12/05/2024

1. 外连接中 where 和 on 的区别?

Where 用在连接完成后作为筛选条件，on 用在连接过程中作为筛选条件。因此 On 用于生成新表的结构，例如合并两个连接表的某一列，而 where 语法顺序在 on 之后用于筛选新表。

2. Map join 是什么?

Map join 称作**内存连接**用于在大数据场景如 hive 下优化连接操作，一个较小的表基于内存而不是磁盘存储和调用并可以连接到所有的数据节点上，这样大数据表在节点上处理数据时只需要访问小表而不是磁盘。

3. 数据域是什么?

数据域是一组具有共同属性或特性的数据集合可以看作是数据的逻辑分组，数据域可以保证数据的 ACID(atomic, consistency, isolated, determined/持久性)

4. 数据热点/数据倾斜在哪些场景会发生?
    a. 在分布式系统如 Borg 上, 部分工作负载 Borglet 加载了过多的资源而其他 borglet 空闲，这影响系统的效率。
    b. 在数据仓库上，如果某些产品的某些键如 ID 明显增多，查询这些键的操作会产生数据倾斜。

    Solutions:

    改变分区策略: 使用 hash 分区或范围分区保证数据·更加均匀的分配到各个节点上。
    使用动态重新分区: spark 在运行时可以动态调整数据的分区以响应数据倾斜发生。
    优化查询逻辑: 对索引热键进行优化或减少对热键的使用。

5. Count(*) 和 count(1)的区别?

Count(*)不关心列的数据哪怕出现了 null，它计算所有的行，count(1)实际上是计算常数 1 出现的次数，由于每行都将为 1 次数，因此它是在每行上计算 1，这同样不会因为列值是否为 null 等而改变，所以两者返回相同的结果。

6. 数据仓库是什么，和数据库的区别是什么?

数据仓库（Data Warehouse）是为了便于进行决策支持和数据分析而设计的系统。它是一个集中式数据存储环境，用来存储来自不同来源的整合数据，这些数据被清洗和处理，以便进行查询和分析。(主题导向：围绕产品或客户构建， 集成性：不同源数据的集合，数据经过处理具有一致性， 长久存储：数仓用于查询和报告而不是数据处理， 时间趋势：提供时间变化下的数据便于时间序列分析)

数据库是操作导向：数据的日常处理 CRUD(创建，读取，更新，删除)，实时与易变性。

7. 如何处理数据缺失值?

删除法: pandas.dropna(), 补充法： pandas.fillna()

8. 如何优化 SQL 查询

优化查询可以提高查询速度和降低数据库的负担,

    a. 创建索引，where on, order by 常用的列可以创建索引
    b. 使用多列索引,
    c. 使用 where 而不是 having，having 在聚合操作完成后才对结果进行过滤而 where 在此之间进行过滤，将子查询改写为 join 连接操作，当必须使用子查询时且子查询返回大量的结果时使用 exists 而不是 in。
    d. 使用正确的 join 以及多表 join 时先处理数据量较小的表。

9. 数据分区的方式与好处?

a. 范围分区：根据某个字段的范围如日期分区
b. 列表分区：根据字段的唯一值分区：如国家，产品型号 ID
c. 散列分区：使用 hash 函数将数据均匀分配到各区避免数据倾斜，不用考虑逻辑
d. 复合分区：综合上述方法

好处：

提高查询性能：仅在分区上执行查询，索引更少，数据量大幅减小

维护更加容易：对大数据表进行分区操作便于更新和备份等操作，各个分区可以设置只读操作。(ACID)

优化存储：根据分区特点决定存储位置是磁盘还是内存。

10. 数据清洗的主要步骤
    a. 数据探索（描述性分析）：探索下数值型数据的分布，是否存在偏态，均值和异常值(箱线图)，确定是否存在缺失，缺失的各列比例如何，人工识别其他错误例如重复，不一致记录。
    b. 数据清理，处理缺失值，纠正类型错误如日期，数值型数据混合了文本，数据缩放(正则化和标准化)

    Scaling has normalisation and standardisation two ways.

    Normalisation will rescale variables to 0-1 but it is poor for outliers.

    Standardisation will ensure variables have mean equal to 0 and fixed variance but bad for long tail variable and it don't give a specific bound for variables that cause problems for some model.

    c. 数据转换：特征工程生成新的变量或是对原来变量进行转换譬如创建比率。如果有必要还需要对数据格式进行转换。
    d. 备份数据
11. 数据建模是什么，常见的数据模型有哪些？

数据建模即创建数据模型，数据模型是对现实世界数据关系的抽象表示

数据模型：

a. 关系数据模型：由表组成，例如 mysql
b. 层次模型，将数据组织为树状的结构，每个节点表示一个数据记录(IBM 的 IMS)
c. 网络模型，是对层次模型的扩展，允许一个自节点拥有多个父节点，图结构表示数据之间的关系
d. ER 模型，实体关系模型，通过实体、实体属性和实体之间的关系来表示数据，实体表示事物，关系表示事物之间的联系。用于数据库设计。
e. 对象关系模型，ER 模型的扩展，添加了继承、多态、与封装等对象特性。

12. 信息检索中如何比较不同排序算法的效果？

排序模型的评估：

1. 相关性判定：三个策略，a) 相关度，b) 按对的偏好，c) 总顺序
2. 评估度量方法：a) Mean Reciprocal Rank; b) Mean Average Precision; c)Discounted Comulative Gain; d)Rank Correlation

a. **平均精确度均值(MAP):** 属于评估测量方法 b，当存在多个查询时，对所有查询的精度取平均值。
b. **Mean Reciprocal Rank (MRR):** 属于评估测量方法 a，当存在一个查询时，用于评估返回的结果列表中第一个相关文档的平均排名的倒数，排名越后，MRR 越小，排名为 1MRR 为 1。

传统排序模型有**相关性排序模型（如 VSM）**和**重要性排序模型（如 pageRank）**：

向量空间模型(VSM): 在欧式空间中将文档与查询词用向量表示，两向量的内积可作为二者相关性。该方法假设 term 之间彼此独立。此向量的计算可借助于 TF-IDF，其中 TF 为 term 在文档中的频率，IDF 表示 term 被文档包含的程度(N 为总文档数，n(t)为包含 termt 的文档数目，$IDF(t) = log\frac{N}{n(t)}$)，即 TF-IDF=TF*IDF。

PageRank: 基于用户随机点击链接抵达某网页的概率进行排序。网页 du 的 PR 值依赖于链接到 du 的网页 dv 的 PR 值，除以 dv 的出链接数。

18/05/2024

13. 常用的数据结构有哪些？

基本的数据结构：

1. 数组(array):存储相同数据类型的元素，线性数据结构，通过索引访问。
2. 链表(linked list)：线性数据结构，元素不必在内存中连续存储，每个节点包含数据和指向下个节点的指针。
3. 栈(stack)：后进先出，添加和删除在一端进行。
4. 队列(queue)：先进先出，添加和删除在两端进行。
5. 哈希表(hash table)
6. 树(tree)：非线性数据结构，用来表示层级数据。
7. 图(graph)：一组由边连接的节点，可以是无相的或单向的。

高级数据结构：了解一些高级数据结构，如平衡树（如 AVL 树、红黑树）、堆、字典树（Trie）、B 树等，特别是如果你的工作涉及到复杂的数据处理或性能优化。

在关系型数据库 mysql 中数据结构有：表、视图(虚拟表，基于多个表来表现查询结果)、索引(辅助数据结构，索引使用一种或多种列作为键)、主键、外键、触发器(triggers)和存储过程(一种预编译的代码块，用于封装逻辑在数据库中执行)。

14.讲述一个做过的数据科学项目，并从数据科学的视角优化该项目？

遵循数据科学

15. 用户生命周期如何刻画？

用户生命周期包括五个阶段，引入期(完成注册成为用户)，成长期(使用产品的各种服务并开始付费)，成熟期(活跃使用，多次付费)，休眠期(一段时间不登陆的成熟用户)，流失期(超过比较长的时间不再使用产品)。

从数据分析的角度可以从用户基础数据和行为数据双角度进行描述性统计来找到新用户到成长用户到成熟用户的一些规律，譬如基础数据下年龄性别区域，用户行为数据中使用时间段，聚焦的服务，使用的时长，在这阶段可以展开 ab test 或机器学习的方式来探索。

同时应该防止成熟期用户衰退为休眠期，可以定义休眠用户的特征(也可以识别阶段之间转换的比率)，设置一些预警机制指标，完成干预(譬如优化各阶段的用户体验提高粘性，优惠卷福利)。

16. 数据分析中擅长的能力？

数据分析中常见的能力项有数据处理与清洗，数据建模与统计分析，数据可视化，业务分析，报告呈现结果，因为在统计与机器学习中学到的知识，在数据清洗中我会根据数据特性和下游使用的模型特点来处理数据，譬如当数值型数据展现长尾分布时，应该使用正则化，当数据存在较多异常值时使用标准化。在模型方面，当使用随机森林时，它对异常值 outlier 不敏感，正则化或标准化是不需要的，而使用监督算法 svm 或无监督算法 k-means 时，缩放就非常重

要，最好使用标准化处理为均值为 0，方差为固定值。此外常规的使用 python 中 fillna 处理缺失值，数据类型转换，数据编码，特征工程创建新的变量等。在数据建模与统计分析方面，我可以使用常用的机器学习模型遵循数据科学的流程来处理，训练，验证，比较模型并给出结果。我还可以使用 r 提供对更复杂的数据譬如时间序列相关进行拟合，我可以拟合 glm, gam, gee 等高级模型，使用常见的指数家族 link 转换，平滑窗口，惩罚项 L1/L2 来更好的分析数据，并依据四种模型假设给出相应的参数假设结果与可视化结果，在使得报告具有更加的可信度和可解释性与机器学习相比。在可视化方面，我拥有 tableau 桌面专家认证证书，同时我学习的可视化课程训练我可以设计合理美观的视图，我熟悉可视化的大拇指准则可以保证报告中图表的规范。

17. t 检验的原理

t 检验以 student 分布为基础，是对一个或两个样本均值进行假设的常用方法。

首先根据检验目的构建原假设 H0 和备择假设 H1,然后构建 t 统计量，这里如果是两个样本要注意方差是否相同，最后根据抽取的样本计算 t 值与临界值比较做出结论。

对于单样本 t 检验，检验一个单个样本的均值是否等于某个数值一般是总体均值。

对于独立样本 t 检验，检验来自两个独立样本之间的均值是否存在显著差异，这种情况我们假设两个样本方差相同，如果 F-test 检验出方差确实不相同，应该转换为使用 welch t 检验。

对于配对样本 t 检验，检验来自一个整体的两个样本之间的均值是否存在显著差异。如果总体方差未知，使用样本方差来估算总体标准差，使用样本方差有效的原因是这里基于自由度的调整，因此不知道总体方差也能做出有效推断。

18. 假设检验的原理，p 值的含义，Z 检验与 T 检验的区别?

假设检验是一种从整体中抽取样本，然后根据样本信息推测总体特征的统计推断方法，本质原理是小概率事件原理(小概率事件在一次实验中基本不可能发生)，如果小概率事件发生了则原假设为伪，否则则不能拒绝原假设。p 值指的是原假设为真时比得到的观测结果更极端的结果出现的概率，一般与显著性水平α比较，若 p 小于α则说明原假设为假，反之则原假设为真。Z 检验适合样本服从或近似 Z 分布的假设检验，T 检验适合样本分布服从 t 分布的假设检验。一般来说总体均值已知方差未知，且样本量小于 30 使用 t 检验，如果样本量大于 30 则使用 Z 检验。当样本数大时根据大数定律样本均值逼近总体均值，因此没必要做 z 检验。当样本数小时不能保证样本服从正态分布所以使用 t 检验合理。

19. 中心极限定理的原理?

对于独立且同样分布的随机变量，即使原始变量本身不是正态分布，标准化均值的抽样分布也趋近于标准正态分布，譬如对于参数为 n, p 的二项分布以均值为 np, 方差为 np(1-p)的正态分布为极限。扩展是林德伯格-莱维定律，独立同分布的变量和仍然满足正态分布，独立不同分布的变量和在满足每个变量三次方的期望有限且所有变量三次方期望之和与变量方差和的极限为 0 时（李亚普诺夫条件）仍然满足正态分布。即 $S_n/\sigma_n \to N(0,1)$, $S_n = \sum_{i=1}^{n} X_i, S_i^2 = Var(X_i), \sigma_n^2 = \sum_{i=1}^{n} S_i^2$

20. 第一类错误和第二类错误是什么?

第一类错误α是弃真错误，即原假设为真，我们却选择了备择假设，一般α为 0.05, 0.1。第二类错误β成为取伪错误，即备择假设为真我们却选择了原假设。

第一类错误和第二类错误是此消彼长的关系，如果样本量不变减小α必然使得β增大。如果想同时减小这两种错误，应该增大样本量。一般第一类错误比第二类错误重要，因为人们希望否定原假设得到备择假设。

21. 概率和似然的区别?

概率是参数已知的情况下预测观测事件的结果，似然则是给出了观测值对参数进行估计，似然是条件概率的逆反。

22. 泊松分布和二项分布的区别?

二项分布的概率密度函数以均值为对称形状，泊松分布的概率密度函数呈右偏的单峰形状，标准差决定了曲线的宽度，当 n 趋向于无穷大时二项分布接近于正态分布，当 n 比较大但 np 比较小时二项分布趋近于泊松分布(np=λ)。**泊松分布是在不知道事件的可能发生总次数时对小概率事件建模，它用于给定空间和时间内随机事件发生的次数，我们用已知的小段的空间和事件发生次数来描述总体发生的次数。**

泊松分布密度函数公式：f(x, λ)=λe^(-λx), x>=0; 0, x<0.

23. 卡方检验?

卡方检验用于

24. 特征筛选的方法有哪几种?

a. 过滤式：先对数据集进行特征选择譬如相关系数，互信息，或卡方检验，然后再训练学习器。
b. 包裹式：直接把最终要使用的学习器的性能作为特征子集的评价标准（譬如回归模型中的 MSE,分类模型中的 recall）。
c. 嵌入式：将特征选择和学习器的训练融为一体，在训练中选择特征的重要性（譬如随机森林中的 RandomForestClassifier 可以在每次迭代中计算 gini impurity），在 fit 结束后使用 feature_importances_调用得到特征重要性得分。

25. 逻辑回归是二元回归，用于分类问题，由于 Y 是定性变量，因此会在回归中使用 sigmoid 将 Y 转换为离散数值，逻辑回归中使用 cross-entropy 作为损失是为了极大化似然函数，似然函数取负就变成了交叉熵损失函数。
26. RFM 模型用来衡量用户价值和用户消费能力，R(recency)：最近一次消费的时间，与用户流失和复购直接相关；F(frecency)：用户消费频率；M(monetary)：用户消费金额。
27. 什么是好的数据指标?

北极星指标（OMTM）,数据指标应该与目标密切相关，准确且稳定（保证长期反应目标）
可持续性（口径的统一（计算相同，同名同义）保证了长期可用），适合横向与纵向比较的，指标分为定性（回答为什么，吸收主观因素）与定量（回答“什么”和“多少”，排斥主观因素）两种：常见定性指标（平均访问时长，转换率，留存率，活跃度），常见定量指标（PV，UV，DAU）

28. GMV 突然下降，为什么?

这是一种异常归因问题，1，检查数据的准确性，从上游表到数据口径确定数据的准确，确定上个月数据源是否正确，各个渠道引流是否正常，2，查询是否有外部因素干扰，包括环境因素、时间因素和竞品因素，譬如在活动发布后的一个月 GMV 上升了 20%这是合理的，观察同类 app 是不是都出现了下降，排除了上述两点后，再进行拆解分析，GMV=订单数*订单均价
3，分析内部因素，针对商品(是否有打折活动商品)、用户类别（新老顾客，是否为会员）、渠道（从哪下单的）对上述两个指标进行拆解，首先看整体下降还是某一类出现了下降，确定大类因素后，接着对内外部因素进行分析，外部使用 pest(政治，经济，社会面，技术面)模型，内部从用户的行动路径譬如从登录到交易/漏斗模型(AARRR—由上至下，有宽至窄(漏斗状)，是拉新[DNU 指的是每日新登录用户数]-促活[DAU,WAU,MAU]-留存[三日留存五日留存]-转化[PR 指的是付费率]-传播[K 因子-每个用户发送好友邀请数量*成功成为新用户的转化率])，这用户维度可以从基本数据和行为数据入手可以参考 RFM 模型，除此之外，在对每个维度拆分时，记得关注时间维度，譬如拆解到每天观察相邻的上月和去年的同期。
如果恰好这里的指标是复合指标：流量*转换率，就要进一步细分(a.流量升转化降—找转化率的异常问题点，b.流量降转换升但总额低，c.流量持平转换降低，d.流量降转换降，e.流量降转化升)，如果是转换率下降那么拆解转换率计算的每个步骤，观测出现显著变化的指标，如果

流量降则按维度拆解。

29. 怎么构建用户画像?

用户画像是交互设计之父 cooper 提出的建立在一系列数据之上的目标用户模型，是从用户群体中抽象出来的典型用户，本质是用来描述用户需求的工具。被称为“真实用户的虚拟代表”。在互联网下是根据用户人口学特征，网络浏览内容，社交活动与消费行为抽象出的标签化用户模型。用户画像一般分为用户社会属性画像，用户消费画像(城市，收入，性别，年龄，使用设备)，用户行为画像(成长期/成熟期用户，价值指数，流失指数，忠诚指数)，用户兴趣画像(价格偏好，类目偏好，特征偏好，下单时间偏好)。如何构建：定量用户画像:定性研究，多个细分假说，通过定量收集细分数据(调查问卷，小组访谈，动态爬取)，统计统计的聚类分析如 K-means 来细分用户，建立细分群体的用户画像。在细分数据收集中应该遵循用户标识+时间+行为类型+接触点建模方式，即某细分下的用户在哪个功能或页面停留了多久做了什么事情(浏览，购买，评论，点击，收藏等等)。要注意，用户标签的权重可能随着时间的增加而衰减，因此定义事件衰减因子为 r，行为类型，接触点决定了权重，因此用户标签权重=r*行为类型权重*接触点类型。

30. 决策树算法有哪些，有什么区别?

ID3, C4.5, CART，ID3 和 C4.5 用于分类而 CART 可以分类和回归，在特征选择方面，ID3 使用信息增益，C4.5 使用信息增益比，CART 在分类中使用 gini 系数，在回归中使用 MSE 或标准差。决策树通过剪枝(正则化)来处理过拟合问题，或者在生成决策树时限定使用的特征数与树的深度。信息增益 g(D|A)=H(D)-H(D|A)，其中 H 是经验熵，g 是在特征 A 集合下对数据集 D 的信息增益。信息增益越大越好。Gini 系数越小越好，说明数据集的不确定性越小。

31. 怎么解决多重共线性?

Pearson 相关性系数取舍变量，特征工程融合生成新变量，逐步回归，岭回归(L1)

32. 最小二乘法和极大似然估计的区别

最小二乘法的目标是最小化 MSE，即模型预测值与真实值的差异，主要用于回归问题，极大似然估计目标是最小化交叉熵损失，基于概率模型，主要用于数据服从某种概率分布。在模型误差为正态分布时，最小二乘法可以做到最优，在大样本中极大似然估计具有一致性和渐进正态性因此会更有效，极大似然估计适用面更广，包括回归分析，时间序列分析等。

33. sql 去重可以使用哪些方法?

使用 distinct 关键字，使用 group by 分组每组具有唯一的不重复的键，使用 row_number()可以为每组分配一个唯一序号

异动分析业务题例:

日活一亿，平均使用时长 60 分钟，在投放广告后观看人数 50 万，平均使用时长 90 分钟，是否广告对平均使用时长具有影响作用?

这是一个因果推断问题，投放广告后增加了使用时长并不能直接说明广告是有作用的，有可能是这部分人本就属于高度用户因此更愿意观看广告，应该使用 AB test 来确定具体的影响因素，首先确定控制变量广告投放，响应变量平均使用时长，然后确定实验所需的最小样本量(八倍的总体方差除以第二类错误的平方)和实验周期，由于这里使用的响应变量是比率指标，所以总体方差的计算为$P_A(1-P_A)+P_B(1-P_B)$，随后对样本进行分割（这里可以使用互斥加正交划分样本，如果我们有两个相关联的渠道来投放广告，譬如从渠道 A 和渠道 B 都能进入广告，那么 A 和 B 在样本上应该使用互斥即一个样本只能在 A 或 B 出出现，对于每一个渠道，譬如有几个独立的层譬如看广告赢取优惠卷，看广告缩短视频等等，这些层之间是独立的没有关联那么一个样本可以在这些层中多次出现，最后保证所有的控制变量都平等的进行了分割，因为这里我们只有一个广告是否投放因此不需要考虑这个）以保持随机性和接近正态分布和避

免辛普森悖论，使用小样本首先进行灰度测试保证选择的控制变量是可靠的，t 检验。

投放广告后平均使用时长下降 10 分钟怎么分析?

这属于指标异动问题，我们可以进行三部操作来查找原因，第一步是确定数据是否获取正确，确保数据口径一致，主要可以拉出上个月的数据表进行核对同时要确保各渠道引流正常。第二步是排除外部因素的影响，譬如在一个活动发布后的一个月或者受到软件版本迭代的影响，使用 PEST 模型来确定一下，也同时需要调出同行竞争 app 是否出现了同样数据下降。在上述两部都确定无误时，拆解指标也就是平均使用时长来分析内部因素影响锁定问题，可以从用户，商品，渠道三个方面拆解，对于用户有基本特征和行为特征两部分，根据公司的用户画像，查看新老顾客，各个时期的用户人数变化，总使用时长变化，这里可以借助一下 RFM 用户价值于用户消费模型，譬如看出高消费用户的人数和使用时长变化。也可以根据用户的行为路径从登录到消费其中包含了观看广告环节下每个流程的人数变化。然后可以对渠道譬如不同页面，同行 app 流量互换跳转下分析用户人数的变化以及使用时长的变化。在异动分析时要注意辛普森悖论问题，譬如两个地区下一个有显著差异变化一个没有，但是融合到一起时显著变换的区域反而不变化了不显著的区域反而发生了大变。

针对是否应该投放广告做 ab test 需要哪些指标呢?

广告可能不仅影响上面提到的用户平均使用时长，这是一个定量指标，回答的 yes or no 的问题，我们也很希望知道对定性指标的影响譬如留存率，转换率，用户忠诚度，这些可以用来回答后续的 why 问题，具有主观解释性为后续决策提供支持，当然对于定量指标常见的肯定要做 DAU,DNU,ARPU 以及 ROI.

34. 怎样用 rand7()生成 rand10()?

Rand7()只能生成 7 个数很明显无法映射到 10 个数上，但是运行两次 rand7()可以产生 49 个数，如果这 49 个数是均匀分布的，舍去 9 个对 10 取模可以映射到 10 个整数上。

(6*(rand(7)-1)+rand(7))%10 + 1