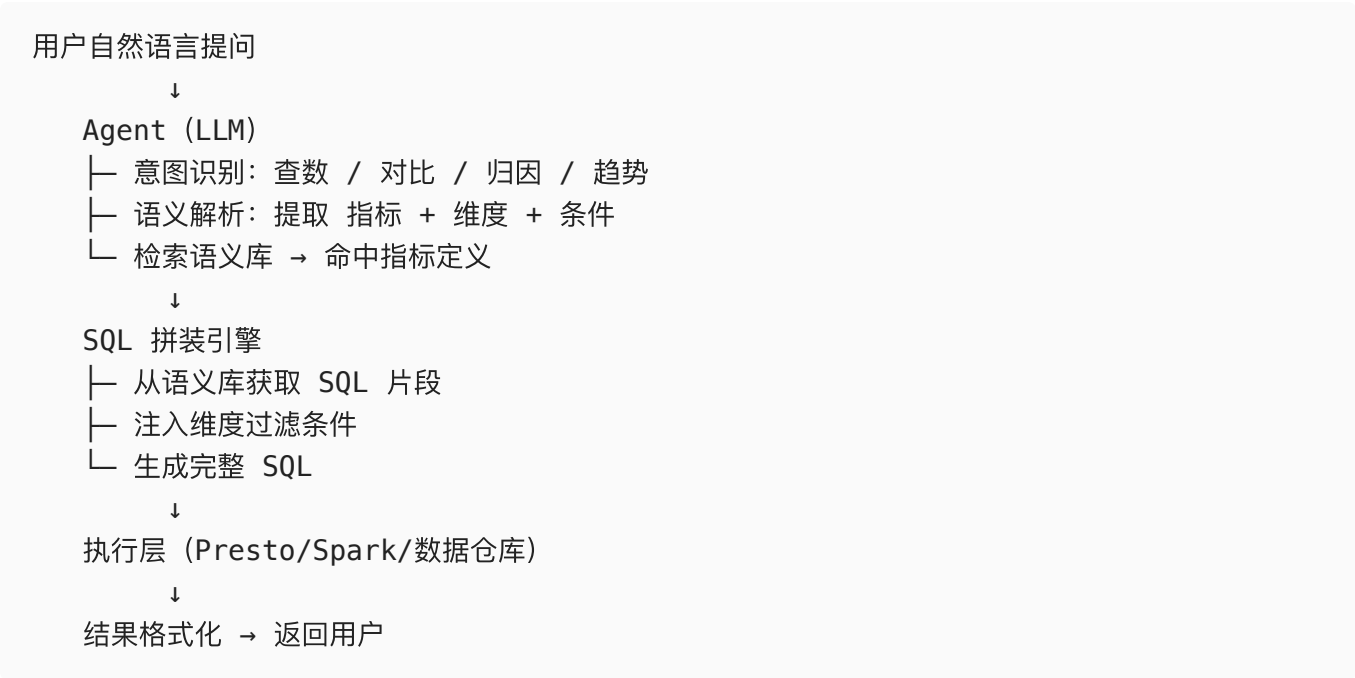


poc-semantic-layer

POC: RCMD Semantic Layer + Agent 数据查询

目标：从 RCMD 团队现有 tracker/SQL 出发，构建指标语义层，让 Agent 能准确回答业务数据问题

1. 整体架构



2. 数据地图构建 (Step 1)

2.1 指标目录提取

来源：

- 现有 tracker 的列定义（已整理 90+ tracker）
- 定时任务 / ETL 脚本中的 SQL
- Confluence 上的指标文档

提取方式：

- 自动：解析 SQL 语句，提取 SELECT 中的聚合表达式 + FROM 表 + WHERE 条件
- 半自动：从 tracker 表头推断指标名，人工补充计算口径
- 人工：业务含义、口径说明、Owner 标注

2.2 RCMD 核心指标清单 (POC Scope)

指标	英文	计算逻辑	常用维度	备注
日活占比	DAU%	$\text{module_dau} / \text{total_dau}$	market, module, date	模块级 DAU 渗透率
订单占比	Order%	$\text{module_orders} / \text{total_orders}$	market, module, date	
曝光占比	Imp%	$\text{module_imp} / \text{total_imp}$	market, module, date	
人均曝光	Imp/UU	$\text{impressions} / \text{unique_users}$	market, module, date	
人均下单	Order/UU	$\text{orders} / \text{unique_users}$	market, module, date	
点击转化率	CTR*CR	$(\text{clicks}/\text{imp}) * (\text{orders}/\text{clicks}) = \text{orders}/\text{imp}$	market, module, algo_ver	端到端转化
GMV	GMV	$\text{SUM}(\text{order_amount})$	market, module, category	含/不含退款需标注
PC2	PC2	platform commission 2	market, module	平台佣金
广告收入	Ads Rev	$\text{SUM}(\text{ad_revenue})$	market, module	
Take Rate	Take Rate	$(\text{PC2} + \text{Ads Rev}) / \text{GMV}$	market	综合货币化率
CTR	CTR	$\text{clicks} / \text{impressions}$	market, module, position	
CR	CR	$\text{orders} / \text{clicks}$	market, module	

2.3 维度定义

维度	英文	取值示例	说明
市场	market	ID, VN, TH, TW, BR, MX, PH, SG, MY	国家站
模块	module	DD, YMAL, PP, Private Domain, Flash Sale	推荐场景
日期	date	2026-01-29	支持 day/week/month 粒度
算法版本	algo_version	v3.2, baseline	AB 实验分桶
品类	category	Electronics, Fashion, FMCG	商品一级品类

维度	英文	取值示例	说明
卡片类型	card_type	product, shop, content	推荐卡片样式
位置	position	top3, top10, all	曝光位置

3. 语义库存储格式 (Step 2)

```
# metrics/ctr_cr.yaml
metric:
  name: CTR*CR
  aliases: ["点击转化率", "端到端转化", "ctr乘cr", "click through conversion"]
  formula: orders / impressions
  # 或分步: (clicks / impressions) * (orders / clicks)
  unit: ratio
  format: percentage # 显示为 0.85%

  source:
    tables:
      - db.rcmd_exposure_log # impressions
      - db.rcmd_order_log # orders
    join_key: request_id

  dimensions:
    required: [market, date]
    optional: [module, algo_version, position, card_type]

  filters:
    default:
      - "status = 'valid'" # 排除无效曝光
      - "is_bot = 0" # 排除机器人

  owner: hongyuan.huang
  team: rcmd-algo
  tracker_ref: "RCMD Algo Key Metric Tracker"

  notes: |
    - DD 和 YMAL 的 impression 定义不同 (DD 是瀑布流曝光, YMAL 是 PDP 页面加载)
    - 含 organic + ads 混合流量, 如需拆分请指定 traffic_type 维度
```

4. Agent 查询流程 (Step 3)

4.1 示例对话

Q: "上周 ID 市场 DD 模块的 CTR*CR 是多少？跟上上周比呢？"

Agent 解析：

```
{
  "intent": "query_compare",
  "metric": "CTR*CR",
  "dimensions": {
    "market": "ID",
    "module": "DD",
    "date_range": "last_week",
    "compare_to": "week_before_last"
  }
}
```


语义库匹配： → metrics/ctr_cr.yaml

SQL 生成：

```
-- 上周
SELECT SUM(orders) / NULLIF(SUM(impressions), 0) AS ctr_cr
FROM db.rcmd_funnel_daily
WHERE market = 'ID' AND module = 'DD'
      AND dt BETWEEN '2026-01-20' AND '2026-01-26'
      AND status = 'valid' AND is_bot = 0;

-- 上上周（同结构，改日期）
```

返回：

上周 ID DD 模块 CTR*CR 为 1.23%，上上周为 1.18%，环比 +0.05pp (+4.2%) 

4.2 复杂查询示例

Q: "最近一个月哪个市场的 YMAL GMV 下降最多？为什么？"

Agent 流程：

1. 识别：指标=GMV，模块=YMAL，时间=最近一个月，意图=归因
2. 先查各市场 GMV 月环比 → 找到下降最大的市场（假设 TH）
3. 对 TH 做维度下钻：按品类、按日期、按算法版本拆解
4. 返回归因结论

5. POC 实施计划

Phase 1：数据地图（1-2 周）

- ☐ 收集 RCMD 团队 top 10 常用 SQL 查询
- ☐ 整理 12 个核心指标的精确计算口径（找 hongyuan.huang 确认）
- ☐ 确认数据源表名、字段名
- ☐ 写成 YAML 语义定义文件

Phase 2：检索 + 拼装（1-2 周）

- ☐ 建指标语义库（向量索引 + 关键词索引双路）
- ☐ 实现 SQL 模板引擎（指标定义 + 维度条件 → SQL）
- ☐ 对接数据查询层（Presto / Spark SQL）

Phase 3：Agent 集成（1 周）

- ☐ LLM prompt 设计：意图分类 + 语义解析
- ☐ 检索 → 拼装 → 执行 → 格式化 pipeline
- ☐ 支持追问 & 对比 & 简单归因

Phase 4：验证 & 迭代

- ☐ 内部试用（先 RCMD algo 团队 5-10 人）
- ☐ 收集 badcase，优化语义匹配和 SQL 生成
- ☐ 逐步扩展到 Search、Ads 团队的指标

6. 技术选型建议

组件	选项	推荐
语义库存储	YAML/JSON + Git	YAML + Git（版本管理、CR friendly）
向量索引	FAISS / Milvus / 简单 embedding	FAISS（POC 够用）
SQL 模板	Jinja2 / 自研	Jinja2（成熟、灵活）
LLM	GPT-4 / Claude / DeepSeek	内部能用啥用啥，POC 阶段 API 就行
执行层	直连 Presto	复用现有查询通道

7. 跟数势方案的区别

	数势 NL2Semantics	SRA 内部版
指标来源	从零定义	从现有 SQL/tracker 反向提取

	数势 NL2Semantics	SRA 内部版
语义层	商业产品（SwiftMetrics）	自建 YAML + Git
口径管理	产品化流程	Owner 标注 + Code Review
适用场景	通用企业 BI	SRA 垂直场景，指标专业度高
优势	产品完整	贴合内部数据架构，可控度高