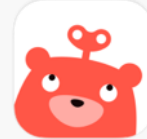


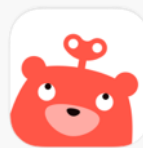


第九届中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2017




C2C市场中推荐系统的 机遇与挑战


张相於



自我介绍

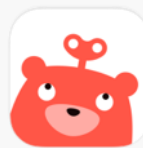
 张相於

 毕业于中国人民大学

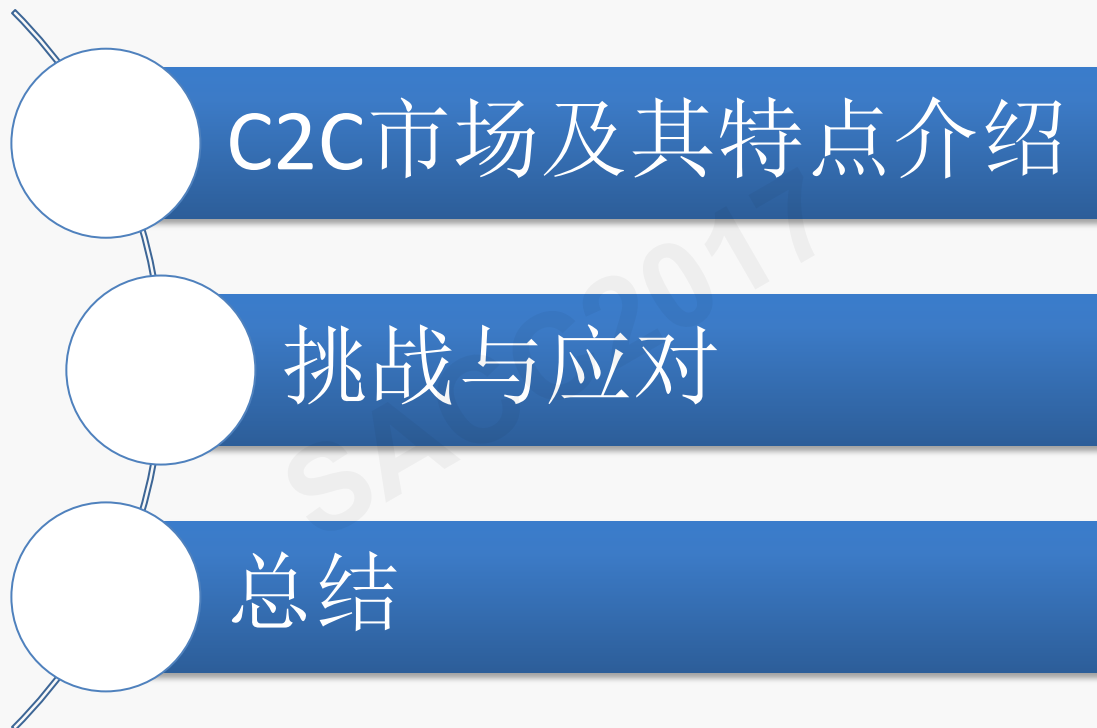
 转转推荐算法部负责人

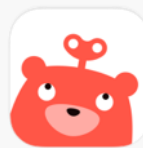
 推荐系统、机器学习系统

 联系方式: zhangxy@live.com

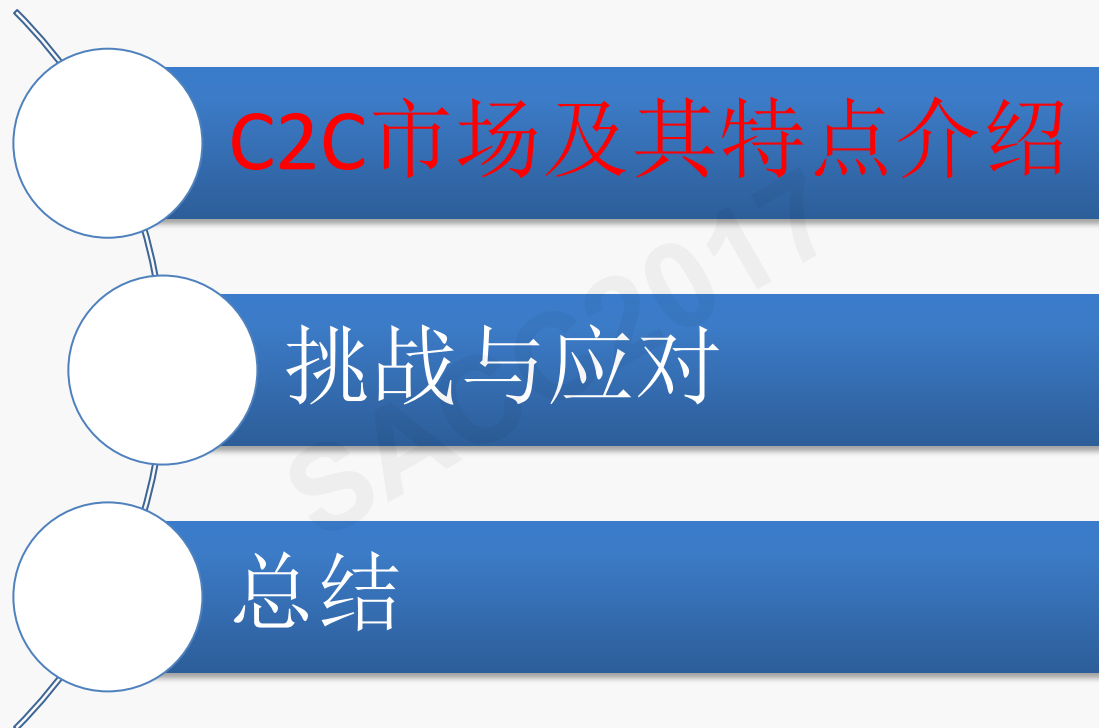


分享提纲






分享提纲






C2C市场

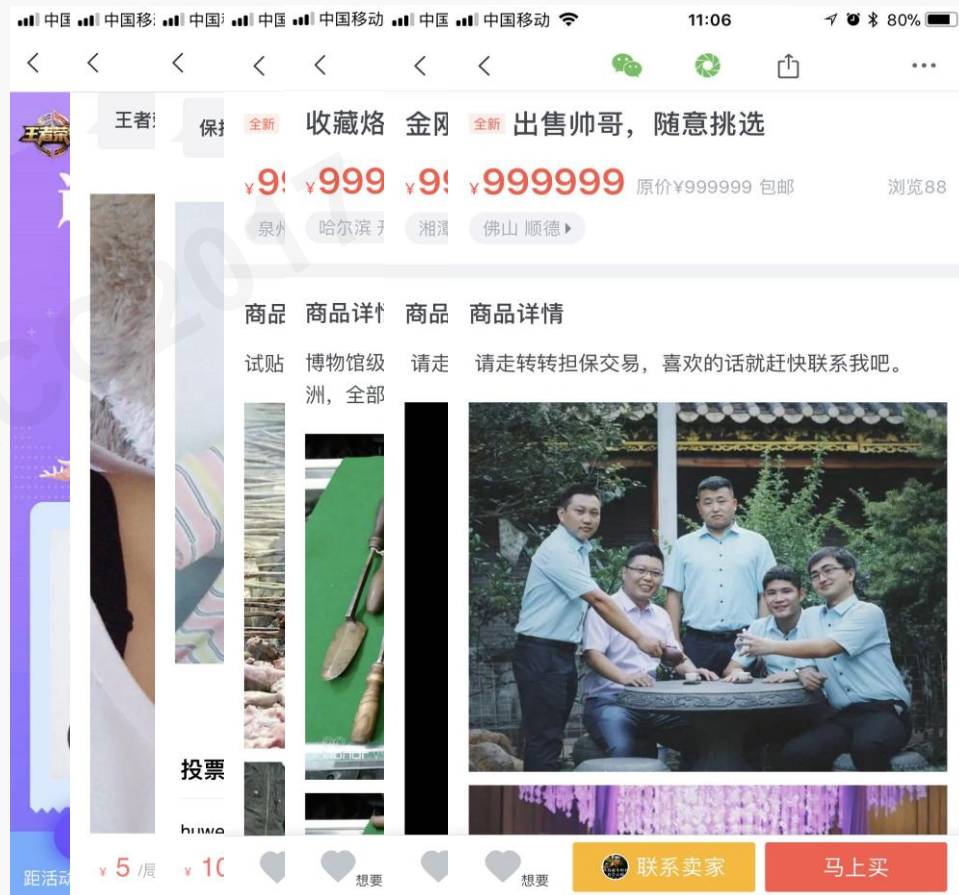
 真·个人对个人的
marketplace

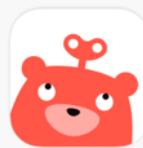
 C2C平台的意义

 物品交易




 技能交换

 发现世界

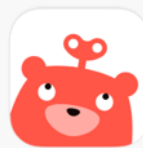




C2C市场的特点

-  信息发布随意性强
-  商品库存唯一性
-  时效敏感性

SACC2017

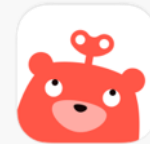


分享提纲





第九届中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2017




挑战1：数据异质性高


SACC2017





数据异质性的含义


 信息发布的随意性


 结构异质性

 结构信息少


 结构信息不确定

 内容异质性

 信息量不确定

 用词多样化

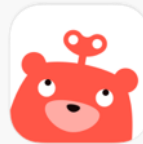
 歧义多

 “iPhone7 128G 国行 无拆无修 发票齐全.....”

 “iPhone6 如图”

 “卖一部iPhone6，要买iPhone7”



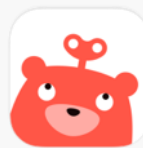


异质数据带来的问题

结构信息少-> 难以制定策略

信息不确定-> 策略覆盖不全

歧义多->策略准确率低



异质数据的优点

A large blue triangle graphic that serves as a background for the three text boxes.

数据量大

多样性丰富

信息及时性



异质数据应对方案

❶ 将非结构化数据转为结构化数据

❷ 按照结构化数据方法来使用

❸ 优点:

❶ 含义明确清晰

❷ 适用范围广

❹ 缺点:

❶ 提取难度高

❷ 信息有损失

❶ 用NLP的方法提取非结构化信息

❷ 用作召回/排序特征

❸ 优点:

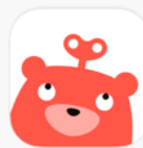
❶ 成熟方法较多

❷ 信息含量大

❹ 缺点:

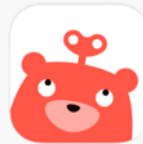
❶ 信息噪音多

❷ 可解释性较弱



数据结构化策略





数据结构化-例子

知识库构建

- 手机：内存、品牌
- 电脑：内存、硬盘

预处理

- 红米 note5A高配版3G

结构映射

- 红米->手机.品牌
- 3G->手机/电脑.内存

结构合并

- 合并：手机.小米.3G



非结构化数据处理

词袋模型

- 适用面广、召回率高、噪音多

文本主题模型（LDA、pLSA）

- 抽象度高、用法多样、实时性能

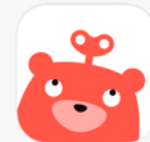
嵌入表示模型（xxx2vec）

- 局部敏感、连续空间、时序敏感

《自然语言处理技术在推荐系统中的应用》
<http://geek.csdn.net/news/detail/208281>

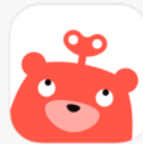


第九届中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2017



挑战2：时效敏感性

SACC2017



时效敏感性的含义

通用时效性

- 对用户的行为作出实时反馈

卖家维度





- 希望自己发布的商品尽快得到注意

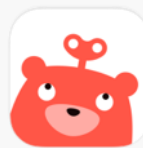
买家维度

- 倾向于与新发布的商品进行交互



时效敏感性的挑战

-  Vanilla CF算法无时效性概念
-  新发布商品行为数据稀疏
-  用户/商品画像离线、分散生成
-  格式、逻辑不统一，实时化难度大



时效性应对方案：CF侧


CF召回策略实时化

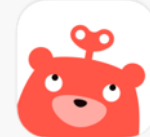
-  基于实时行为召回CF相关商品

CF算法时效性优化

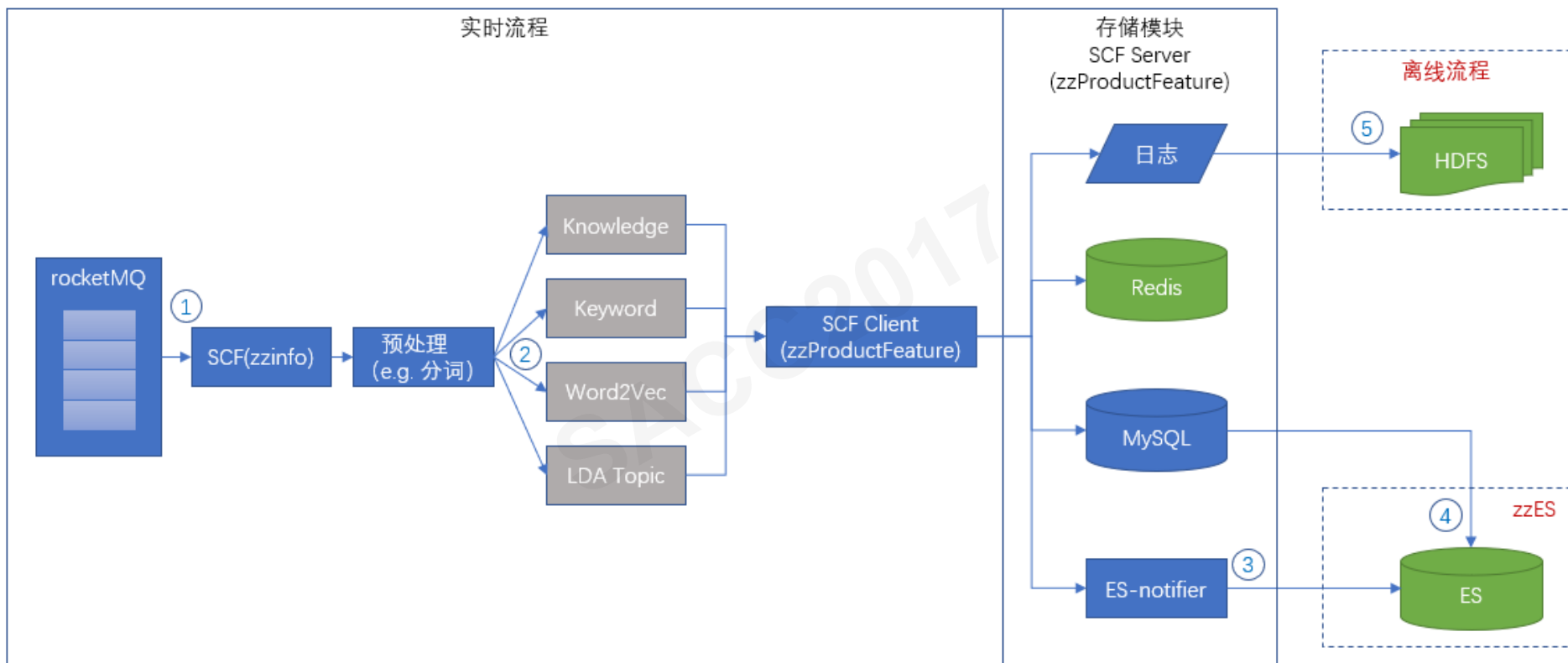
-  使用时效性更强的数据

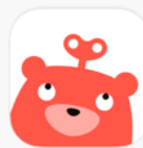
-  鼓励行为时间间隔短的行为

-  使用nearline方式计算近实时增量Cf数据



时效性应对方案：画像侧





时效性应对方案：综合

数据层

- 数据生成实时化
- 生成策略时效性优化

策略层

- 挖掘实时行为
- 商品时效性限定

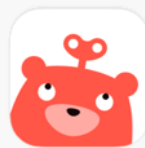


第九届中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2017



挑战3：复杂策略下的性能压力

SACC2017



复杂策略

相关性召回

CF策略 × 6

用户画像策略 × 6

托底策略.....



模型排序

特征查找 × 2

模型预测 × 2

日志记录.....



业务规则

商品过滤

业务降权

信息拼接.....



性能压力

召回

- 外部存储、网络交互
- 策略设计、多步交互

排序

- 特征运算、特征查找
- 模型预测、日志记录


业务

- 属性过滤、规则降权
- 信息拼接、其他需求




Pull-based架构特点

中心思想

 所有操作均在用户请求发生时实时进行

优点

 时效性、新鲜度

缺点

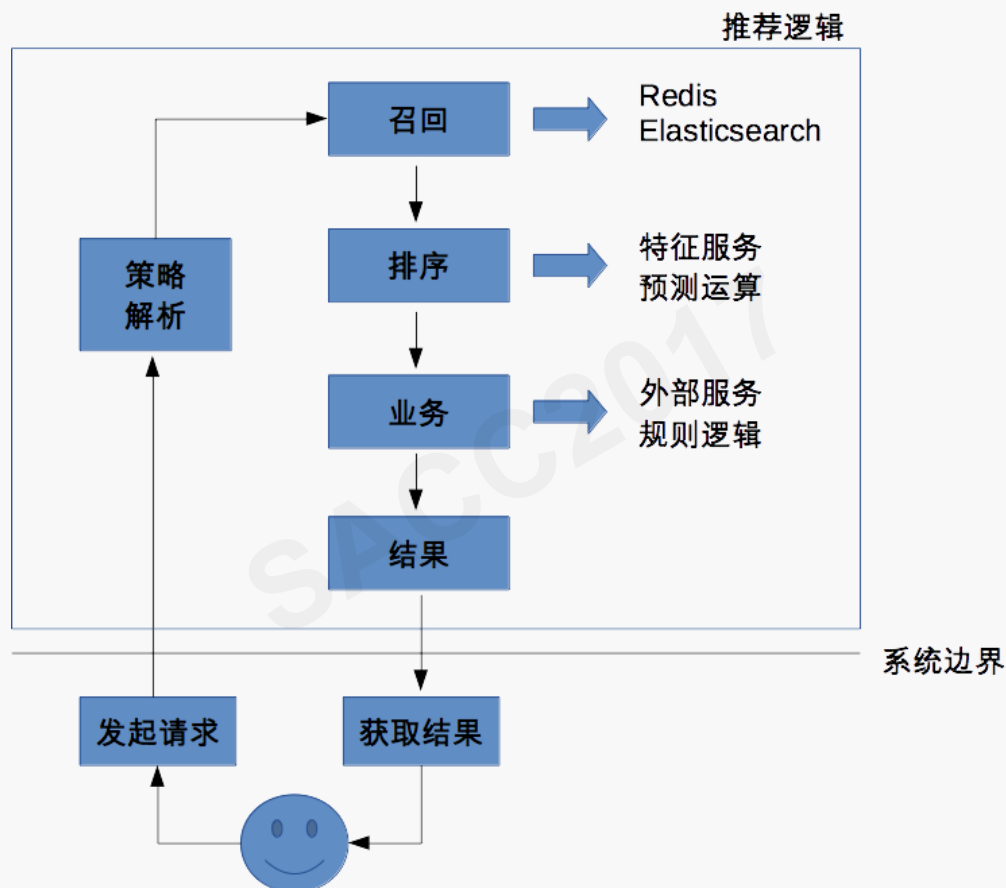
 性能压力大

 扩展难度高

SACC2017



Pull-based架构示意





思考：实时计算的必要性

🐻 是否每个步骤都必须实时计算？

🐻 离线相关策略每天计算一次即可

🐻 实时相关策略可提前进行计算

🐻 还有哪些可行的计算触发时机？

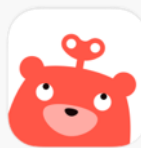
🐻 离线：凌晨计算、定期更新

🐻 在线：行为发生时计算

🐻 牺牲的时效性/新鲜度如何弥补？

🐻 缓存过期

🐻 定时更新



新方案：推拉结合

最终目标


-  将逻辑计算与请求处理尽量分离

-  赋予系统更强的计算能力

推

-  多维度触发时机主动推送数据变更

拉

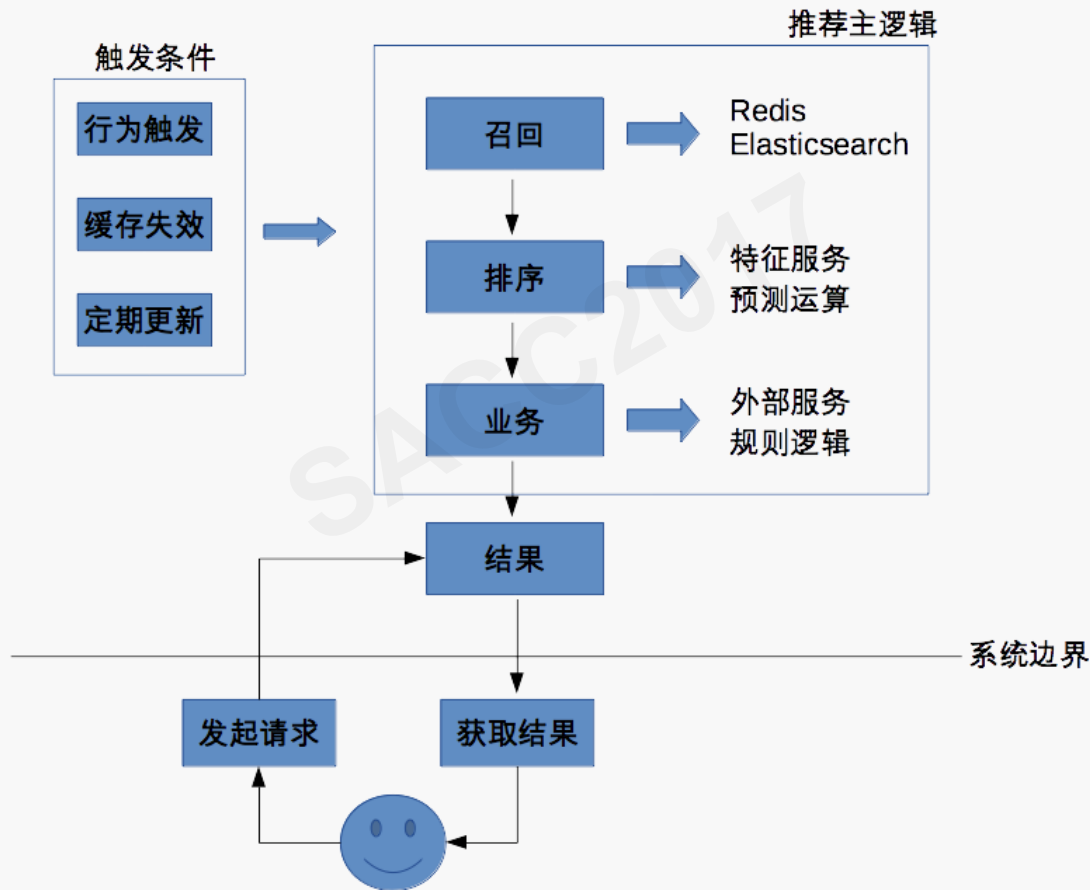
-  请求到来时直接获取计算好的数据

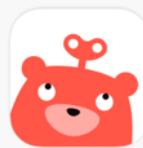
细节

-  缓存过期、活跃度预测.....



推拉结合方案架构示意





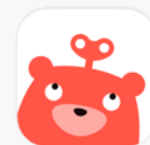
推拉结合优缺点分析

优点

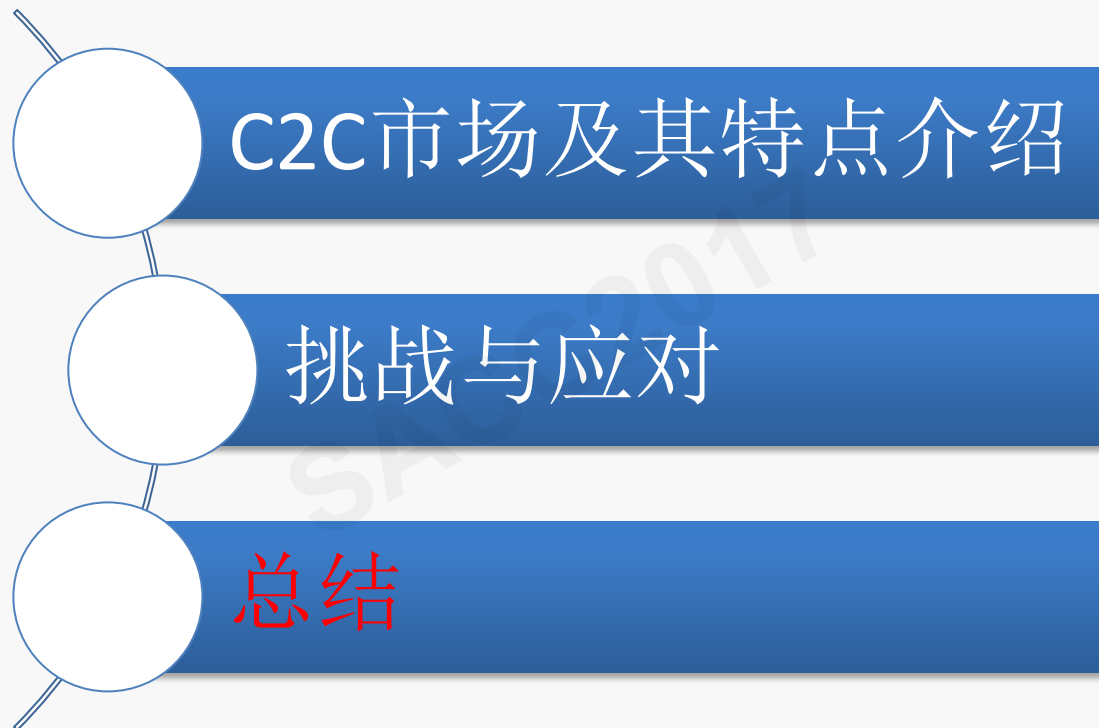
- 计算分离，性能提升
- 近线计算，算能扩容

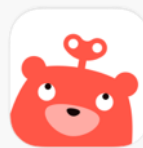
缺点

- 设计复杂，细节繁多
- 新鲜度缺乏足够保证



分享提纲





总结

挑战1：用户发布的数据异质性

挑战2：买卖双方的时效敏感性

挑战3：复杂策略下的性能压力

THANKS

zhangxy@live.com

