

众安金融实时特征平台实践

郭育波 大数据开发高级专家



目录 CONTENT

- 01 众安金融MLOps简介
- 02 实时特征平台架构设计
- 03 实时业务特征计算详解
- 04 反欺诈场景的特征应用

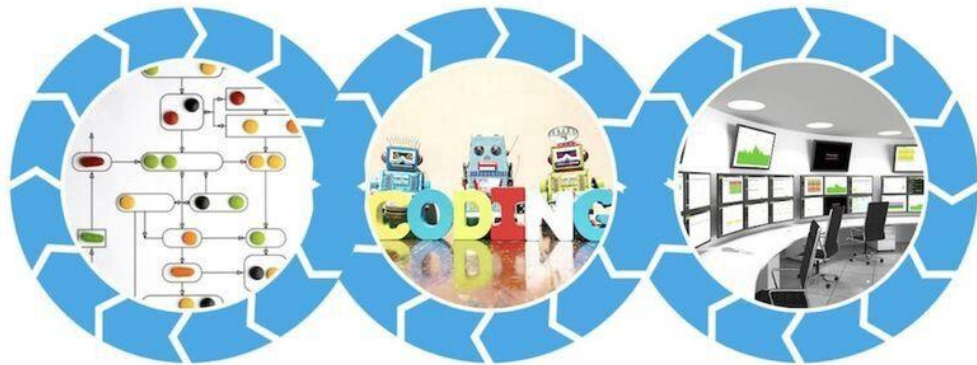
01

众安金融MLOps简介



什么是MLOps?

$$\text{MLOPS} = \text{ML} + \text{Dev} + \text{Ops}$$



ML

Experiment, Train & Tune
ML to find
Optimal Model

Dev

Develop CI / CD
Pipelines to deploy
Models

Ops

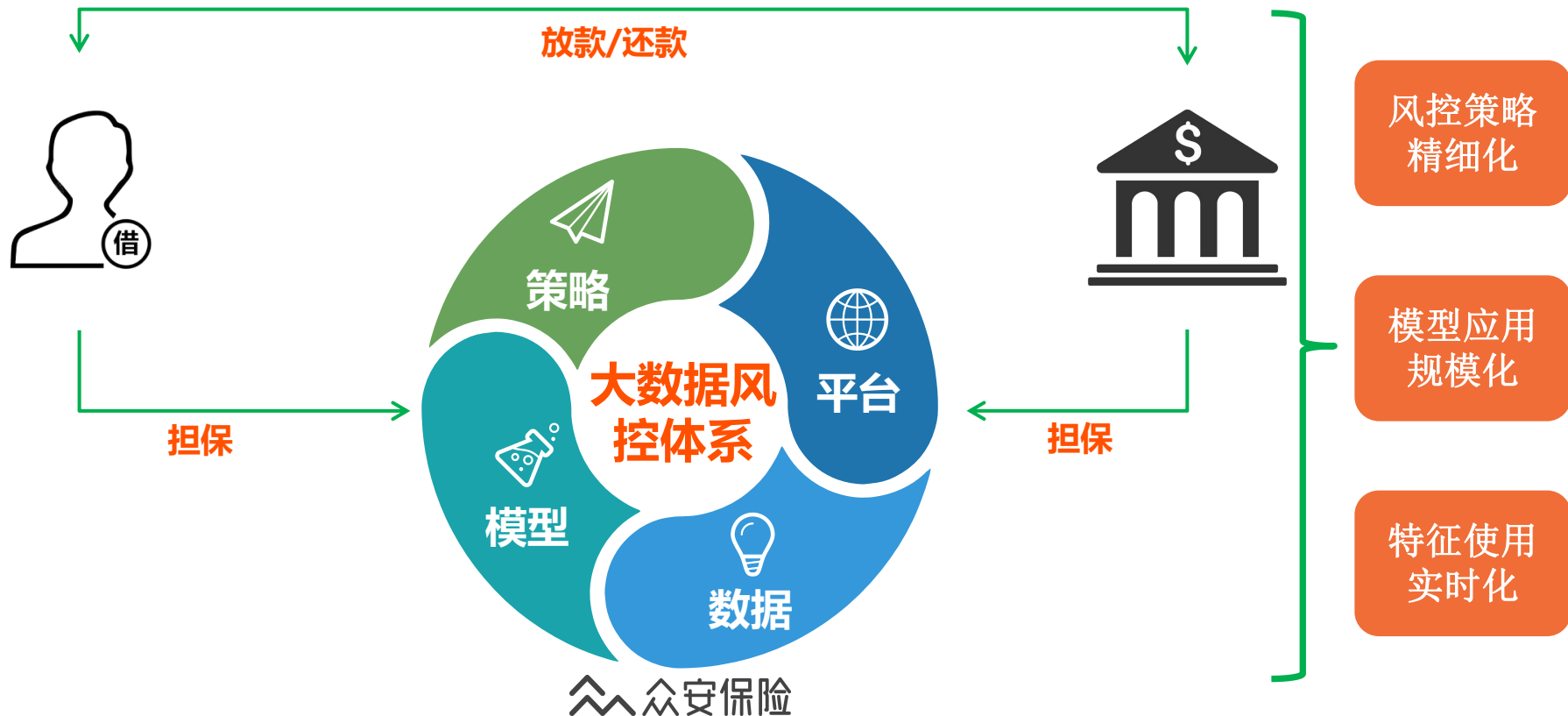
Monitor & Manage
Models
at scale

MLOps是将机器学习、数据工程和 DevOps 融合在一起，从而实现机器学习模型的高效迭代，持续稳定的应用于生产业务的一套方法架构。

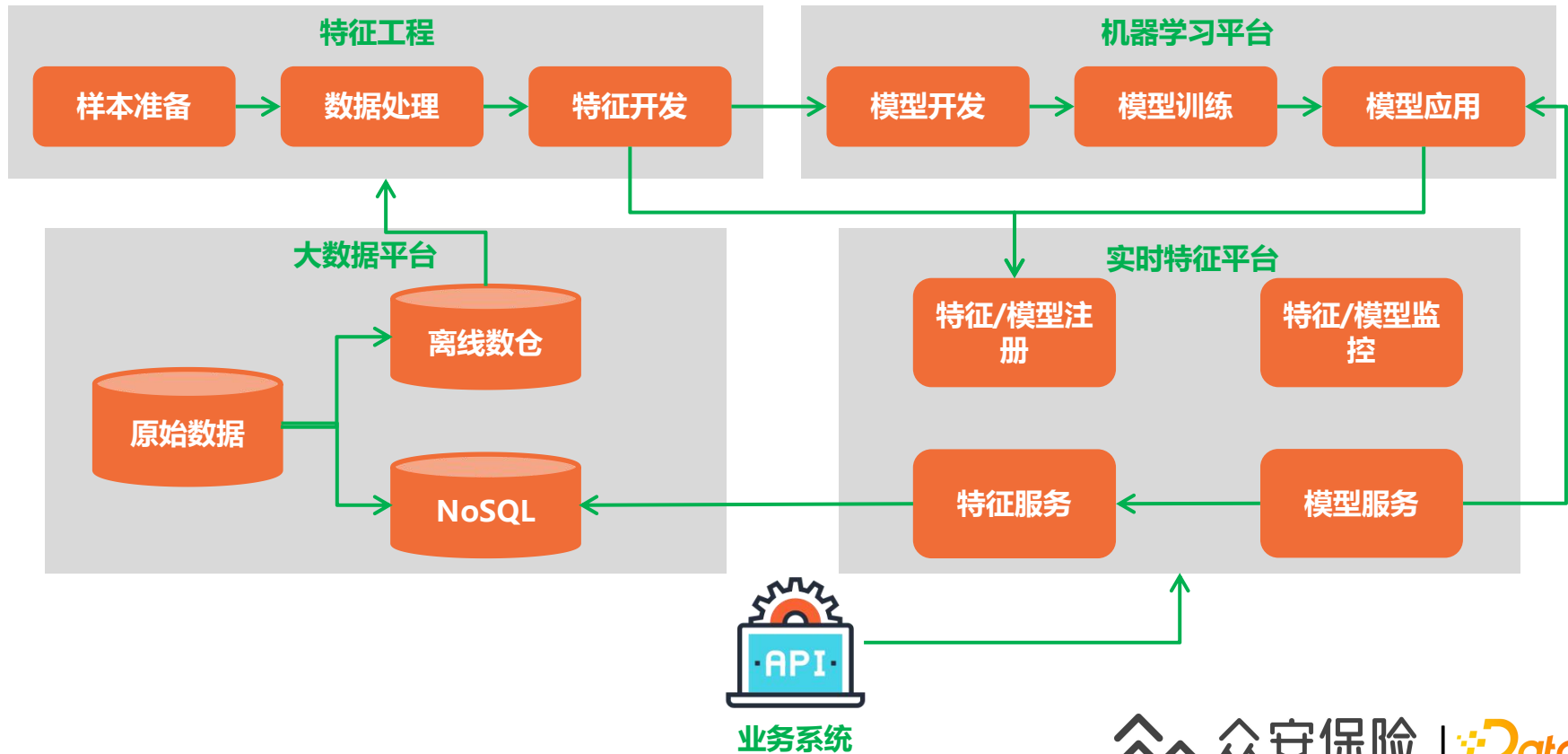
众安金融MLOps流程说明



众安金融为什么需要建设MLOps?



众安金融MLOps体系



02

实时特征平台架构设计



众安金融特征数据分类



交易行为数据

借款申请行为数据、还款行为数据、调额和逾期数据



三方征信数据

三方征信数据里面丰富的个人征信数据进行金融特征衍生



设备抓取数据

用户授权的位置信息、IP地址、设备指纹等数据



用户行为数据

用户行为埋点数据进行特征的衍生

众安金融特征应用场景



实时特征平台核心能力

丰富的数据接入能力

支持配置化的接入丰富的数据源

实时的数据处理能力

实现原数据的实时收集和特征实时计算



快速的系统响应能力

提供高吞吐和低延迟的系统服务能力

高效的特征配置化能力

支持可配置化的特征加工能力

实时特征平台的业务架构

业务应用

前置平台

反欺诈平台

决策平台

额度平台

资金平台

营销平台

保后平台

特征网关(鉴权限流, 特征编排, 数据路由)

特征配置

三方特征配置

实时特征配置

互斥特征配置

模型特征配置

特征计算

三方特征计算

实时特征计算

反欺诈特征计算

模型特征计算

特征管理

特征变量管理

模型特征管理

关系图谱管理

特征跑批管理

特征监控

特征调用链路

特征告警监控

特征统计大盘

效果统计分析

实时特征平台

特征数据源

征信数据...

离线/实时数仓...

三方数据...

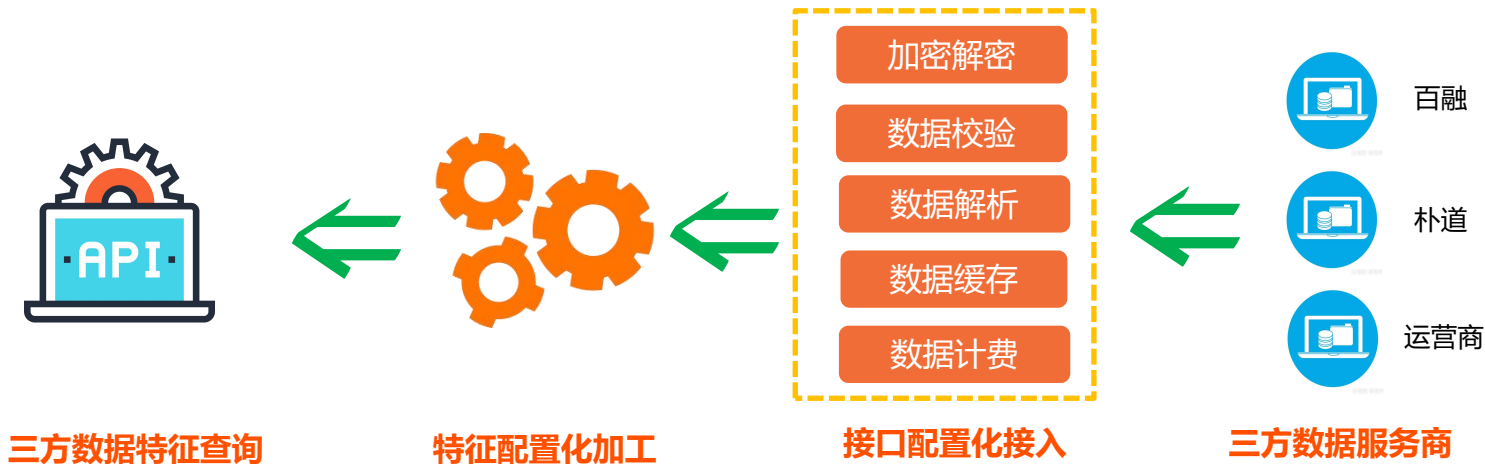
征信数据网关

离线调度平台

实时计算平台

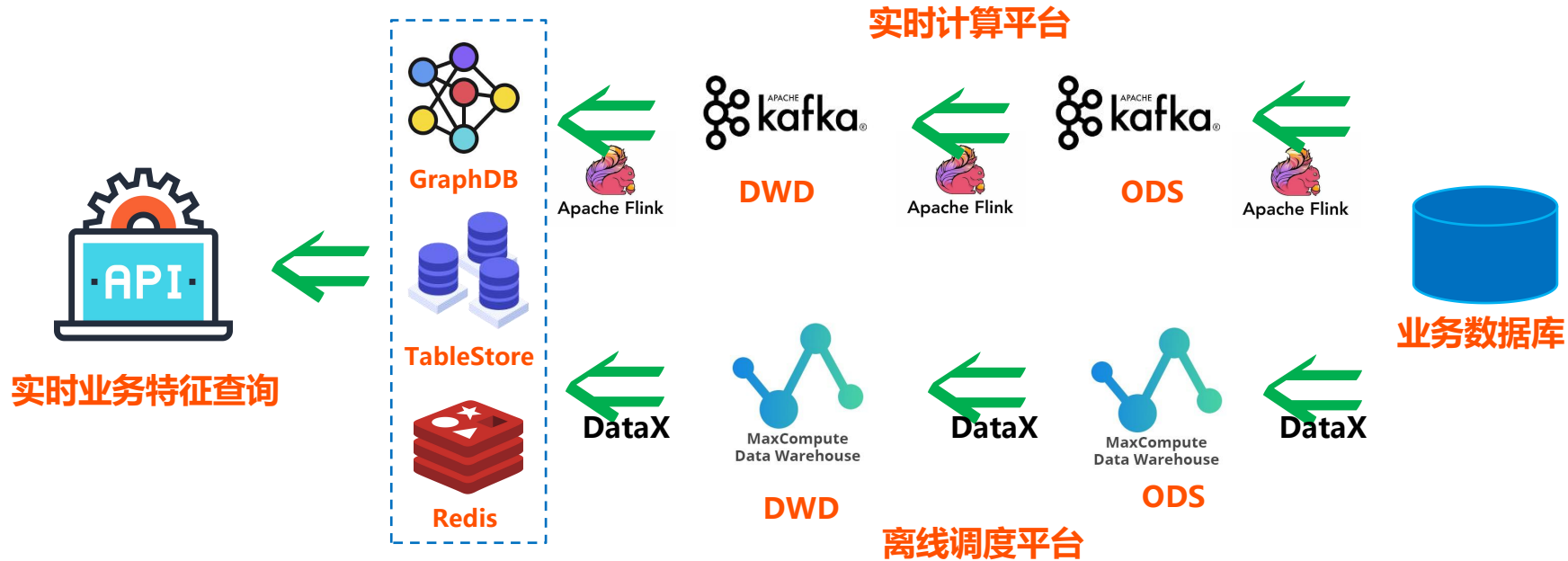
三方数据平台

三方数据实时接入



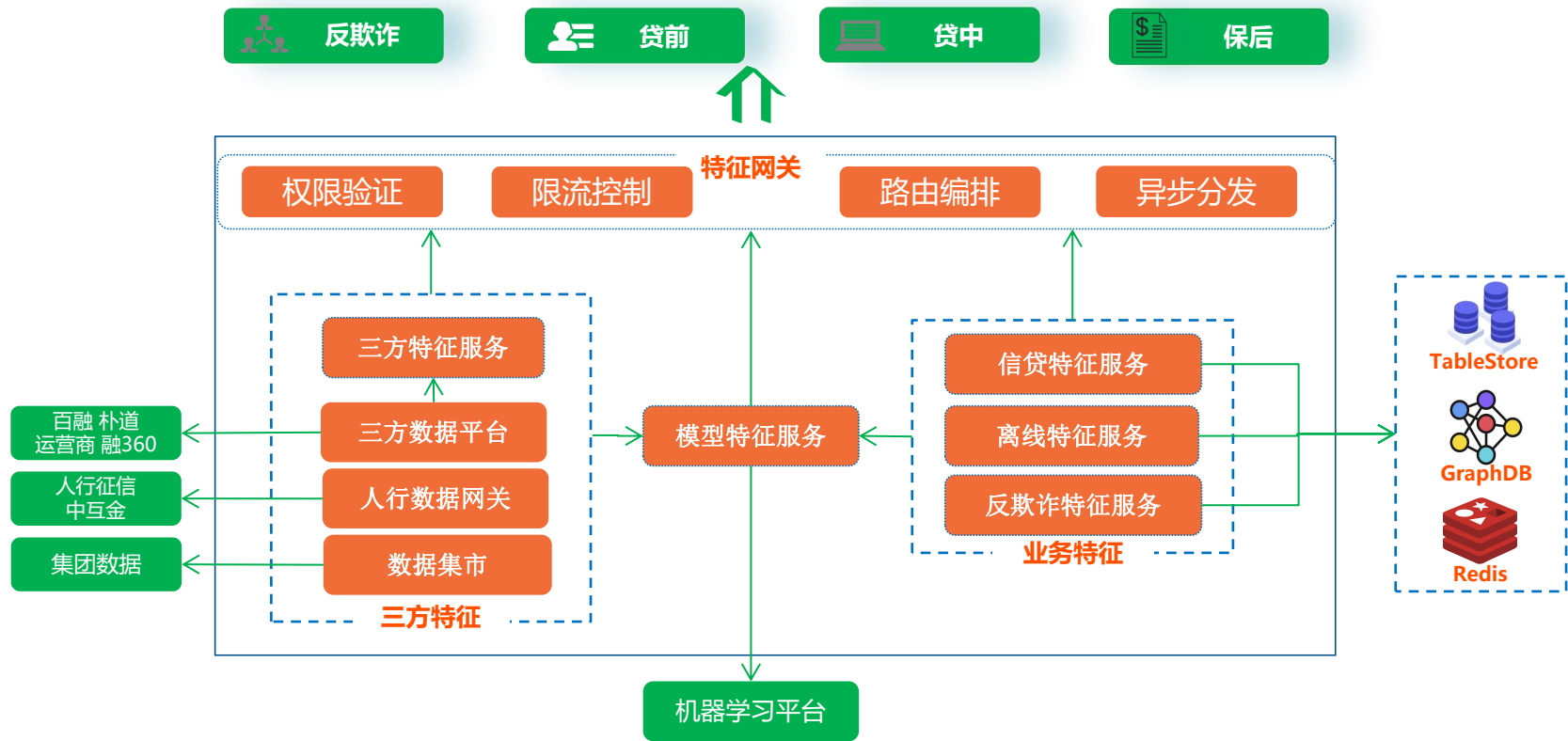
实时接口调用  缓存数据查询  三方数据获取

业务数据实时接入



批量数据处理 + 实时数据同步 + 高速查询引擎 \Rightarrow Lambda架构

实时特征平台的系统架构



03

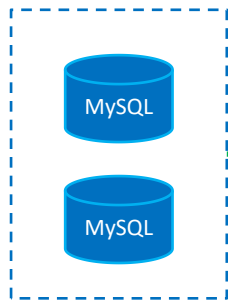
实时业务特征计算详解



特征实时计算方案选型



实时数据核心数据设计



业务数据库



ID-Mapping

src_entity: 原始主键
entity_type: 实体类型
des_entity: 目标主键

业务明细数据表

用户明细表

授信明细表

支用明细表

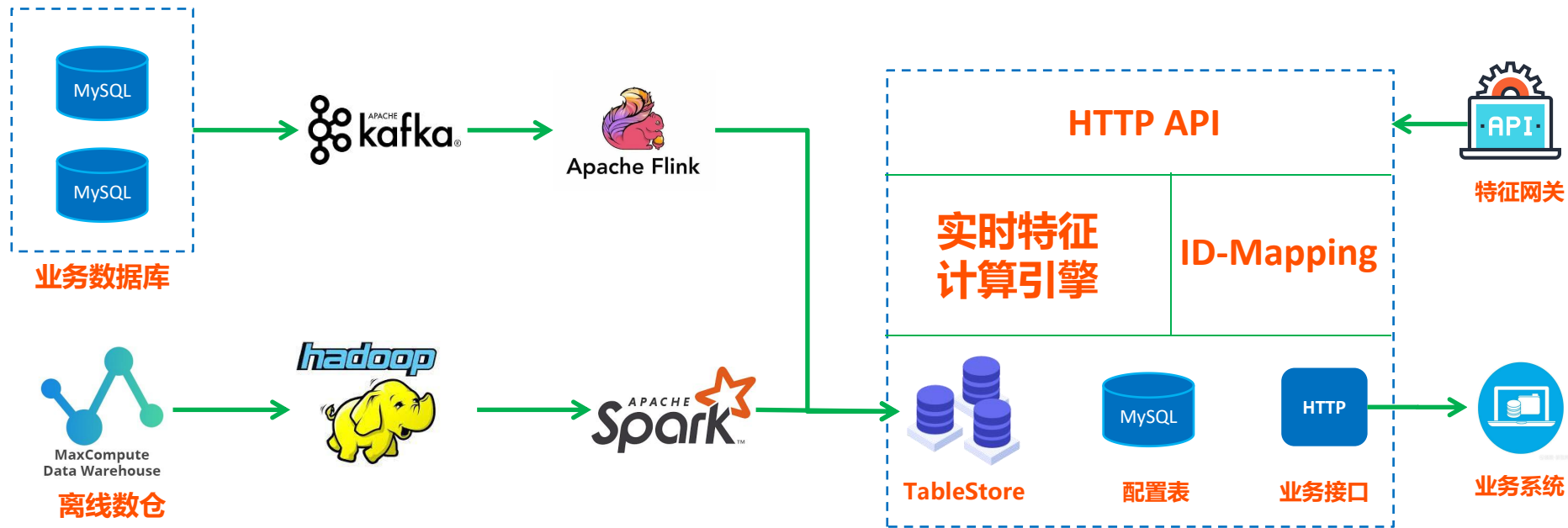
还款明细表

额度明细表

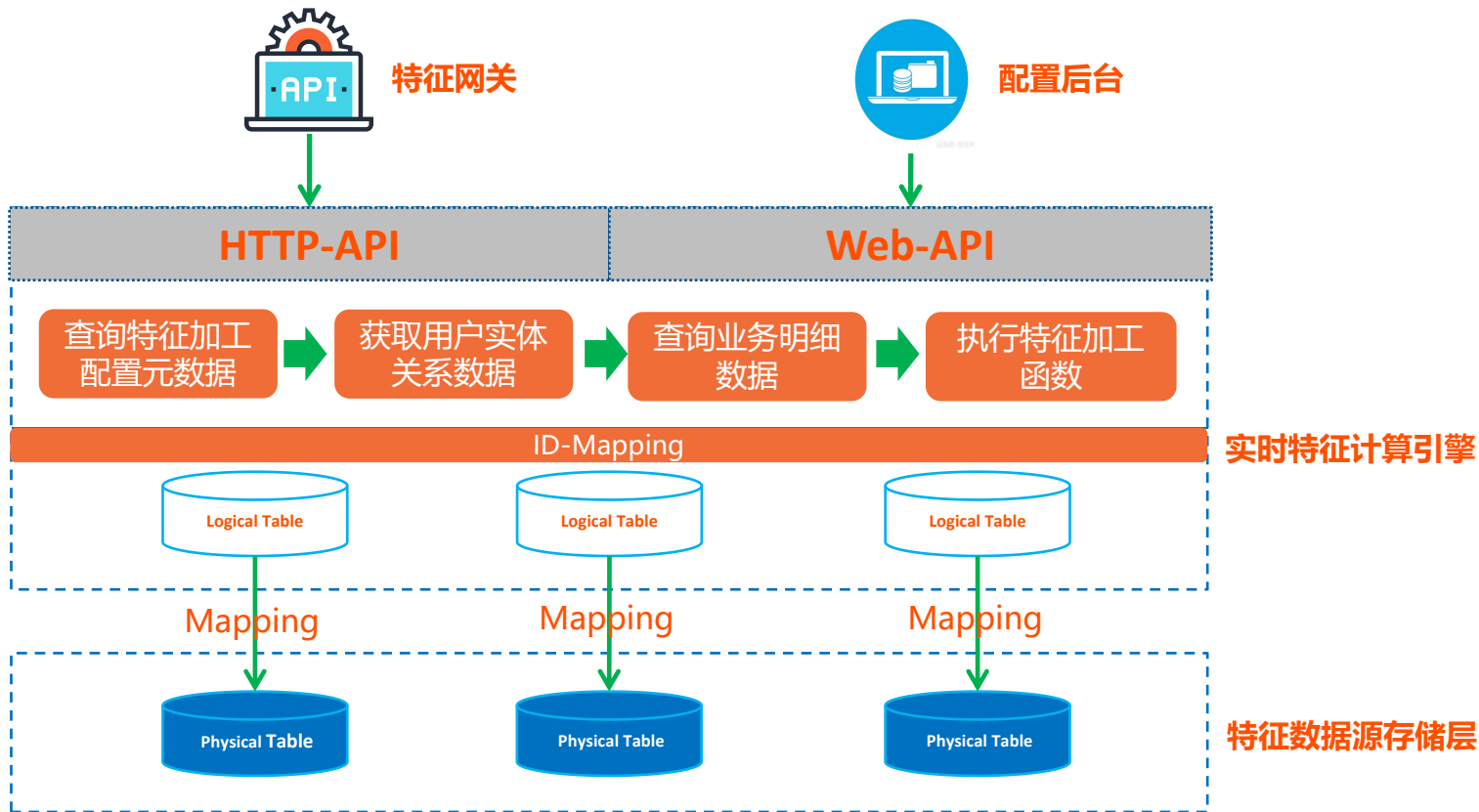
逾期明细表

实时业务特征数据流

实时数据同步 + 离线数据回补 \Rightarrow 全量时序数据采集 特征计算引擎 + ID-Mapping关联表 \Rightarrow 特征配置化生成



实时特征计算引擎



04

反欺诈场景的特征应用



反欺诈特征分类

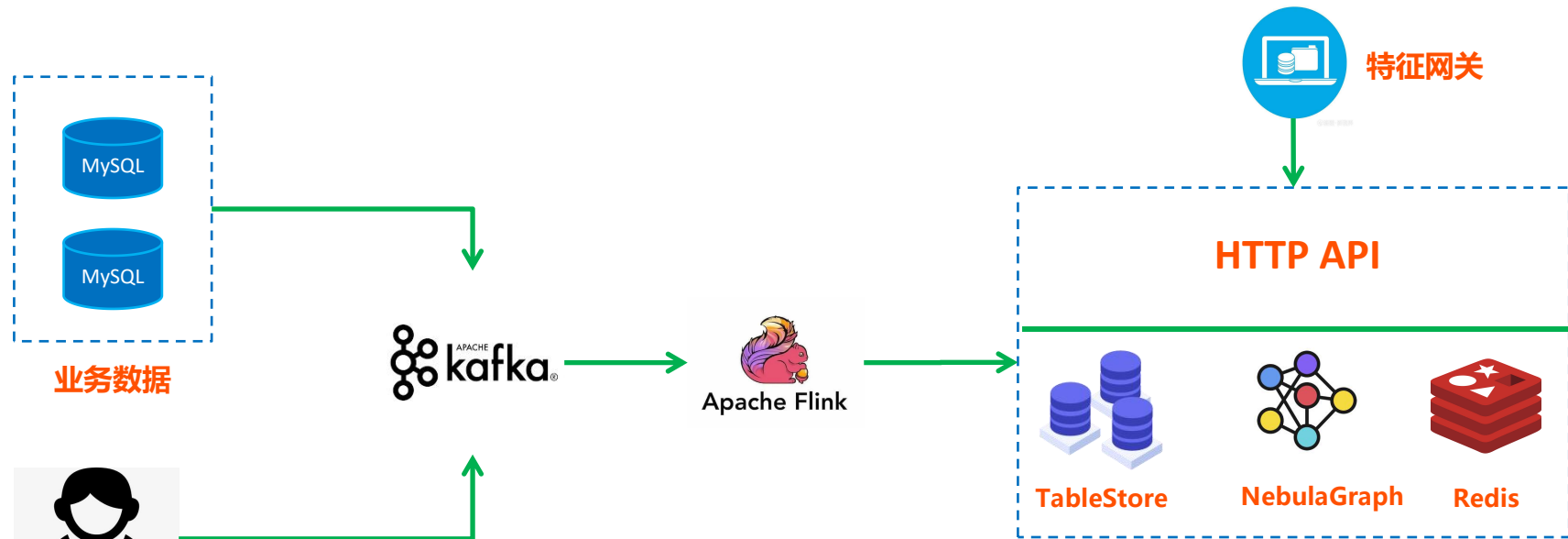


变量英文名	变量中文名	统计口径
certno_network_group_num	身份证_客户所在群组规模	身份证号码关联的关系网络群组的节点数
certno_network_group_self_num	身份证_客户所在群组本产品申请客户数量	身份证号码关联的关系网络群组的已注册客户数
mobile_network_group_num	手机号_客户所在群组规模	手机号码关联的关系网络群组的节点数
mobile_network_group_self_num	手机号_客户所在群组本产品申请客户数量	手机号码关联的关系网络群组的已注册客户数
td_deviceid_self_cnt_180d	近180天内_同一同盾设备指纹关联本产品账号个数	近180天本产品设备记录中入参同盾设备指纹对应的账号(注册手机号)个数
td_deviceid_self_cnt_90d	近90天内_同一同盾设备指纹关联本产品账号个数	近90天本产品设备记录中入参同盾设备指纹对应的账号(注册手机号)个数
td_deviceid_self_cnt_30d	近30天内_同一同盾设备指纹关联本产品账号个数	近30天本产品授信申请过程设备记录中入参同盾设备指纹对应的账号(注册手机号)个数
td_deviceid_self_cnt_3d	近3天内_同一同盾设备指纹关联本产品账号个数	近3天本产品授信申请过程设备记录中入参同盾设备指纹对应的账号(注册手机号)个数

- 1 用户埋点行为类型特征
- 2 GPS+GeoHash位置类型特征
- 3 同一设备关联的用户数业务数等实时数据
- 4 关系图谱邻边关系：节点度数，是否关联黑名单等图谱特征
- 5 关系图谱社群大小，社群特性等社群特征

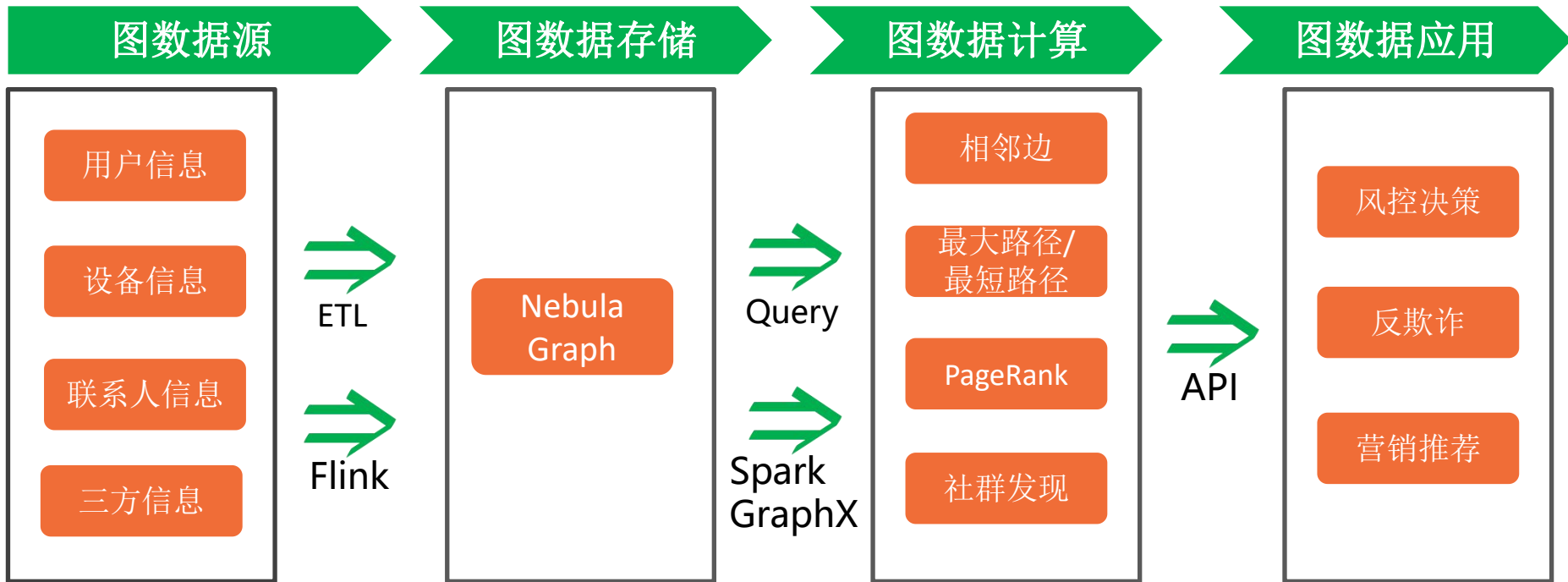
变量中文名	变量英文名	统计口径
自赏_身份证_本产品_近5分钟内同一GPS格子内申请次数	inde_certno_self_same_gps_grid_crd_apply_times_5mins	根据身份证及产品信息，计算本用户授信时点GPS所在1km*1km格子编号，再计算近5分钟内，同一格子编号下，本产品在该格子内的所有授信申请次数（不包含用户自身） 若取不到当前用户GPS信息，赋值-90
自赏_身份证_本产品_近10分钟内同一GPS格子内部申请次数	inde_certno_self_same_gps_grid_crd_apply_times_10mins	根据身份证及产品信息，计算本用户授信时点GPS所在1km*1km格子编号，再计算近10分钟内，同一格子编号下，本产品在该格子内的所有授信申请次数（不包含用户自身） 若取不到当前用户GPS信息，赋值-90
自赏_身份证_本产品_近30分钟内同一GPS格子内部申请次数	inde_certno_self_same_gps_grid_crd_apply_times_30mins	根据身份证及产品信息，计算本用户授信时点GPS所在1km*1km格子编号，再计算近30分钟内，同一格子编号下，本产品在该格子内的所有授信申请次数（不包含用户自身） 若取不到当前用户GPS信息，赋值-90
自赏_身份证_本产品_近60分钟内同一GPS格子内部申请次数	inde_certno_self_same_gps_grid_crd_apply_times_60mins	根据身份证及产品信息，计算本用户授信时点GPS所在1km*1km格子编号，再计算近60分钟内，同一格子编号下，本产品在该格子内的所有授信申请次数（不包含用户自身） 若取不到当前用户GPS信息，赋值-90
自赏_身份证_本产品_近2小时内同一GPS格子内部申请次数	inde_certno_self_same_gps_grid_crd_apply_times_2hrs	根据身份证及产品信息，计算本用户授信时点GPS所在1km*1km格子编号，再计算近2小时内，同一格子编号下，本产品在该格子内的所有授信申请次数（不包含用户自身） 若取不到当前用户GPS信息，赋值-90

实时反欺诈特征数据流

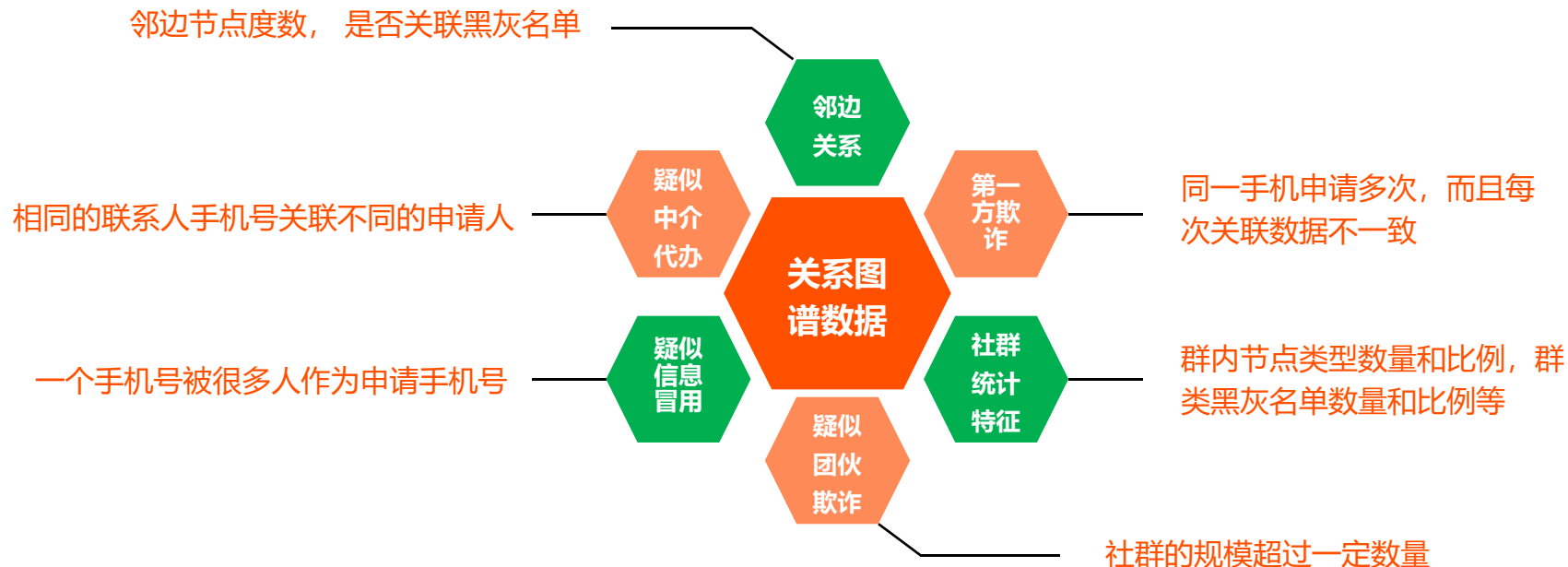


TableStore: 借助Tablestore的高性能查询能力和LBS函数能力实现位置聚集类特征计算
NebulaGraph: 使用图数据库实现用户关系，设备关系等关联实体的数据存储
Redis: 提供实时的用户行为特征数据存储

关系图谱架构图



反欺诈图特征提取



非常感谢您的观看

众安保险 | DataFun.

