携程实时用户意图&AB Test

IT基础业务研发部.BI
元凌峰

技术研发中心 . 基础业务研发部
大数据产品研发团队

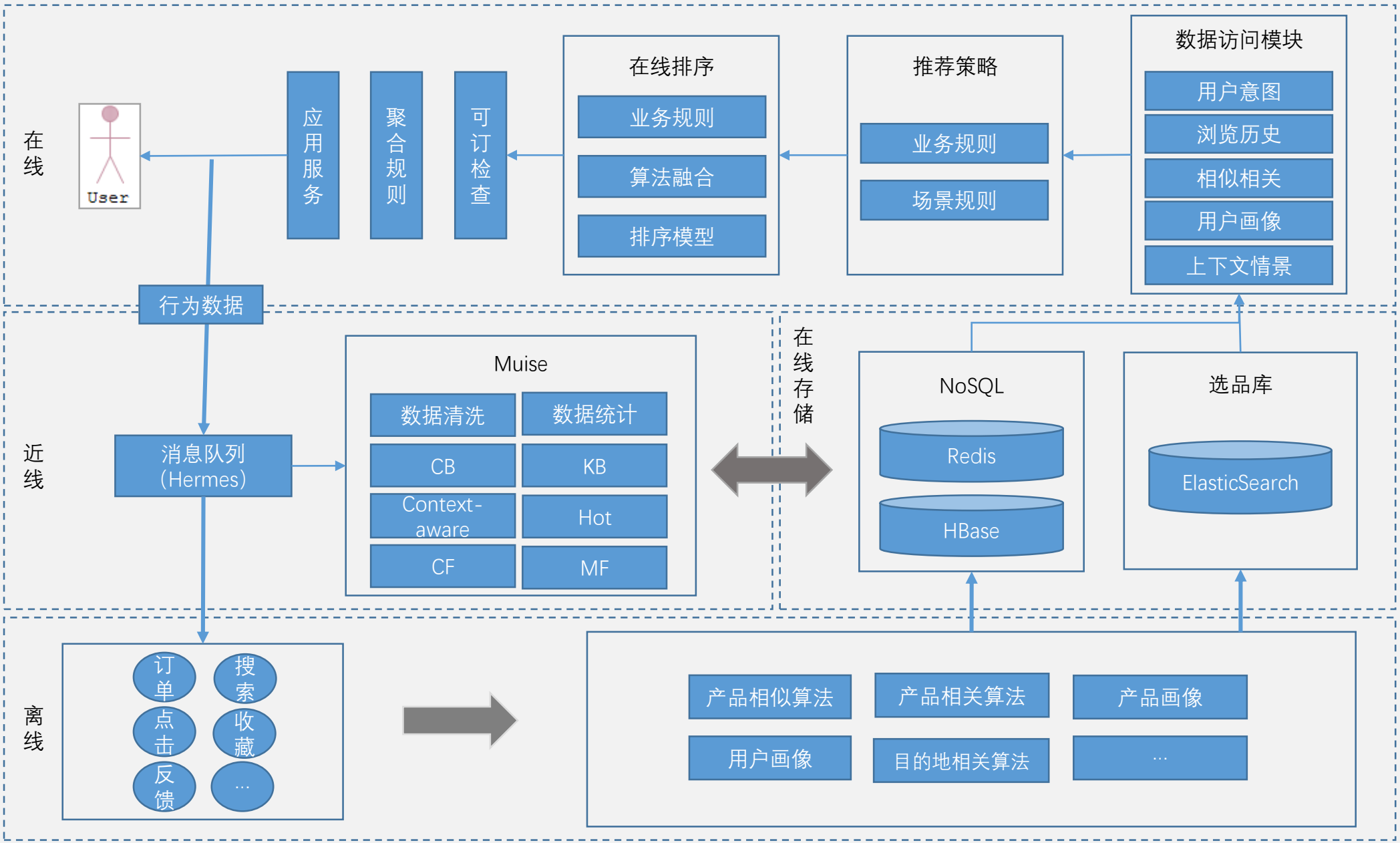
业务发展带来的挑战

- 业务对推荐栏位的需求快速发展
- 推荐系统对用户意图，推荐算法的实时性需求
- 适应多业务，多栏位，多维数据源的用户精准化运营需求
- 业务对AB实验的实时监控

目录

- 1、streaming应用一——推荐场景下的流式计算
- 2、streaming应用二——实时AB Test监控
- Q&A

推荐场景下的流式计算


在线
User
应用服务
聚合规则
可订检查
在线排序
业务规则
算法融合
排序模型
推荐策略
业务规则
场景规则
数据访问模块
用户意图
浏览历史
相似相关
用户画像
上下文情景
行为数据
近线
消息队列(Hermes)
Muise
数据清洗
数据统计
CB
KB
Context-aware
Hot
CF
MF
在线存储
NoSQL
Redis
HBase
选品库
ElasticSearch
离线
订单
搜索
点击
收藏
反馈
...
产品相似算法
产品相关算法
产品画像
用户画像
目的地相关算法
...

推荐场景下的流式计算——实时意图

假设用户是常驻地：南京
(1) 用户已购买北京酒店
10.1入住，10.6离店
(2) 用户在浏览曼谷的攻略
(3) 用户当前在上海开会

探索

曼谷的酒店（市区，四星级）

曼谷的景点

曼谷的玩乐

曼谷的机票

曼谷的用车，美食等

曼谷的度假产品（四钻级，5天左右）

行前

北京当地及周边玩乐

南京<->北京的飞机，火车

北京的景点

北京的用车

行中

上海的景点

上海及周边的玩乐

上海->南京的火车

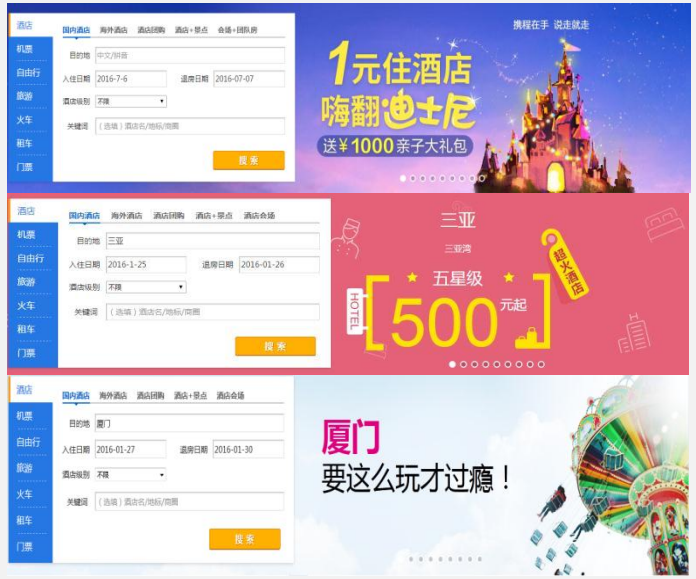
本地

南京的展览

南京当地新奇的玩乐

推荐场景下的流式计算——实时意图

个性化推荐—携程站内动态广告



A版：普通广告　　B版：动态广告

广告转化率

目前，在携程APP、PC、H5上很多广告栏位，都会展示该类型广告

主要使用预测的用户实时意图，实时拼接广告素材，生成动态广告

正在加入用户画像、上下文情境(地理位置，主题)生成一些动态广告。

推荐场景下的流式计算——实时意图

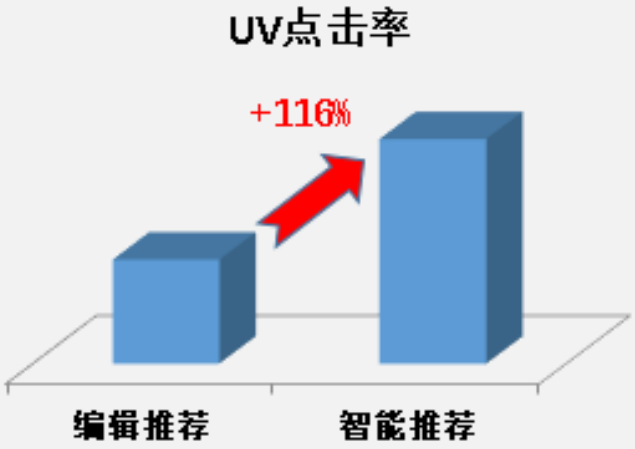
个性推荐—攻略：首页目的地推荐

6.16版本，推荐目的地模块接AB test进行分流，测试模块点击转化率

A版：编辑推荐（**特点：人为编辑更新热门目的地**）

B版：智能推荐（根据用户在携程的浏览、搜索、订单提交等行为，分析挖掘，提供符合用户需求的目的地，**特点：个性化**）

算法策略：用户实时意图+目的地相关+用户画像



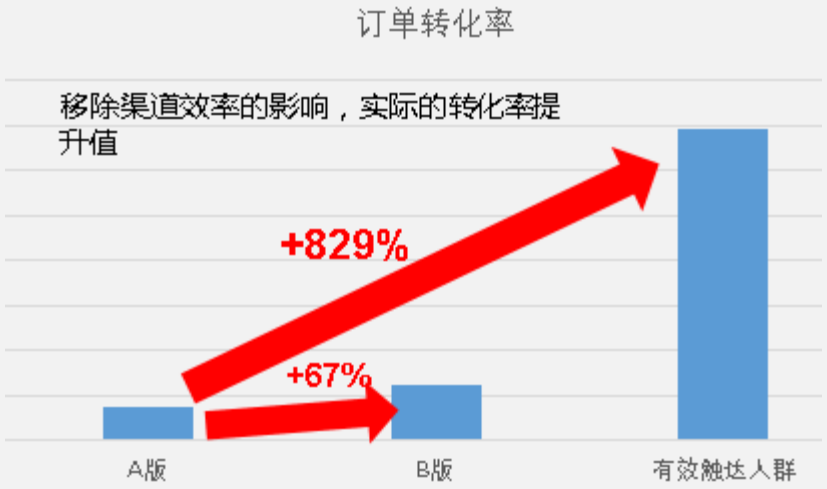
推荐场景下的流式计算——实时意图

个性化推荐—度假：历史同期出行用户营销

用户：历史同期出行的用户

A版：不营销

B版：智能营销触达（基于用户意图提供符合需求的目的地旅行产品，**个性化，千人千面**）



较不触达用户提升**67%，**抛出渠道效率的影响，较不触达用户提升**829%**

推荐场景下的流式计算——实时意图

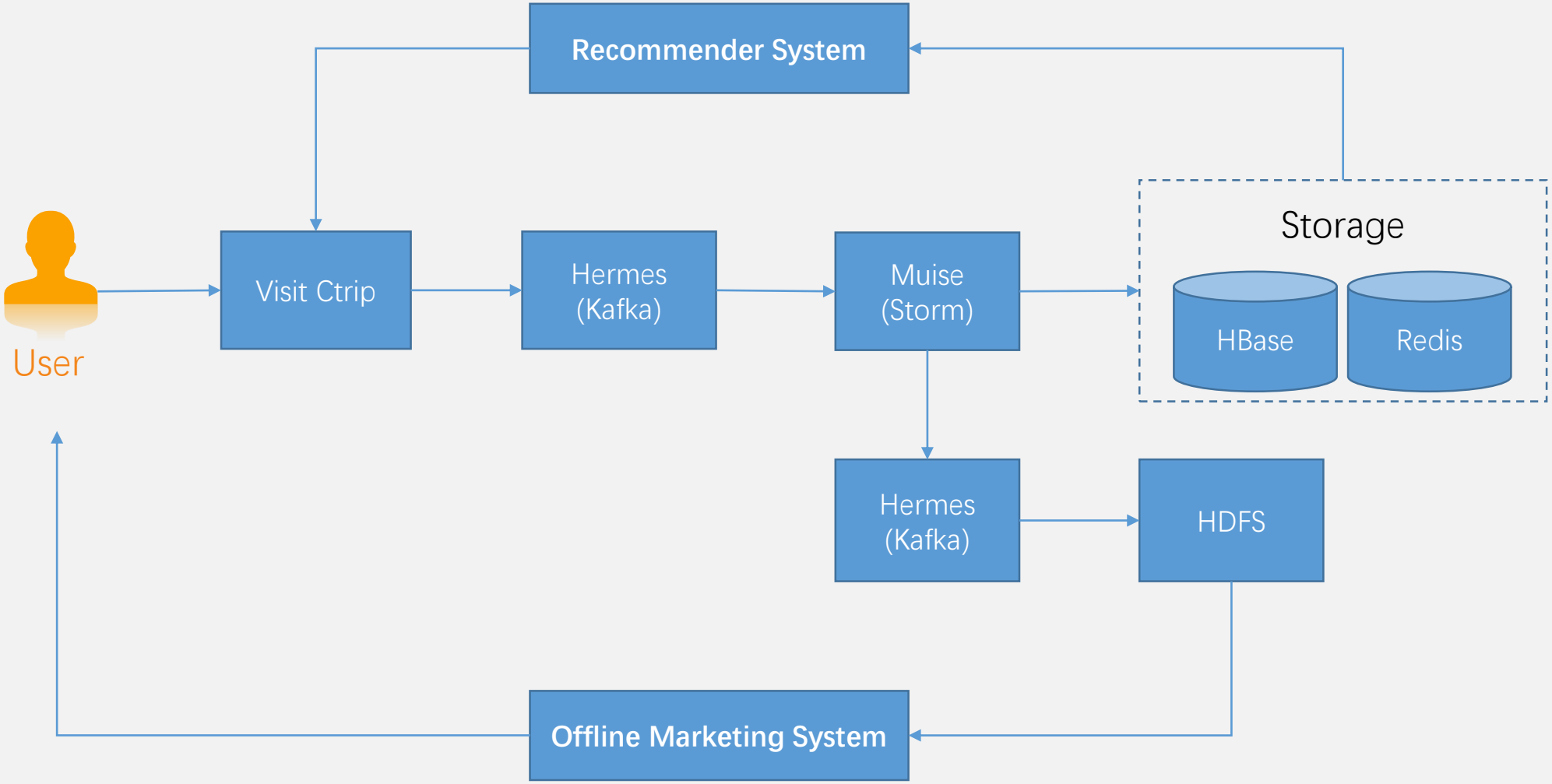
- 范围：实时用户意图目前涉及**12+个业务线**，及相应用户**行为和订单数据**
- 内容：实时意图预测，实时LBS推荐，交叉推荐，订单反向推荐，出行状态推荐，用户权重，行程推荐等
- 应用：目前意图使用场景：**度假个性化首页（牵手游），发现频道，攻略目的地推荐，动态广告，站外广告，站内营销拉新场景等**
- 性能：每天用户行为意图更新在**百毫秒级**
- 采用Redis+Hbase双写，读性能在**几十毫秒**

每日实时处理数据概况	
埋点行为数据量	近十亿条
写入意图数量	数亿条
实时跨平台打通行为，涉及打通数据的读写更新	近十亿条
请求产品维表每天有	千万次
用户基础画像表交互	千万次
实时流处理数据量	数百GB

基础数据	
各业务线维表	近千万级
城市维表	万级
用户画像	亿级
行为特征	千万级
中间表	万级

推荐场景下的流式计算——实时意图

架构及实现



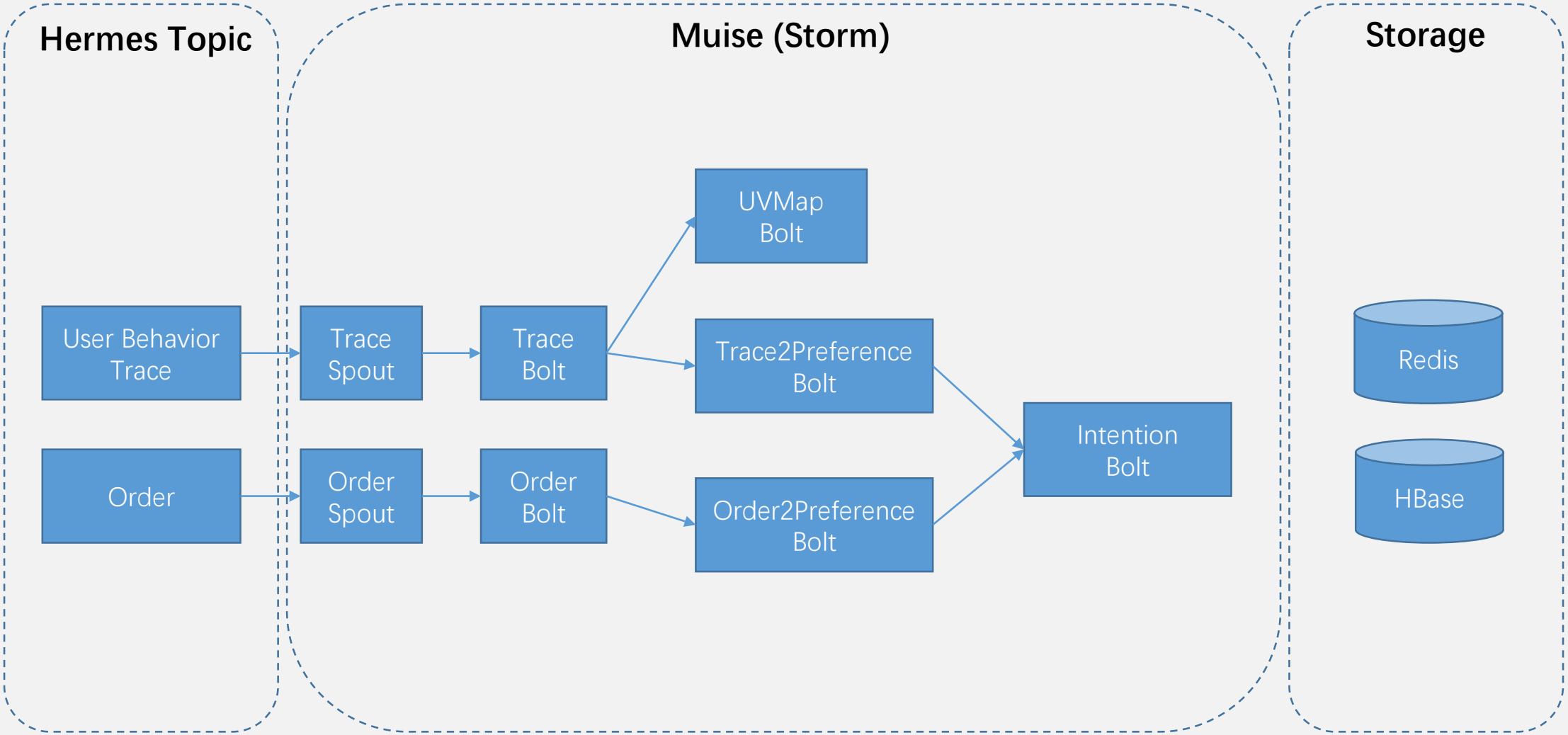
推荐场景下的流式计算——实时意图

架构及实现

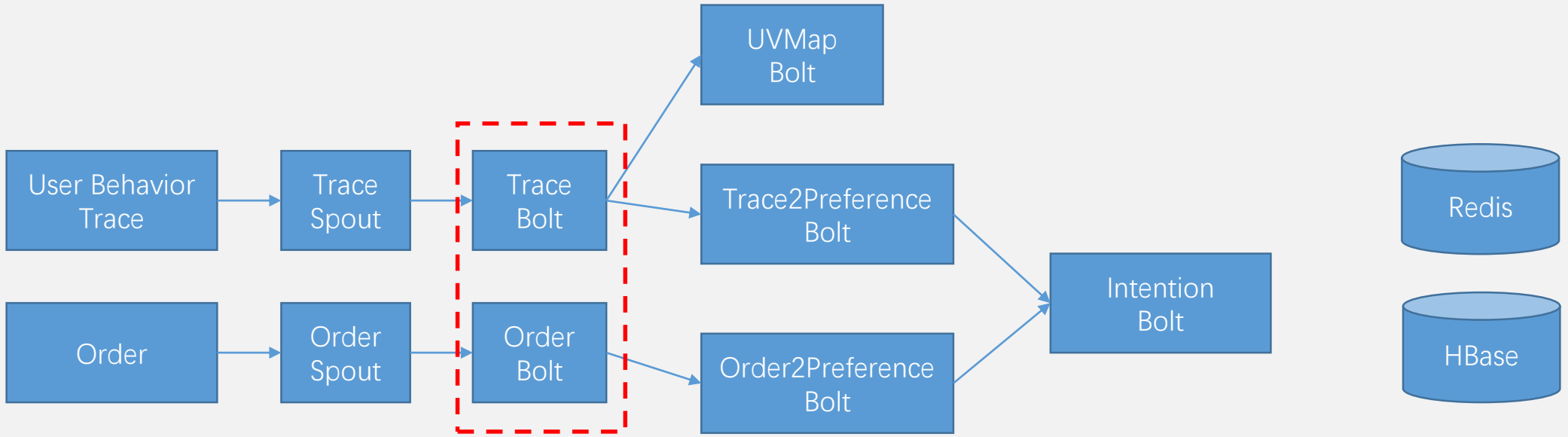
	意图v1.0	意图v2.0
架构	Storm + Redis	Storm + Hbase + Redis
计算内容	行为特征	行为特征、意图预测等
优点	计算量小，速度快	简化online计算量
缺点	Redis要求高可用；online计算量大	多模型+规则引擎，计算较复杂；Hbase的读写IO大

推荐场景下的流式计算——实时意图

架构及实现

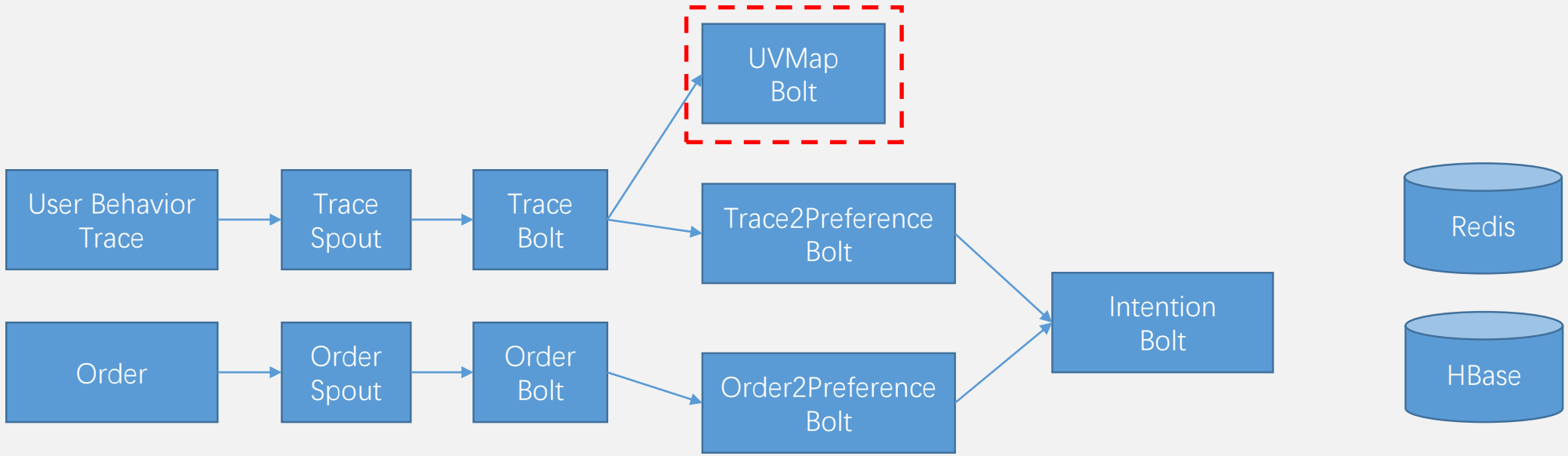


推荐场景下的流式计算——实时意图



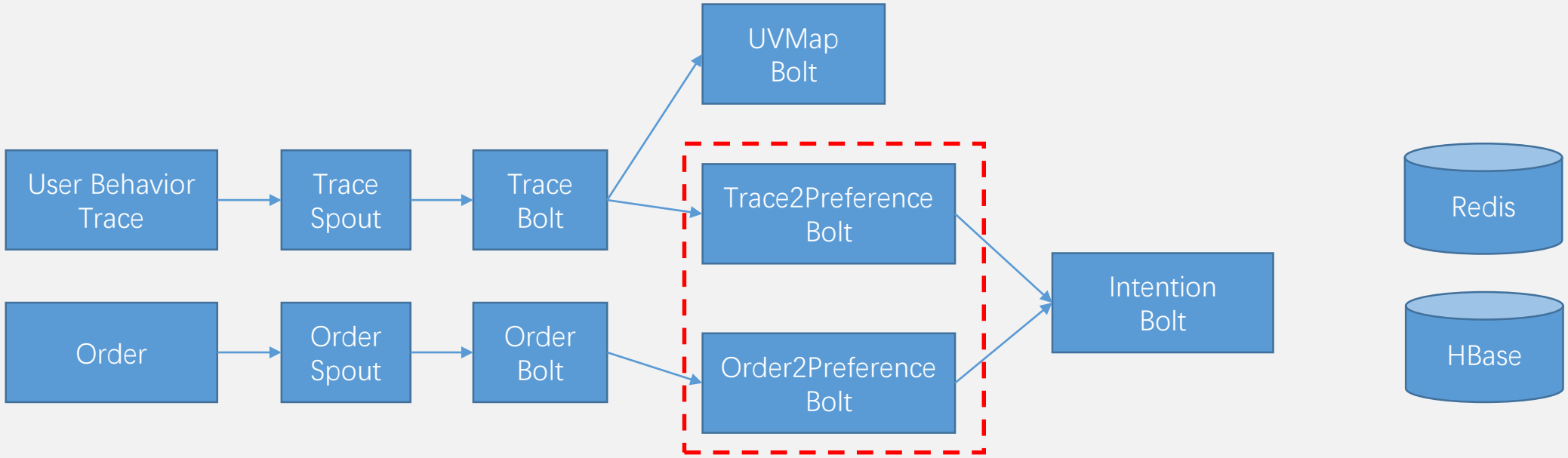
- 实时ETL
  - 行为数据和订单数据的信息抽取
  - 数据过滤
  - 数据Join
  - 数据转化

推荐场景下的流式计算——实时意图



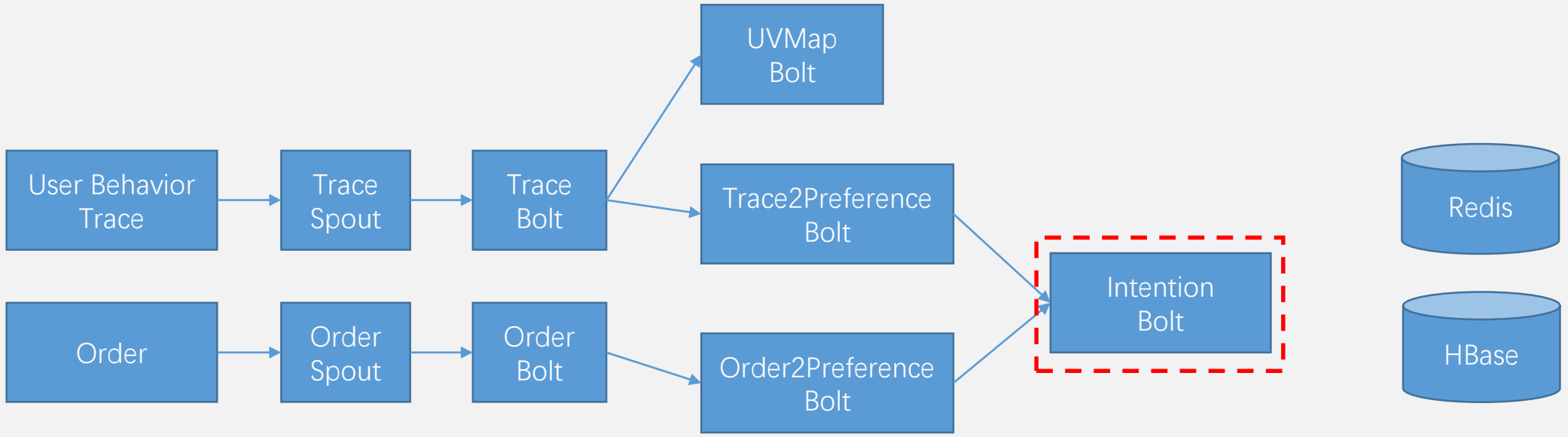
- 实时跨设备平台
  - 更新UID与设备的关系
  - 更新设备与UID的关系

推荐场景下的流式计算——实时意图



- 实时计算用户行为特征
  - 基于滑动时间窗的行为特征
  - 基于牛顿冷却定律的时间衰减
  - 噪声数据剔除
  - 更新用户context信息

推荐场景下的流式计算——实时意图

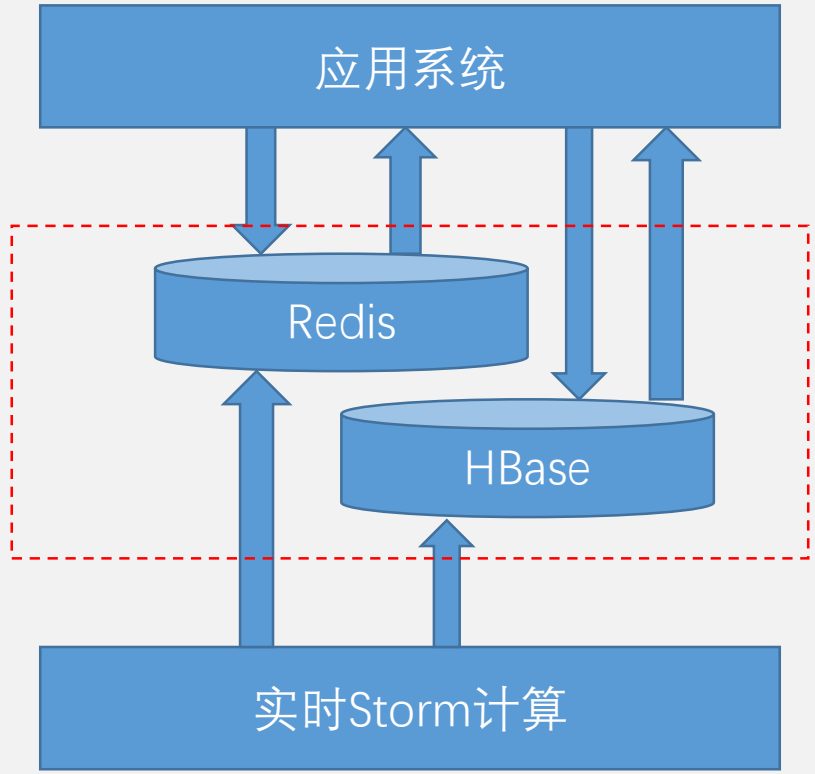


- 实时计算用户意图
  - 基于马尔科夫预测模型的cross-selling，up-selling
  - 基于LBS的推荐
  - 基于行程状态的推荐
  - 规则引擎
    - 订单反向推荐
    - 常驻地推荐

推荐场景下的流式计算——实时意图

存储

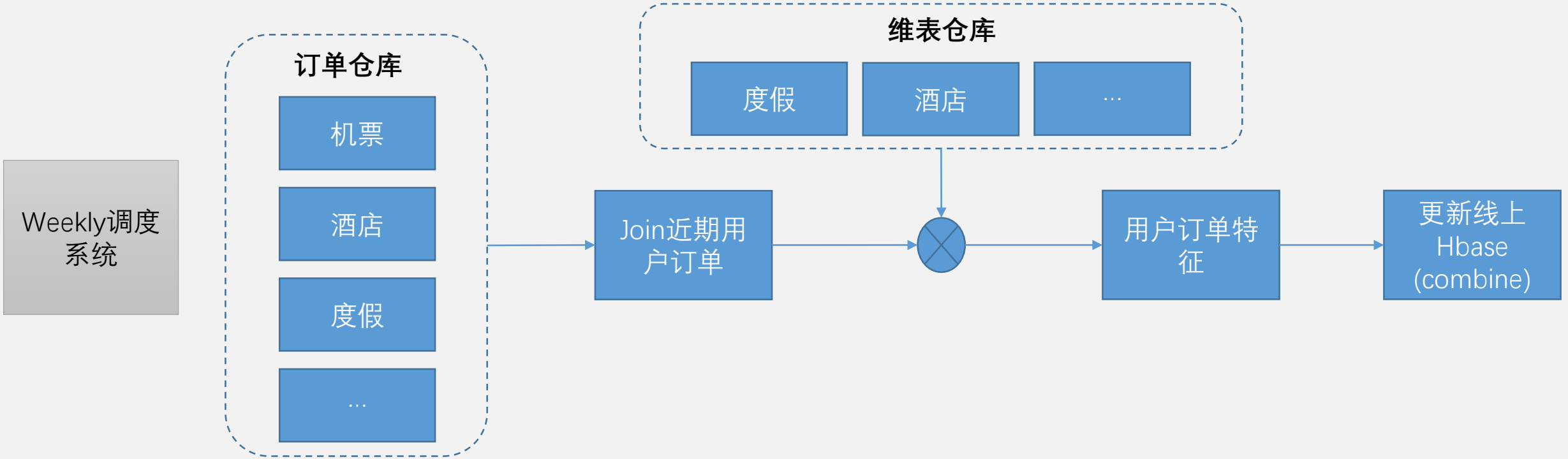
- Storm是无状态的，需要外部存储进行状态保存
- Hbase
  - User Profile，各业务线维表，基础维表
  - 用户实时特征表
  - 用户实时意图表
  - 用户设备关联表
- Redis
  - 用户意图热点数据
- 坑
  - 并发读写问题
  - 网络IO



推荐场景下的流式计算——实时意图

数据回补机制

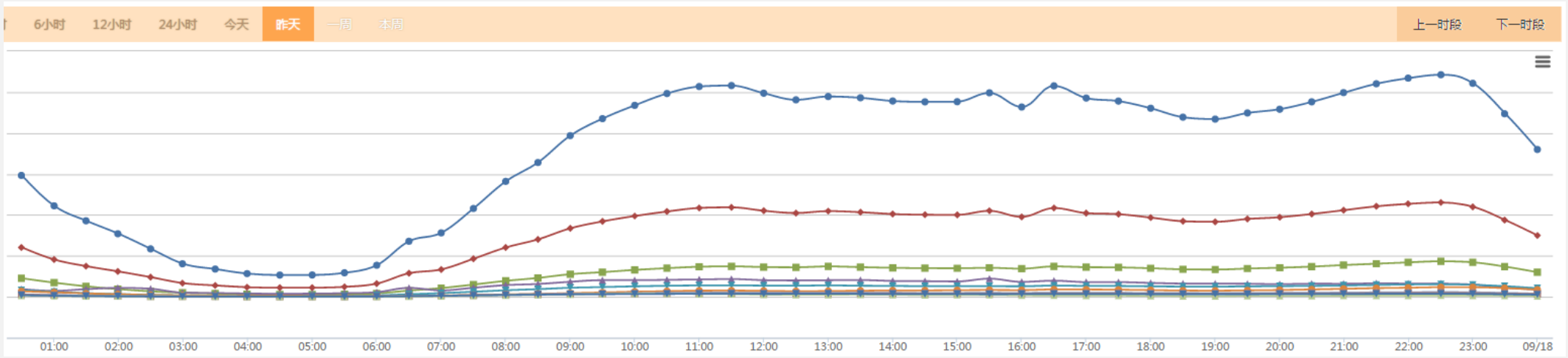
- 线上关键数据与非关键数据
  - 失败重试
- 定期矫正数据



推荐场景下的流式计算——实时意图

监控

- 1、Topology运行情况
- 2、定位问题
- 3、发现计算瓶颈



推荐场景下的流式计算——其他流式推荐模型

基于实时用户行为的CF

- 时间窗
- 短聚合

Item

User

$$\begin{bmatrix} r_{1p} & r_{1q} \\ \dots & \dots \\ r_{up} & r_{uq} \\ \dots & \dots \\ r_{mp} & r_{mq} \end{bmatrix}$$

$$sim(i_p, i_q) = \frac{(i_p, i_q)}{\|i_p\|\|i_q\|}$$

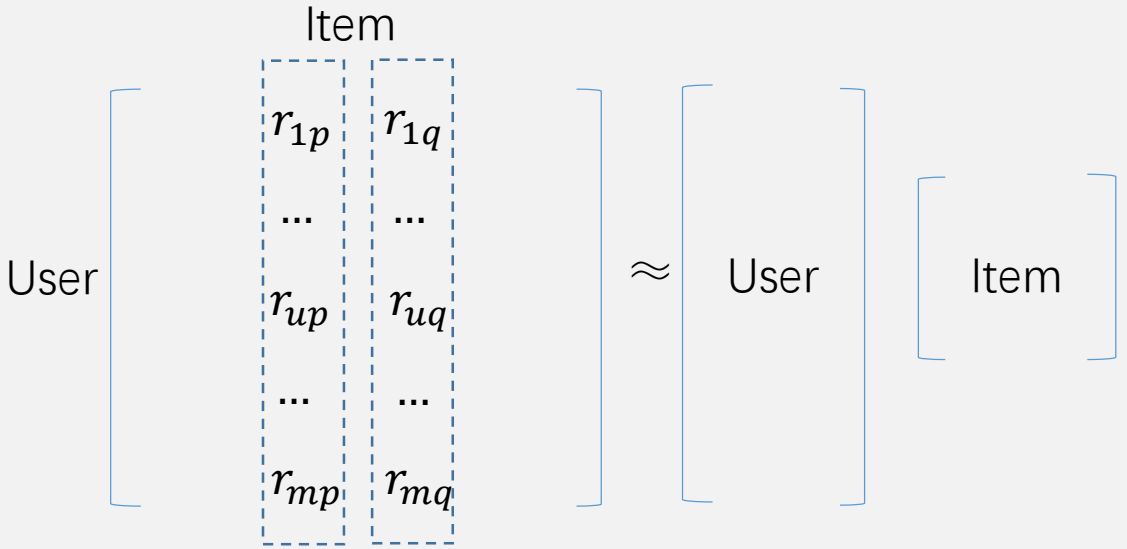
$$sim(i_p, i_q) = \frac{\sum_u r_{up} r_{uq}}{\sqrt{\sum_u r_{up}}\sqrt{\sum_u r_{uq}}} = \frac{PairCount(i_p, i_q)}{\sqrt{ItemCount(i_p)}\sqrt{ItemCount(i_q)}}$$

增量更新： $ItemCount(i_p)' = ItemCount(i_p) + \Delta r_{up}$ $PairCount(i_p, i_q)' = PairCount(i_p, i_q) + \Delta r_{up} r_{uq}$

【Reference】Huang Y, Cui B, Zhang W, et al. Tencentrec: Real-time stream recommendation in practice[C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015: 227-238.

推荐场景下的流式计算——其他流式推荐模型

基于实时用户行为的MF

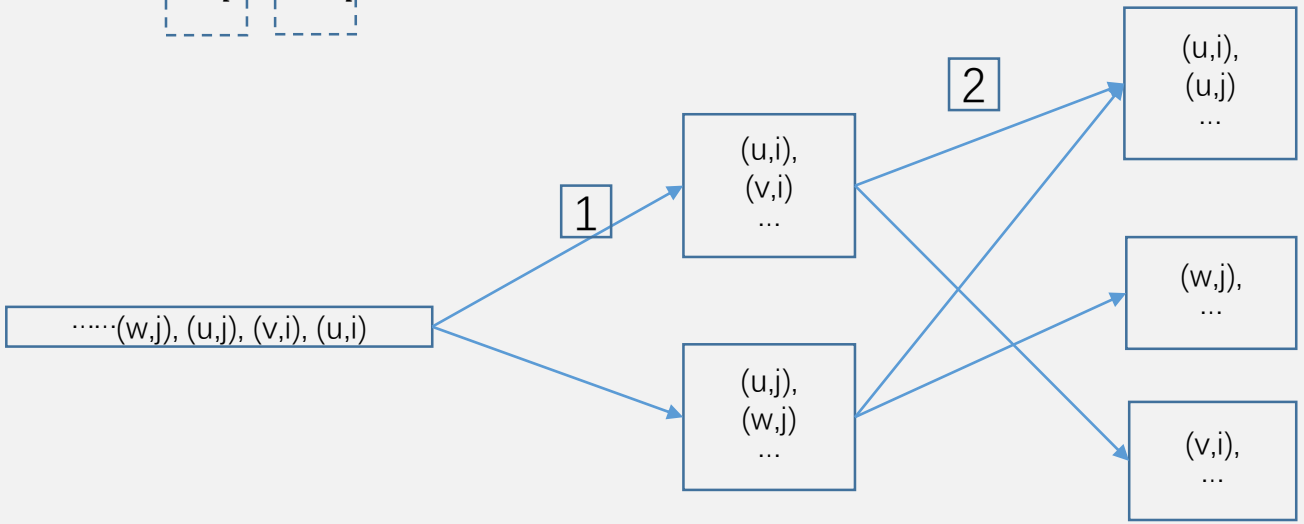


$$R \approx PQ^T$$

$$L = \sum_u (r_{ui} - q_i{}^T p_u)^2 + \lambda[\sum_u \| p_u \|^2 + \sum_i \| q_i \|^2]$$

$$e_{ui} = r_{ui} - q_i^T p_u \qquad q_i = q_i + \eta(e_{ui} p_u - \lambda q_i)$$

$$p_u = p_u + \eta(e_{ui} q_i - \lambda p_u)$$



In parallel way, our streaming computation:

1、group by pid, update q_i according to p_u, p_v

2、group by uid, update p_u according to q_i, q_j

【Reference】Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

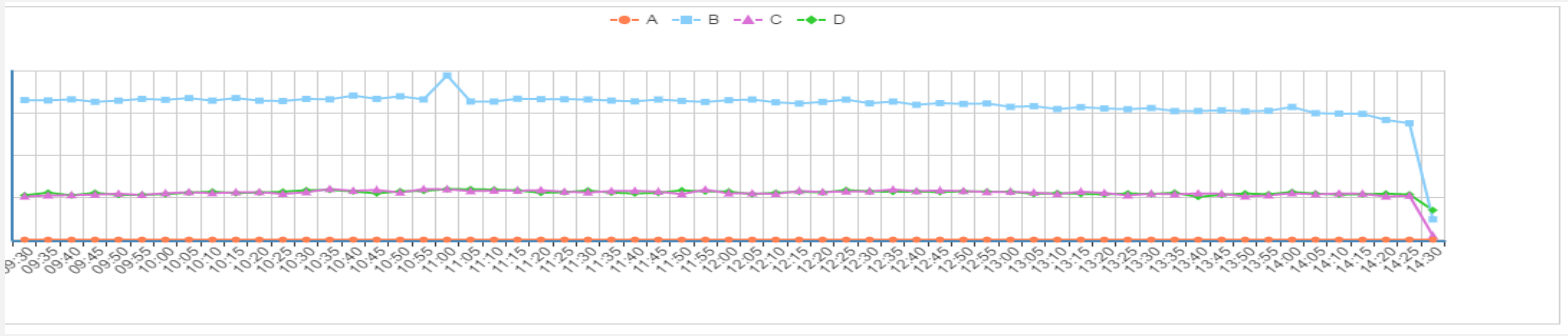
实时AB Test监控

- 目前AB Test每日线上实验情况
  - 每天在线实验：数百个
  - 实时日志数量：十亿级

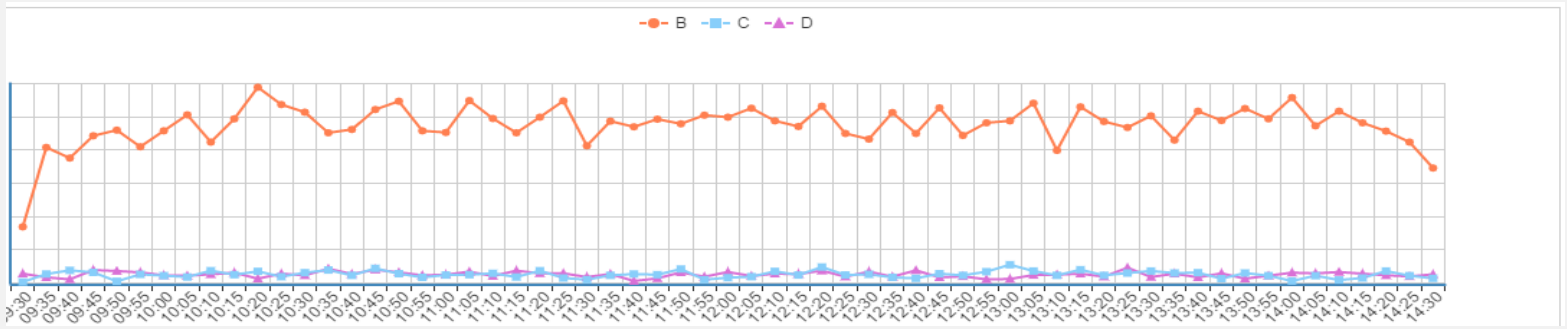
- 线上实验对实时流量及订单的监控需求

实时AB Test监控

- UV量

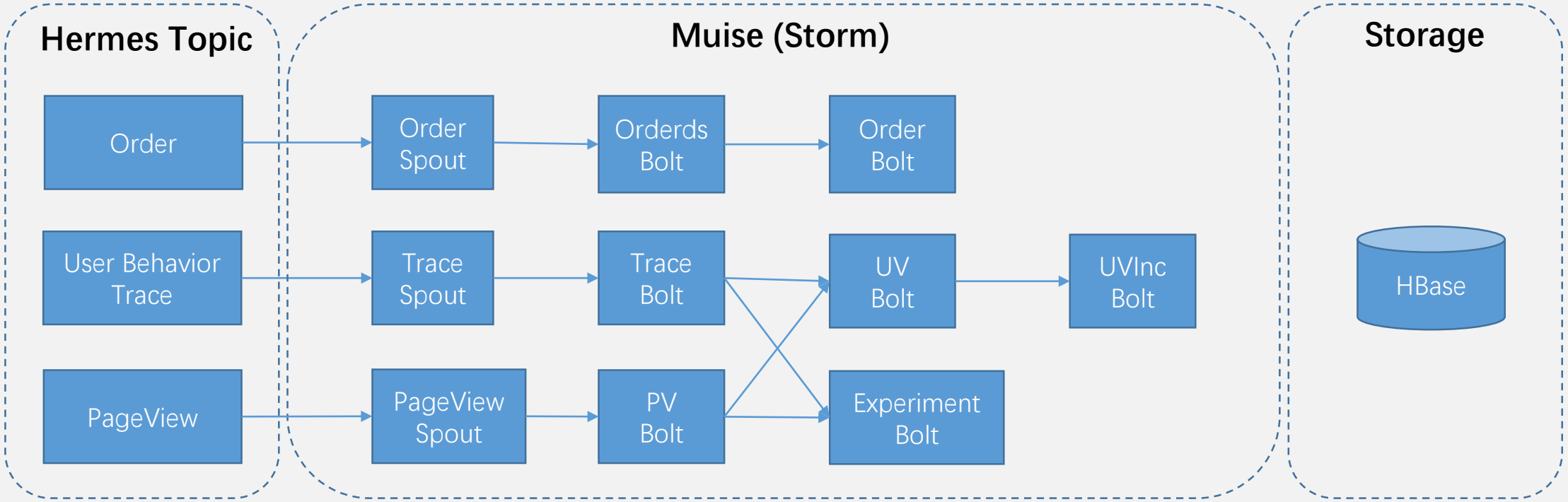


- 订单量



实时AB Test监控——计算流程

- 存储：HBASE
  - ROWKEY设计
  - 批量写
- 计算：STORM
  - 分流数据提取解析
  - BLOOM过滤器去重，UV、订单
  - UV短时间内的聚合



Ctrip 携程

Thank you

Q&A

技术研发中心 . 基础业务研发部

大数据产品研发团队