

17 | 数据获取与分析：常见的SQL 技巧和分析方法

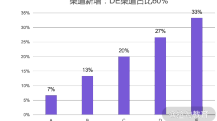
1. 数据获取前期准备

- 1.1. 了解业务方和研发说的是哪张表？ 那份日志？
- 1.2. 了解表和日志的筛选条件；为什么这样筛选？
- 1.3. 了解表和日志是否曾经有缺失？
- 1.4. 验证现在是否同样有坑？ 跑一遍数据

2. SQL提数常见问题

- 2.1. 非技术问题
    - 老是被各种事打断
    - 遇到一些坑
    - 突然发现一个新点就深挖
    - 用好“时间管理”：分清事情主次；用好早上时间；提前了解会议主题；周末多下功夫
    - 可观人问事；最好文档化！
    - 跳出思维误区，做好问题拆解！
  - 2.2. 技术问题
    - 不会提数和分析
    - 那就多学，多向同事请教；
    - 这是一些常见的SQL提数会遇到的问题
- 先聚合再计算**  
如果要计算某个维度下的用户数，不要直接count (distinct imei)  
而是Select city,count (1) as uv from (select city,imei,count (1) from a group by city,imei) t1 group by city
- 一列度多行**  
Ab测试中会对一个用户打很多标签，而这些标签都是存在一个字段中  
所以要查看标签维度指标，就要对该字段进行列变行拆解  
Select",b from t1 Lateral view explode (a) table as b
- 取TOP**  
要看某分类下的Top10消费数字分类（金额一致就并列）  
Select",rank () over (partition by a order by b desc) as rank from table t1

渠道	新增用户数
A	1万
B	2万
C	3万
D	4万
E	5万
合计	15万



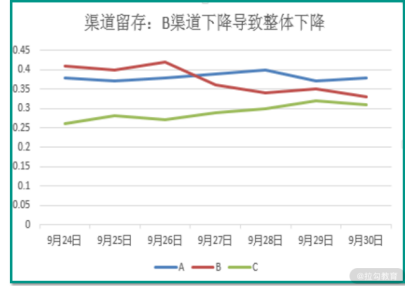
(比如这样的“结构分析”：查看渠道分布占比)

3.1. 结构分析

微视与抖音的重合用户消费分类	消费CTR	微视大点消费分类	消费CTR	diff相对值
搞笑	0.08	搞笑	0.06	33%
游戏	0.06	游戏	0.07	-14%
舞蹈	0.1	舞蹈	0.05	100%
正能量	0.03	正能量	0.05	-40%
萌系	0.02	萌系	0.03	-33%
明星	0.08	明星	0.06	33%
美食	0.03	美食	0.01	200%
音乐	0.04	音乐	0.05	-20%
二次元	0.05	二次元	0.08	-36%
运动	0.03	运动	0.06	-50%

- 数据对比才有意义，常见的是对比大量，比如右侧，其中 diff = (第二列的消费 CTR / 第四列的消费 CTR) - 1

3.2. 对比分析



(像这样看指标趋势，波动则进行维度拆解！)

3.3. 时序分析

- 3.4. 相关性分析
- 3.5. 机器学习

3. 常用的分析方法

4. 总结“5W1H”分析模板

- 5.1. Who: 指用户基础属性、用户画像
- 5.2. Where: 渠道分析，渠道入口，用户从哪里来
- 5.3. When: 时间上的特征
- 5.4. What: 用户使用了什么功能，哪些行为更加重要
- 5.5. Why: 为什么要这么做，用户是主动还是被动做的
- 5.6. How: 怎么做的，行为路径是什么

• 提数分析完成后，别急着写报告！！  
• 先同步给业务方，看看有无问题；看看他们的落地方案再写报告