108.76	male	晨光中性笔	202104
1653.89	female	百马钢笔	202105
280.23	male	瑞幸咖啡	202102
132.21	male	晨光中性笔	202103
896.38	male	瑞幸咖啡	202106
200.35	female	yonex羽毛球拍	202102
400.36	male	喜之郎果冻	202102
3201.59	male	百马钢笔	202103
442.66	female	瑞幸咖啡	202104
673.25	male	喜之郎果冻	202104

将指定月份设为查询条件

```
select sum(cast(a.price as decimal))as turnover,user_gender as gender,good_name
as goods from(
week5.purchase a left join week5.users b on a.user_id=b.user_id left join
week5.goods c on a.good_id=c.good_id
)where purchase_date='202103' group by b.user_gender,goods
```



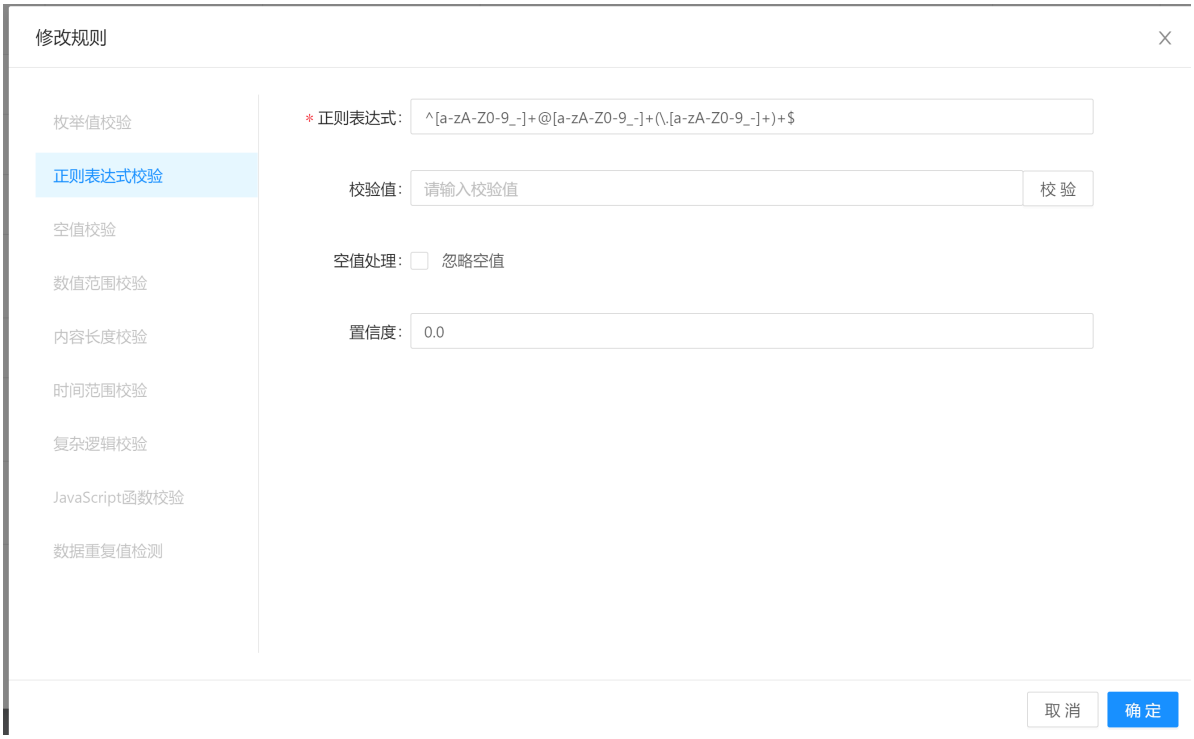
turnover	gender	goods
21023.59	female	yonex羽毛球拍
1123.55	male	双鱼乒乓球拍
3201.59	male	百马钢笔
132.21	male	晨光中性笔

数据质量任务核查

目前用户数据

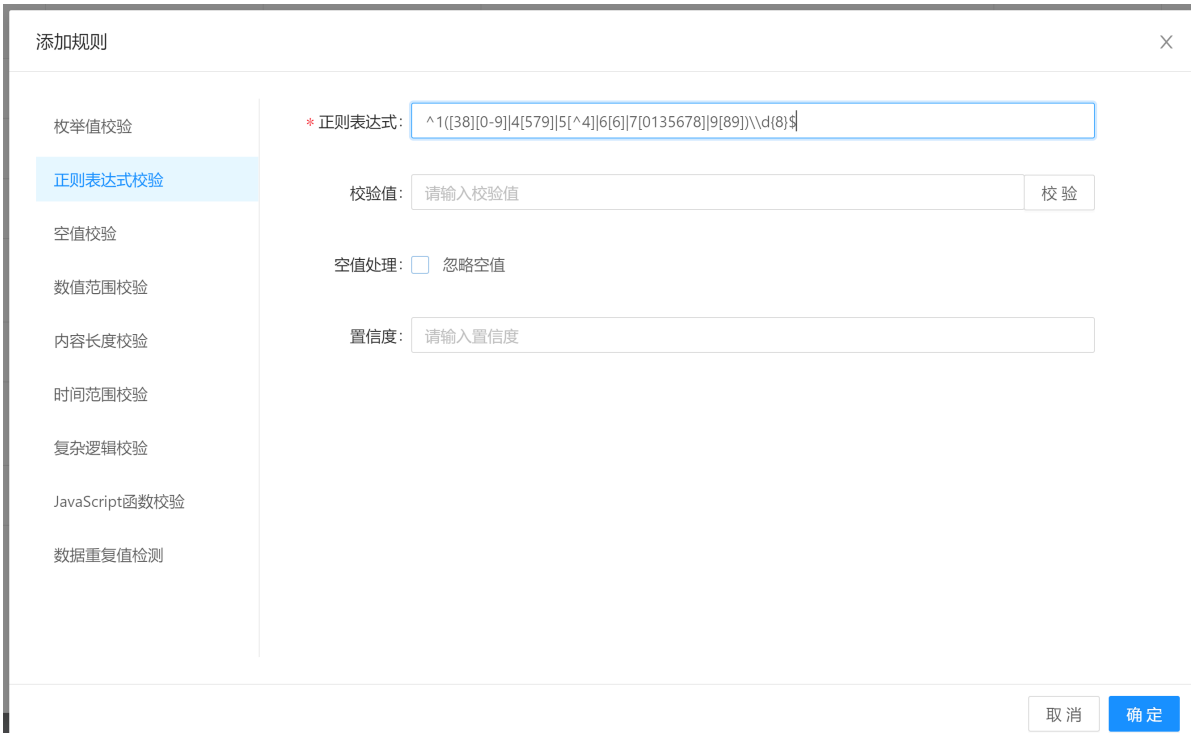
user_id	user_name	user_gender	register_dat	user_years	email	phone
003	内马尔	male	202103	2003	1uu456@126.com	1344544536
007	德克力	male	202103	2003	1uu45126.com	134444536
010	梁何利	female	202103	2013	1uud456@126.com	1344544836
002	露易丝	female	202006	1995	12pp56@126.com	13977254456
004	巴克洛夫	male	202103	2003	1uu456@126.com	1344544536
005	特洛伊	male	202103	2009	1uu4126.com	134454456
008	张华西	female	202103	1996	1u56@126.com	134544536
011	罗曼园	female	202103	2001	1us456@126.com	1344544596
001	爱德华	male	202005	2000	123456@126.com	13522254456
006	何罗喜	male	202103	2004	1uu456@126.com	134454k4536
009	诸葛曦	female	202103	2008	1uu456@126.55.com	1344574536

设置邮箱格式检验规则



设置年龄检验规则

设置手机格式检验规则


添加规则
枚举值校验
正则表达式校验
空值校验
数值范围校验
内容长度校验
时间范围校验
复杂逻辑校验
JavaScript函数校验
数据重复值检测
* 正则表达式:
^1([38][0-9]|4[579]|5[^4]|6[6]|7[0135678]|9[89])\\d{8}$
校验值:
请输入校验值
校 验
空值处理:
忽略空值
置信度:
请输入置信度
取 消
确 定

检验结果


greenplum
读取
成功
用户
11条 (0.55KB)
11 条/秒
数据检测
成功
数据检测
11条 (0.59KB)
脏数据：11条


节点：数据检测
开始时间：2021-07-02 16:38:53
结束时间：2021-07-02 16:38:53
日志
输出参数
脏数据
名称
数量
操作
phone [用户手机号码]_正则表达式校验自定义
11
下载
email [用户邮箱]_正则表达式校验自定义
9
下载
user_years [用户年龄]_数值范围校验自定义
4
下载
全表脏数据
11
下载