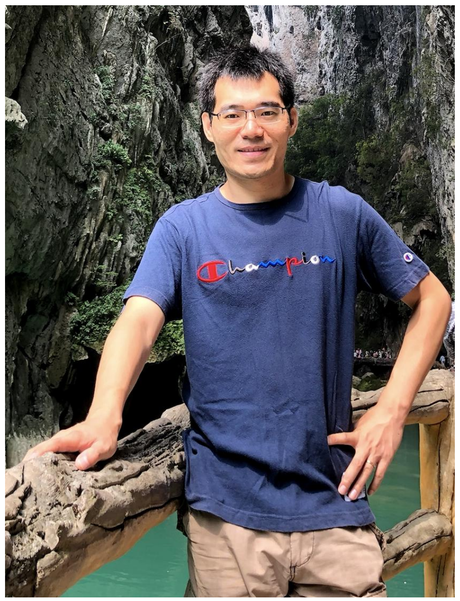
小红书推荐系统介绍

讲师简介



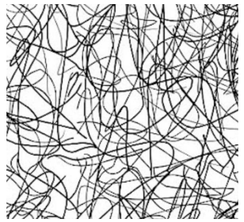
秦波（星爵）
小红书技术部智能分发部-平台架构组

- 小红书推荐引擎(北京)工程负责人
- 全程参与小红书推荐平台的建设，目前正致力于公司内多业务域推广/技术支持中台化推荐平台服务。

大纲

- 小红书推荐引擎介绍
- 小红书推荐引擎核心实现
- 展望

推荐引擎开发背景

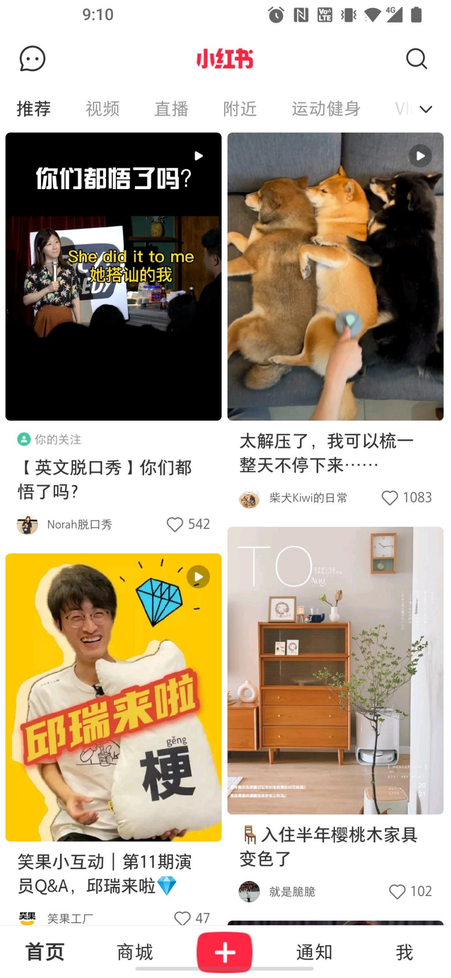


需求繁复，
响应慢

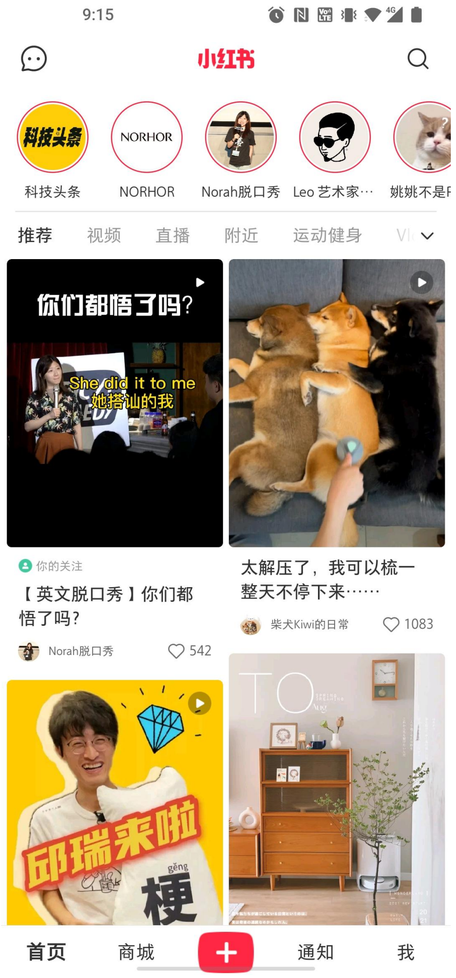
基础框架/引擎沉淀少，
开发姿势各异，维护成本高

人效比变低

多样的推荐业务

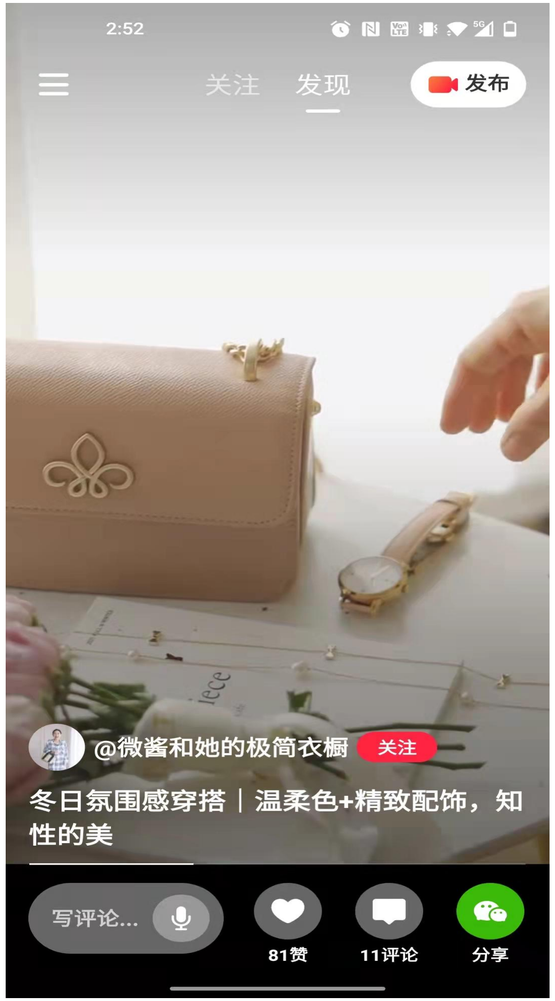


主推荐/相关推荐



人的推荐

电商推荐



新业务推荐

推荐平台建设的核心问题

- 推荐平台的定位
  - 支持主端推荐业务快速迭代
  - 支持新推荐业务/新推荐场景快速上线

推荐引擎建设目标

- 存量业务/新业务统一开发方式
  - 基础框架及各类引擎抽象

    对推荐链路各个服务统一抽象，抽象出底层开发框架及长层若干引擎，实现开发方式的标准化。
  - 新业务上线周期缩到**2周~4周**

    考虑新业务账号及内容独立性，对召回数据/特征模型/冷启动策略等按需拼装，实现新业务快速上线。
  - 同业务域多场景推荐实现一套代码开发

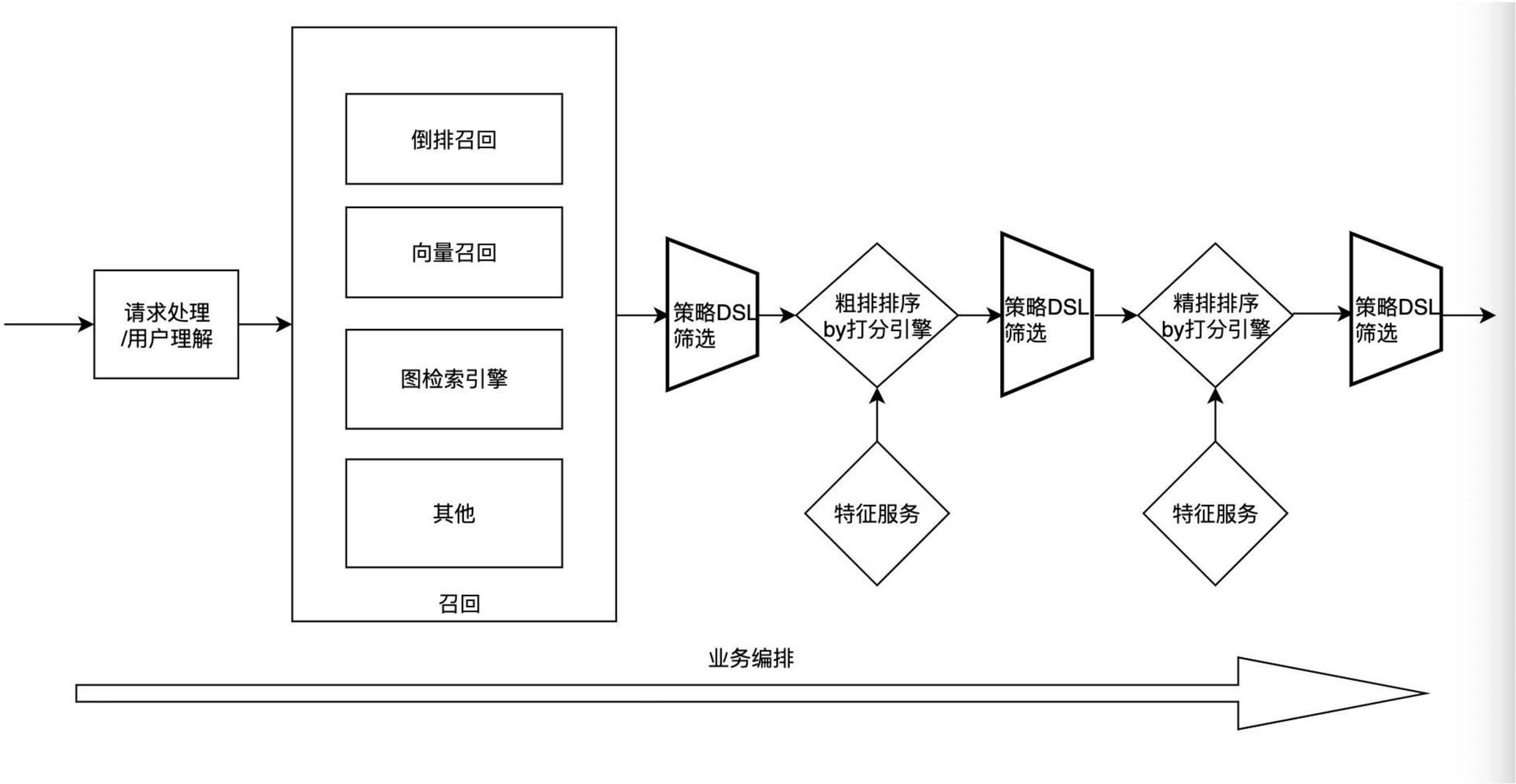
    homefeed/relatedfeed/peoplefeed等推荐业务一套代码支持，多种角色部署。

- 业务开发方式从杂糅式变为配置化/积木式开发

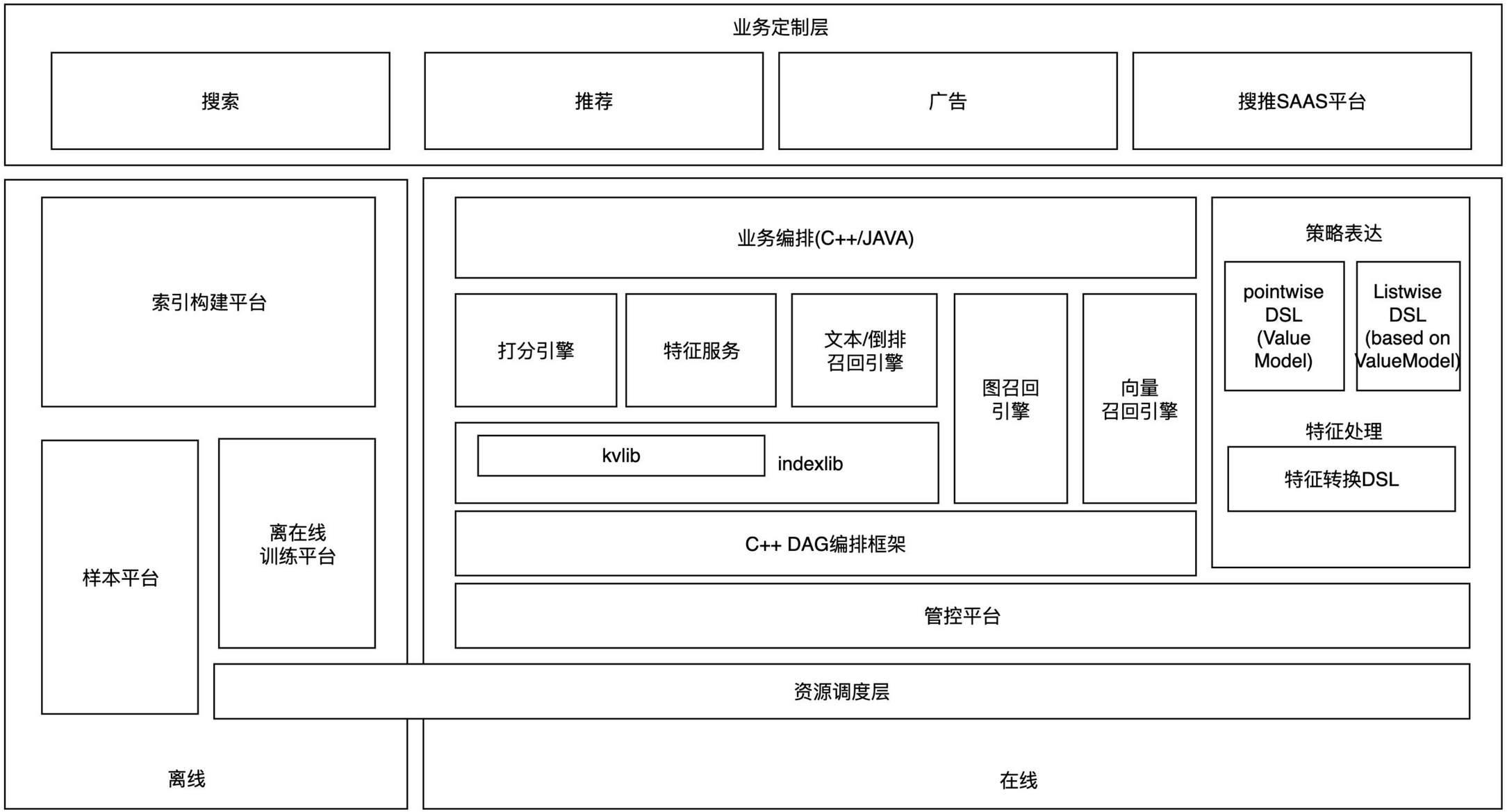
大纲

- 小红书推荐引擎介绍
- 小红书推荐引擎核心实现
- 展望

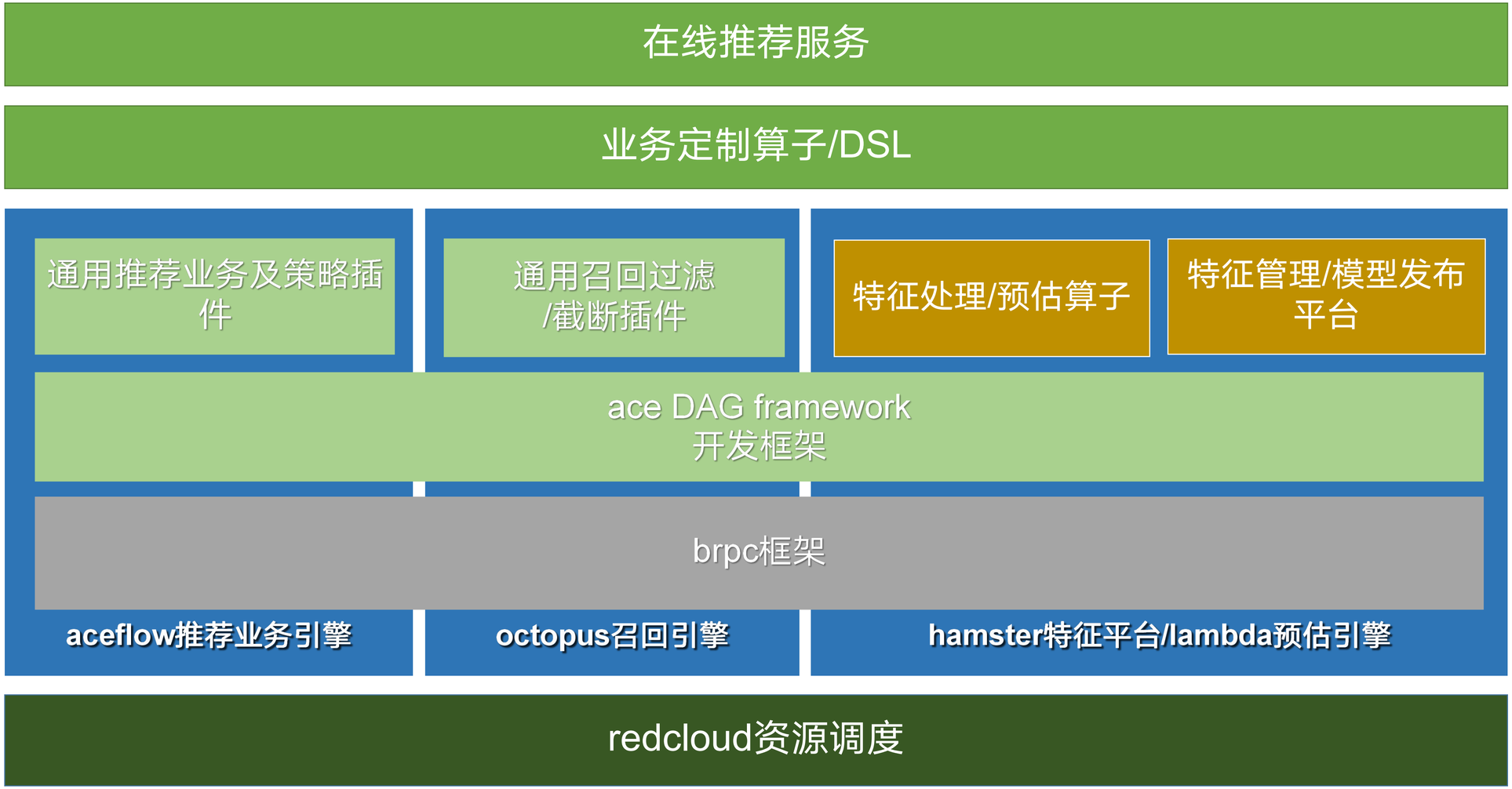
业务视角的推荐流程


请求处理
/用户理解
倒排召回
向量召回
图检索引擎
其他
召回
策略DSL
筛选
粗排排序
by打分引擎
特征服务
策略DSL
筛选
精排排序
by打分引擎
特征服务
策略DSL
筛选
业务编排

核心系统一览


业务定制层
搜索
推荐
广告
搜推SAAS平台
索引构建平台
样本平台
离在线
训练平台
业务编排(C++/JAVA)
打分引擎
特征服务
文本/倒排
召回引擎
图召回
引擎
向量
召回引擎
kvlib
indexlib
C++ DAG编排框架
管控平台
资源调度层
策略表达
pointwise
DSL
(Value
Model)
Listwise
DSL
(based on
ValueModel)
特征处理
特征转换DSL
离线
在线

推荐在线核心服务


在线推荐服务
业务定制算子/DSL
通用推荐业务及策略插件
通用召回过滤/截断插件
特征处理/预估算子
特征管理/模型发布平台
ace DAG framework
开发框架
brpc框架
aceflow推荐业务引擎
octopus召回引擎
hamster特征平台/lambda预估引擎
redcloud资源调度

技术准备-基础框架

- ace DAG framework
  - 基于brpc实现的支持代码逻辑动态配置/加载的通用插件式开发框架。
  - 统一管理外部rpc资源（redis/http/thrift等）/集成词典加载逻辑/业务代码so化/集成部分个性化降级策略/各类标准化打点监控等功能。
  - 业务逻辑以processor方式动态DAG化配置、注册、运行及生命周期管理。
  - C++开发门槛降低到STL数据和容器开发水平。

```
{
    "global": {
        "resource": [
            {
                "resource_type": "redis",
                "resource_name": "user_follow",
                "spec": "corvus-sns-user-follow.service.consul:12345",
                "load_balance": "la",
                "timeout": 50,
                "max_retry": 2
            }
        ],
        "dataloader": [
        ]
    },
    "modules": [
        {
            "module_name": "HelloWorldModule.so",
            "module_path": "/REC-ace/sample/bazel-bin/helloworld_module.so",
            "parameters": {}
        }
    ],
    "processors": [
        {
            "processor_name": "ReqParserProcessor",
            "parameters": {
            }
        },
        {
            "processor_name": "HelloWorldProcessor",
            "parameters": {
                "resource_follow": "user_follow",
            }
        },
        {
            "processor_name": "ResponseProcessor",
            "parameters": {}
        }
    ],
    "apps": [
        {
            "app_name": "hello_world",
            "context_name": "HelloWorldContext",
            "phases": [
                {
                    "phase_stage": 1,
                    "processors": [
                        "ReqParserProcessor",
                        "HelloWorldProcessor",
                        "ResponseProcessor"
                    ]
                }
            ]
        }
    ]
}
```

技术准备-协议和数据统一

- 服务协议标准化
  - 统一的推荐业务引擎服务协议/召回协议/预估协议等。
  - 个性化传参通过协议中扩展字段实现。

- 数据schemaless
  - 离在线用数据datalake格式做schemaless
  - 字段id化管理，字段类型固化为若干类型
  - 字段集中管理，id->name通过配置mapping

```
message RepeatedFloatValue {
    repeated float float_val = 1;
}
message RepeatedIntValue {
    repeated int32 int_val = 1;
}
message RepeatedStringValue {
    repeated string string_val= 1;
}
message RepeatedLongValue {
    repeated int64 long_val = 1;
}

message PayloadV3 {
    map<int32, float> float_map = 1;
    map<int32, int32> int_map = 2;
    map<int32, string> string_map = 3;
    map<int32, int64> long_map = 4;
    map<int32, RepeatedFloatValue> repeated_float_map = 5;
    map<int32, RepeatedIntValue> repeated_int_map = 6;
    map<int32, RepeatedStringValue> repeated_string_map = 7;
    map<int32, RepeatedLongValue> repeated_long_map = 8;
}
```

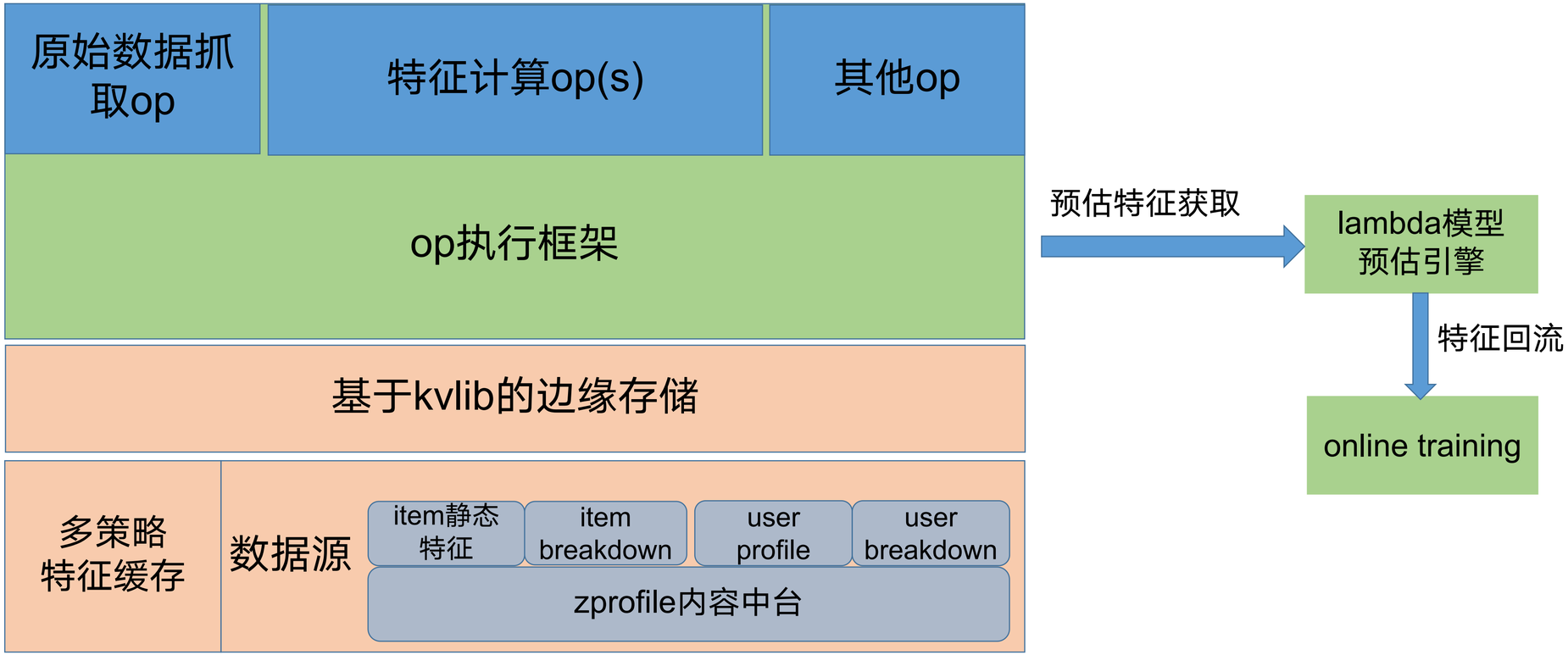
适合推荐的召回引擎

- 解决召回的灵活性问题
  - 索引类型多样性：倒排检索、kv检索、向量检索集成等
  - 过滤的多样性：DSL filter语法，支持自定义的filter插件
  - 截断的多样性：提供基建（统一的词典/schema管理），支持自定义的召回内rank截断插件
- 基于索引构建平台实现索引标准化
  - 通过算法侧沉淀，将召回方式抽象为若干种
  - 打通算法中台，将索引的生产标准化
- 自运维
  - 各项配置自动校验，自动上线

hamster特征平台

- 基于web特征统一管理平台，方便特征管理
  - 统一的特征元数据管理/缓存策略/特征监控等
  - 方便特征跨业务复用
  - 支持离在线特征统一等
- 特征抓取下移到预测引擎，降低预测服务接入成本
  - 笔记预测时只需要传入笔记id即可。
  - 用户侧特征仅需要传入context特征及用户id即可。

hamster特征平台



- 管理所有特征(包括user、item静态动态特征/交叉特征计算), 打造特征全周期一站式管理
- 千级别量级特征管理能力大幅提高
- 离在线统一/特征跨业务复用

hamster特征平台

小红书 Hamster特征平台

- 通用
  - 数据源管理
  - 特征组管理
  - 特征管理
  - 模型管理
  - 操作审计

请选择特征组 *

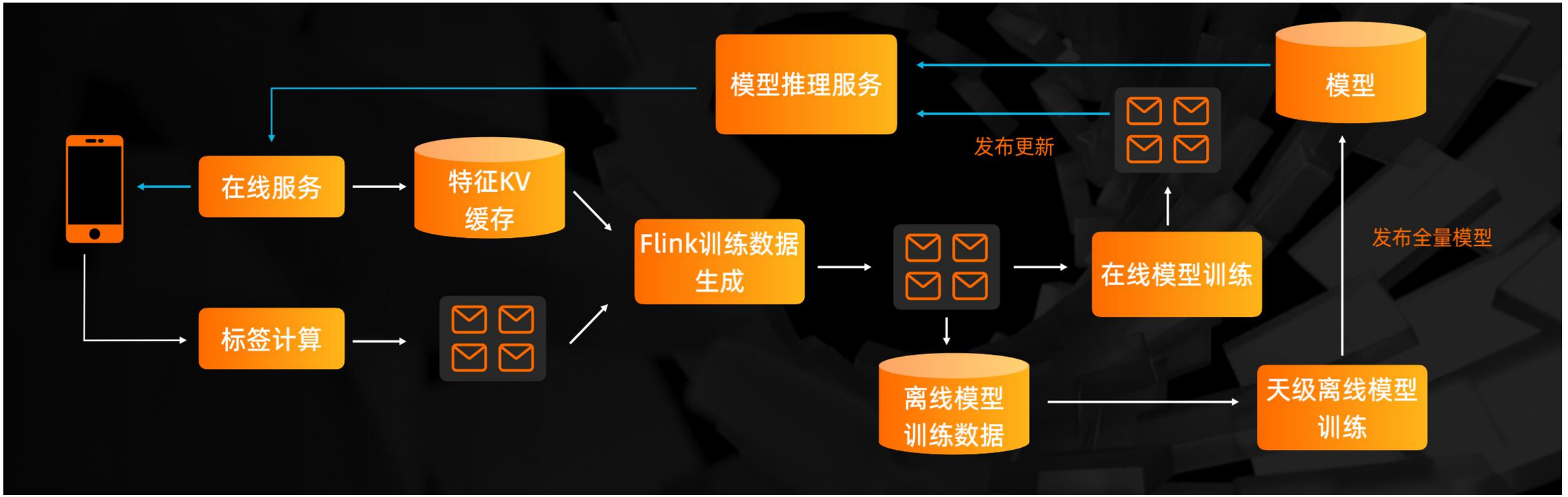
homefeed_note_feature

添加特征

id	名称	字段类型	描述	状态	操作
512	videoAuthorActivePlatform	int	-	在线	上线 下线
768	notePageTime0Exp	int	-	在线	上线 下线
1024	notePlatformAndroidCommentLikeRatioExp	float	-	在线	上线 下线
1280	noteAgeClickYesterday28DExp	int	-	在线	上线 下线
1536	notePlatformAgeClickExp	int	-	在线	上线 下线

5条/页 共 446 页 1 2 3 4 5 跳转至 1 页

larc训练及lambda预测引擎

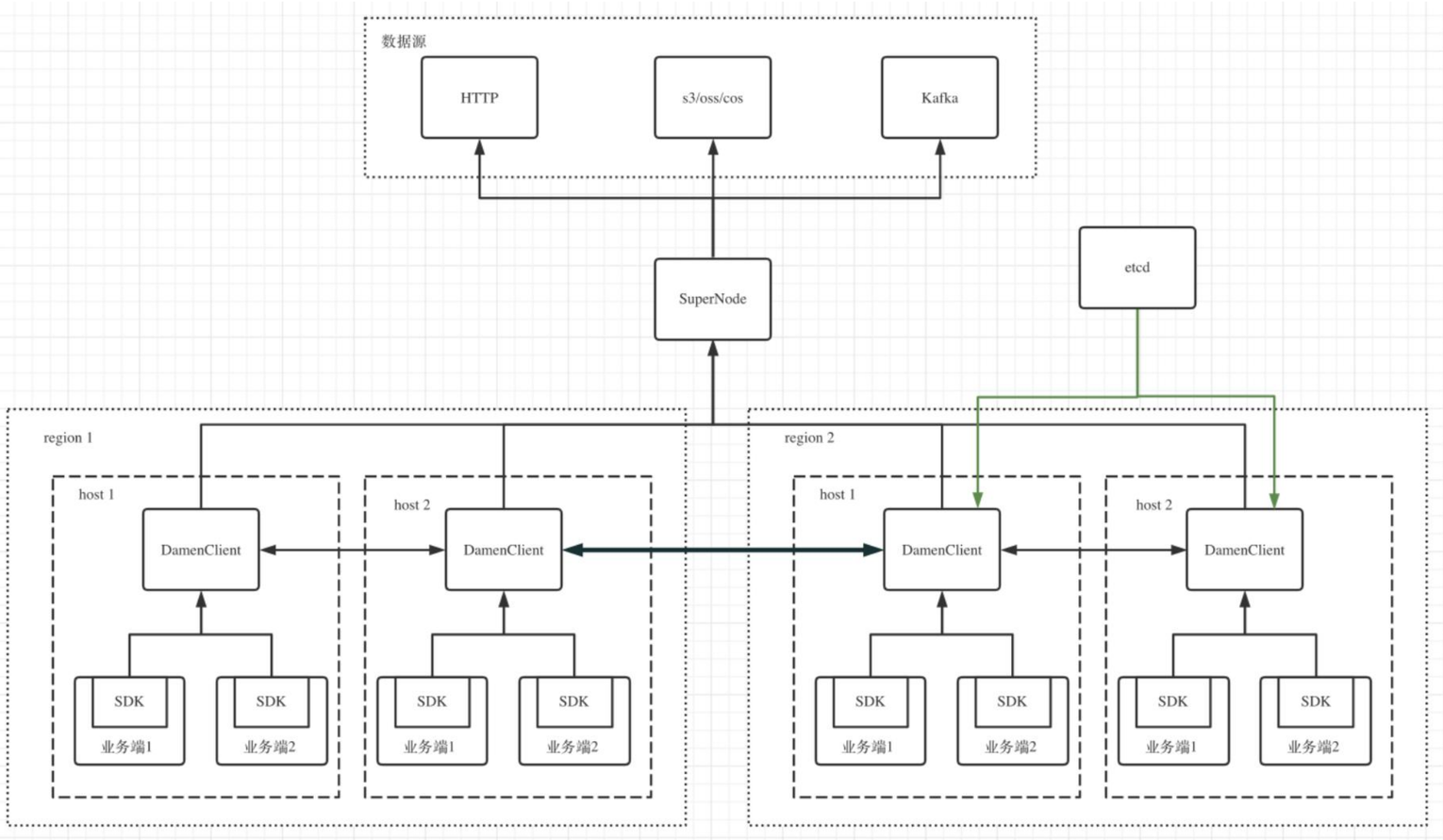


- online training/分布式ps/training ps与online ps参数实时同步
- CPU/GPU在线预估服务
- 弹性/云原生

其他

- redcast数据分发服务
  - P2P方式文件分发/流式的批流一体数据分发，让数据分发更快/成本更低
  - 支持索引/词典/天级模型/实时模型参数等各种数据P2P方式分发，多云场景下大幅降低带宽成本(带宽下降80%)
- EdgeStorage边缘存储系统
  - 支持多种数据类型(kv/kkv)的数据本地读服务建设
  - 离线数据构建打通索引构建平台，并结合redcast支持web化的配置且正排schema实时更新
  - 理论上所有需要rpc方式访问的数据都可以通过EdgeStorage系统转换为本地数据读取。大幅减少在线服务的时延。

redcast数据分发系统



- 支持跨机房，网络传输默认同机房优先级高，跨机房优先级低
- 点对点的网络传输，可设置黑白名单，人工干预网络传播
- 支持按机器，按机房网络限速
- 支持文件分发，支持文件夹同步
- 文件分发支持http, cos, oss, s3等下载方式
- 支持流式分发，数据源支持kafka，客户端接口与kafka相似
- SDK支持命令模式，方便业务方控制文件加载的过程
- 支持丰富的监控，peer状态，数据传播过程，延时，速度等

大纲

- 小红书推荐引擎介绍
- 小红书推荐引擎核心实现
- 展望

展望

- 业界领先的推荐业务引擎/召回引擎/特征服务/模型预估引擎建设
- 在线服务DAG化以及自运维系统建设
- 更高效低成本的数据分发方式的迭代和推广（P2P分发+边缘存储方案）