

金融反欺诈场景下的Spark实践

王婷 数据科学家 宜人贷

2017-05-19

个人简介

宜信
CreditEase

宜人贷
www.yirendai.com

- 计算机专业PH.D.
- 近5年从事数据挖掘、大规模社交网络分析、社会计算、知识图谱等机器学习算法实践工作
- 现任宜人贷数据科学家，从事反欺诈建模和创新技术自动化风控系统，已成功申请2项反欺诈技术专利



- ① 金融科技企业面临的欺诈风险
- ② 在线反欺诈中的Spark算法实践
- ③ 基于Spark架构的实时反欺诈平台

- ① 金融科技企业面临的欺诈风险
- ② 在线反欺诈中的Spark算法实践
- ③ 基于Spark架构的实时反欺诈平台

- 金融的本质：资源的最合理化应用
- 互联网技术：交易的边界成本趋向“零”
- 金融科技（FinTech）：通过技术手段推动金融创新，形成对金融市场、机构及金融服务产生重大影响的业务模式、技术应用以及流程和产品



Volume

每天生成
T级数据



Velocity

最高每分钟
50+申请



Variety

网络, 设备, 行为,
渠道, PII, 社交,
三方, 等类别



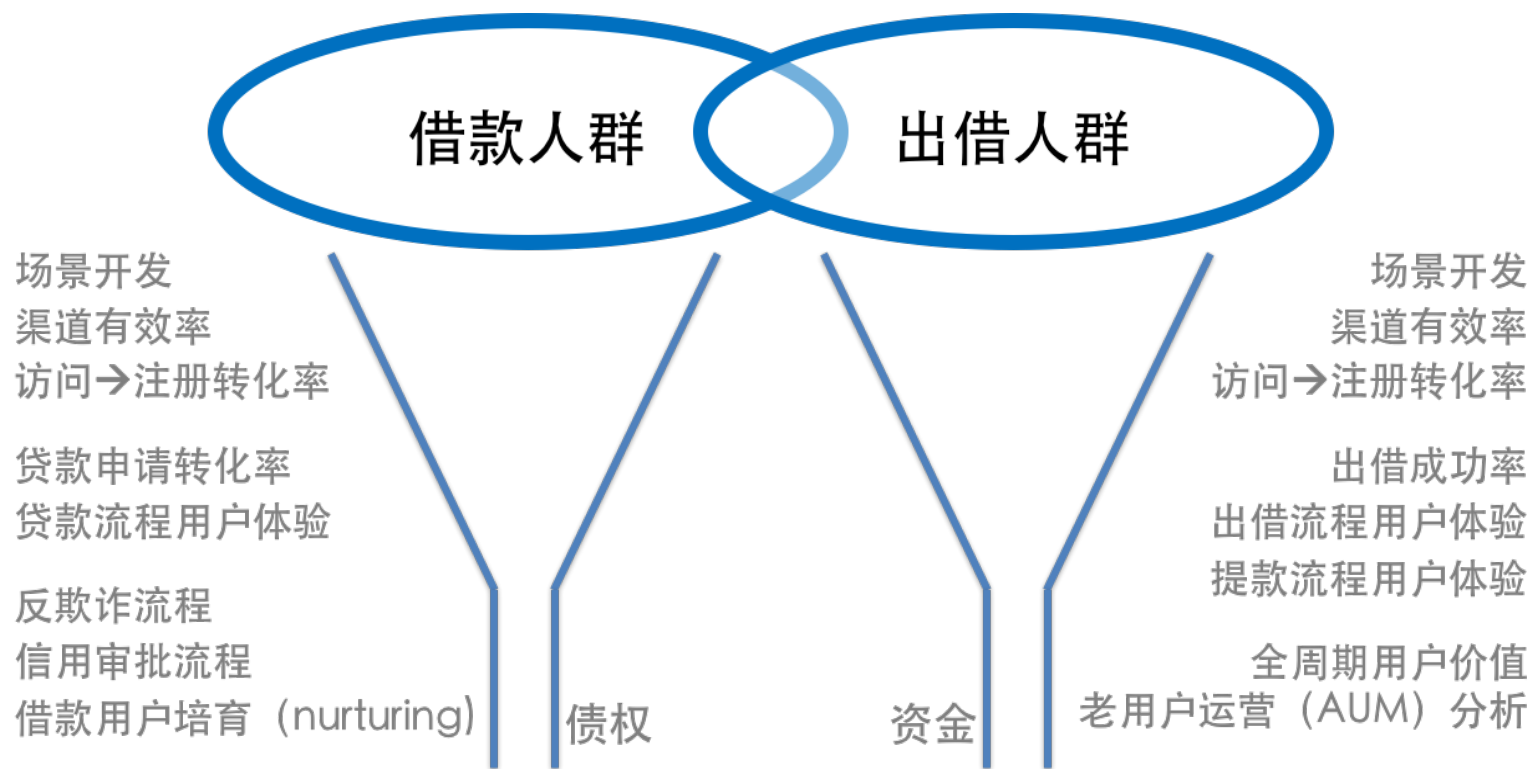
Veracity

完整度和质量
经常残差不齐

金融科技-个人对个人的信用贷款

宜信
CreditEase

宜人贷
www.yirendai.com



宜人贷：P2P 借款与理财咨询服务平台



宜人贷借款APP



宜人理财APP



宜人贷官方网站

宜人贷：极速信任-自动化信用评估

宜信
CreditEase

宜人贷 
www.yirendai.com

全流程线上借款与理财咨询服务

客户获取

信用评估

交易促成

客户服务

线上

线下



场景不同

人群不同

数据获取方式不同

数据维度不同

数据深度不同

信用评估机制不同

欺诈风险

是互联网金融
线上信贷工厂模式
最大的挑战

金融科技企业面临的欺诈风险

风险	遇到的问题	业界通常解决方法	业界的方法为什么无效
信用风险	还款能力	收集收入水平、消费水平、负债情况等对用户进行风险评分	无权威数据、数据收集难度大、传统评分卡有效特征挖掘难度大
欺诈风险	伪冒申请和欺诈交易	人工审查、信用黑名单、基于规则	人工效率低、无权威黑名单、无法自动发现异常、欺诈手段更新快

人群团体化



地区集中化



方式多样化



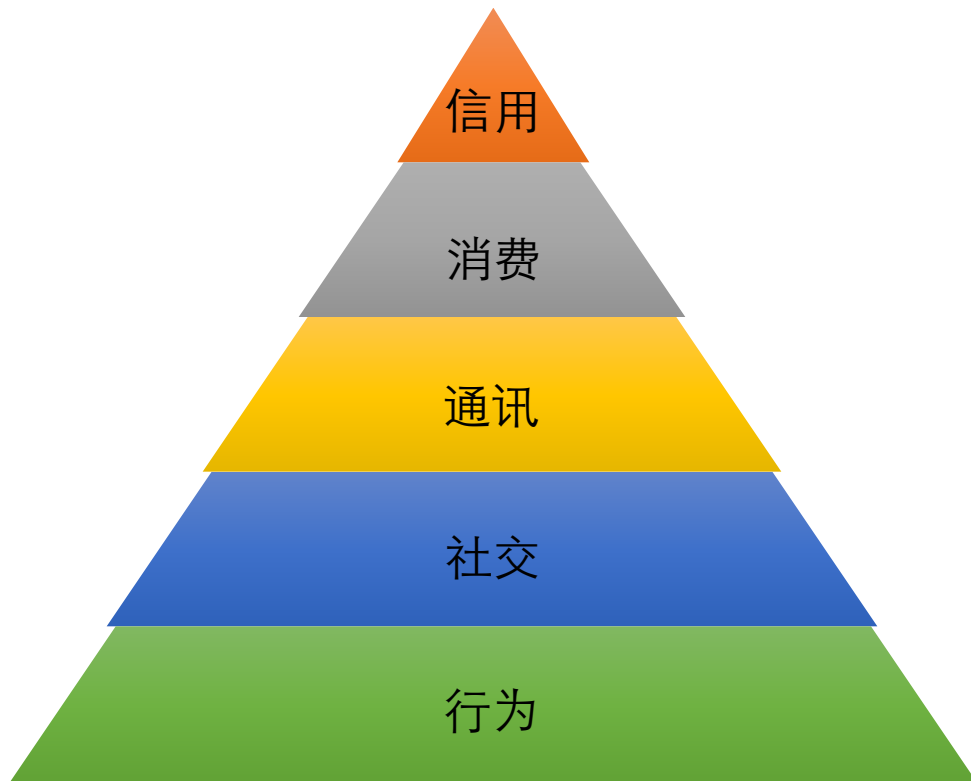
工具智能化



- ① 金融科技企业面临的欺诈风险
- ② 在线反欺诈中的Spark算法实践
- ③ 基于Spark架构的实时反欺诈平台

反欺诈也是一种机器学习过程

- Y目标: Benchmark选取
 - 好、坏用户定义
 - 训练、测试和跨时间验证样本
- X变量: 特征工程
 - 人工特征工程
 - 图谱特征挖掘技术
 - ✓ 知识图谱技术
 - ✓ 图挖掘技术



风险控制数据金字塔

构建金融知识图谱：FinGraph

宜信
CreditEase

宜人贷
www.yirendai.com

FinGraph 平台系统

■ 10种实体

- 电话、身份证、银行卡、信用卡、IP、设备号、地理位置等

■ 约2.6亿节点

■ 约10亿边关系

应用场景层面

智能搜索、反欺诈、贷后管理、营销分析、运营支撑 等

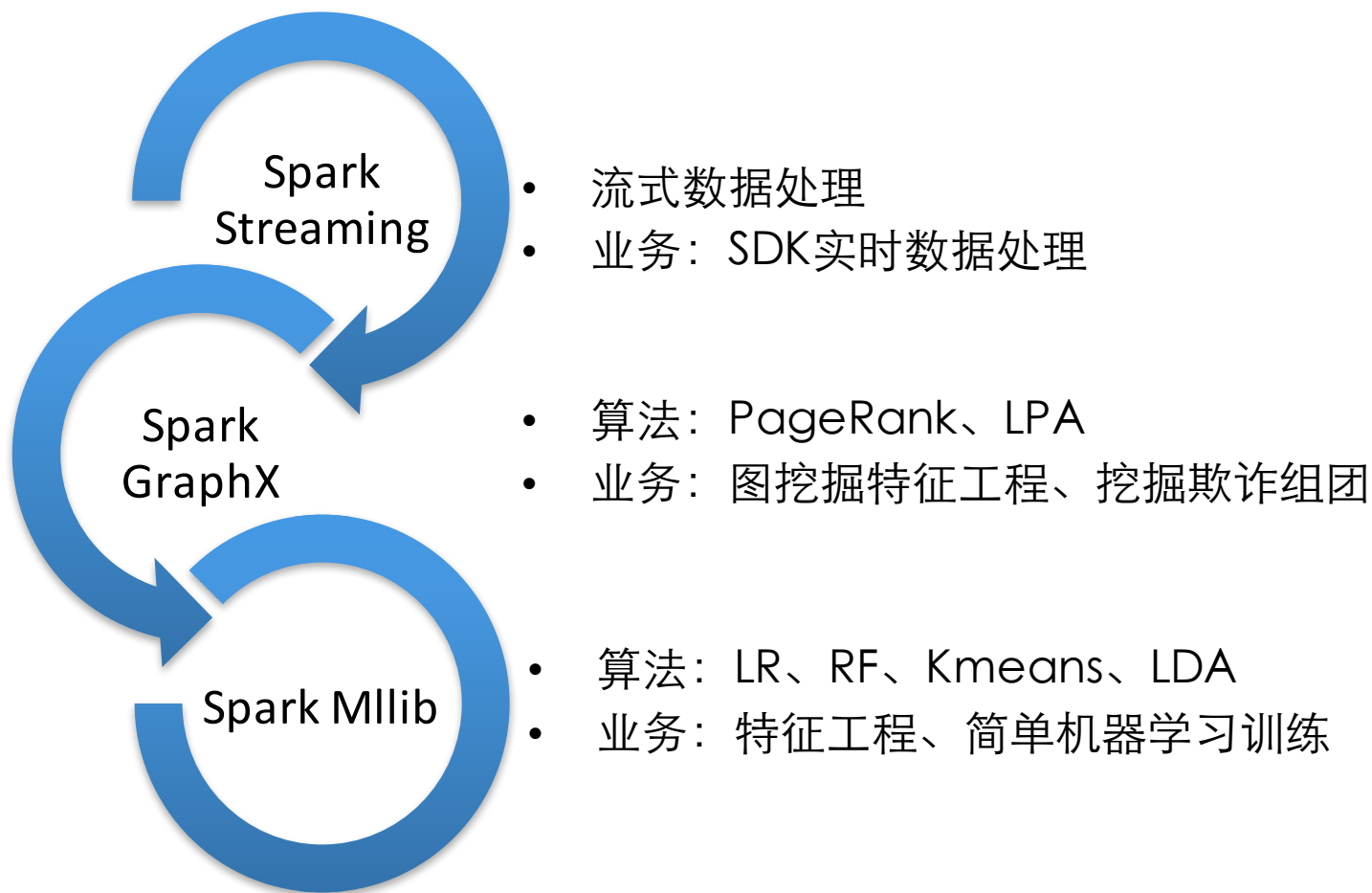
系统支持层面

特征工程、模型开发、异常监控、推荐系统 等
Spark+Hadoop+GraphX+Mllib+Streaming+TensorFlow

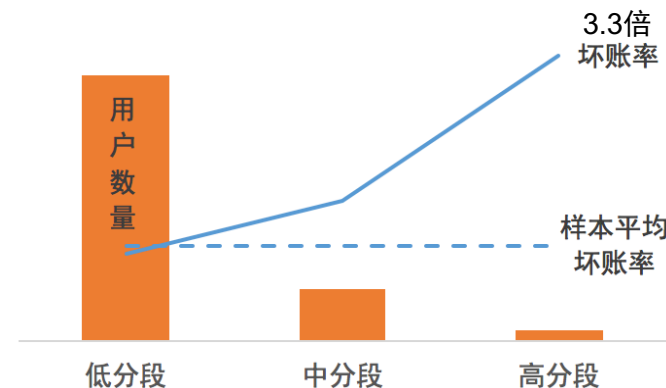
数据整合层面

信用数据、金融消费数据、行为数据、社交数据、
网络安全、第三方数据 等
图数据库neo4j

反欺诈场景下Spark三板斧



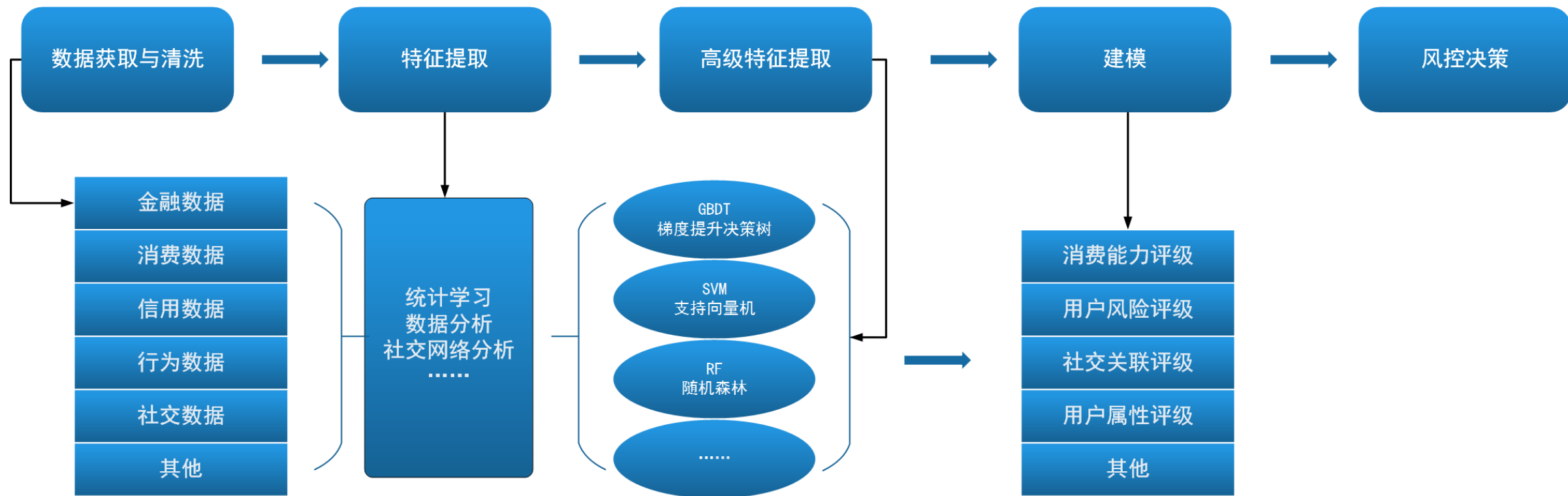
- 反欺诈分析案例：借款用户通信社交网络与欺诈风险
 - 结论：PageRank高分段用户的坏账率是低分段用户的3.3倍



反欺诈建模中的数据科学

宜信
CreditEase

宜人贷
www.yirendai.com

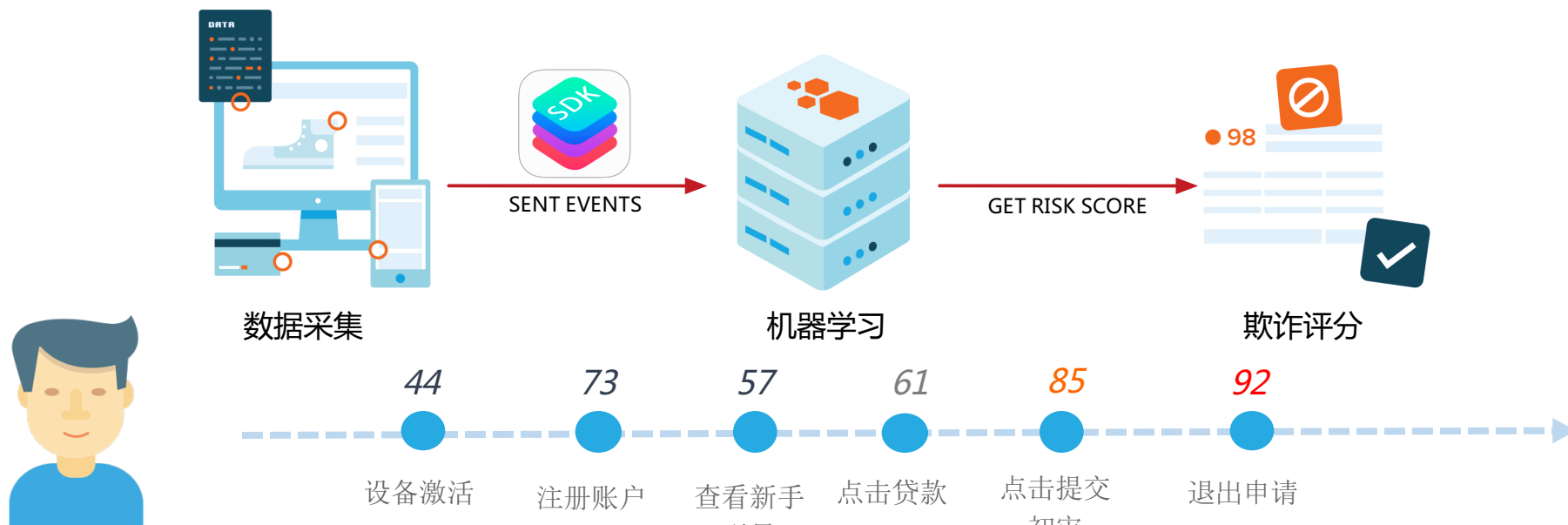


- ① 金融科技企业面临的欺诈风险
- ② 在线反欺诈中的Spark算法实践
- ③ 基于Spark架构的实时反欺诈平台

对不同事件得出实时欺诈评分

宜信
CreditEase

宜人贷
www.yirendai.com



PHONE:18612586949
NAME: Mike
ADDRESS: Chaoyang, Beijing
IP: 123.89.21.10
IMEI : 447769804451095
Mac: 00-80-C2-00-00-1A
OS : android 4.3
Model: Oppo R7

.....

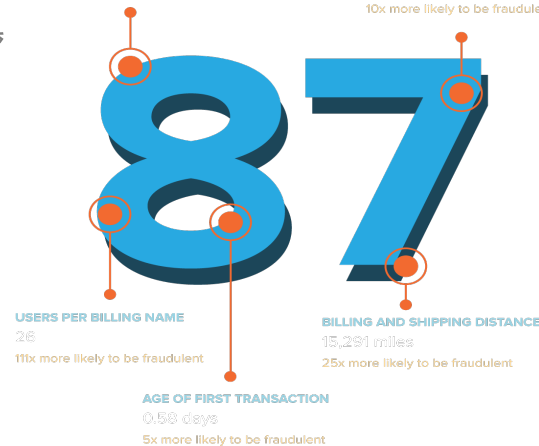
安装贷款类app过多

联系人一度触灰

初审速度过于频繁
银行卡交易流水异常

DEVICE FINGERPRINT
v2vu4la15nm12cc6nf0bs3v90
52x more likely to be fraudulent

IP ADDRESS
190.128.105.159
10x more likely to be fraudulent



通过SDK采集欺诈事件

设备数据

- ✓ 手机品牌
- ✓ 手机型号
- ✓ 操作系统
- ✓ 本机号码
- ✓ 设备ID
- ✓ App安装列表

行为数据

- ✓ 账号登录
- ✓ 页面进入
- ✓ 按钮点击
- ✓ 信息输入
- ✓ 广告浏览
- ✓ 操作时间

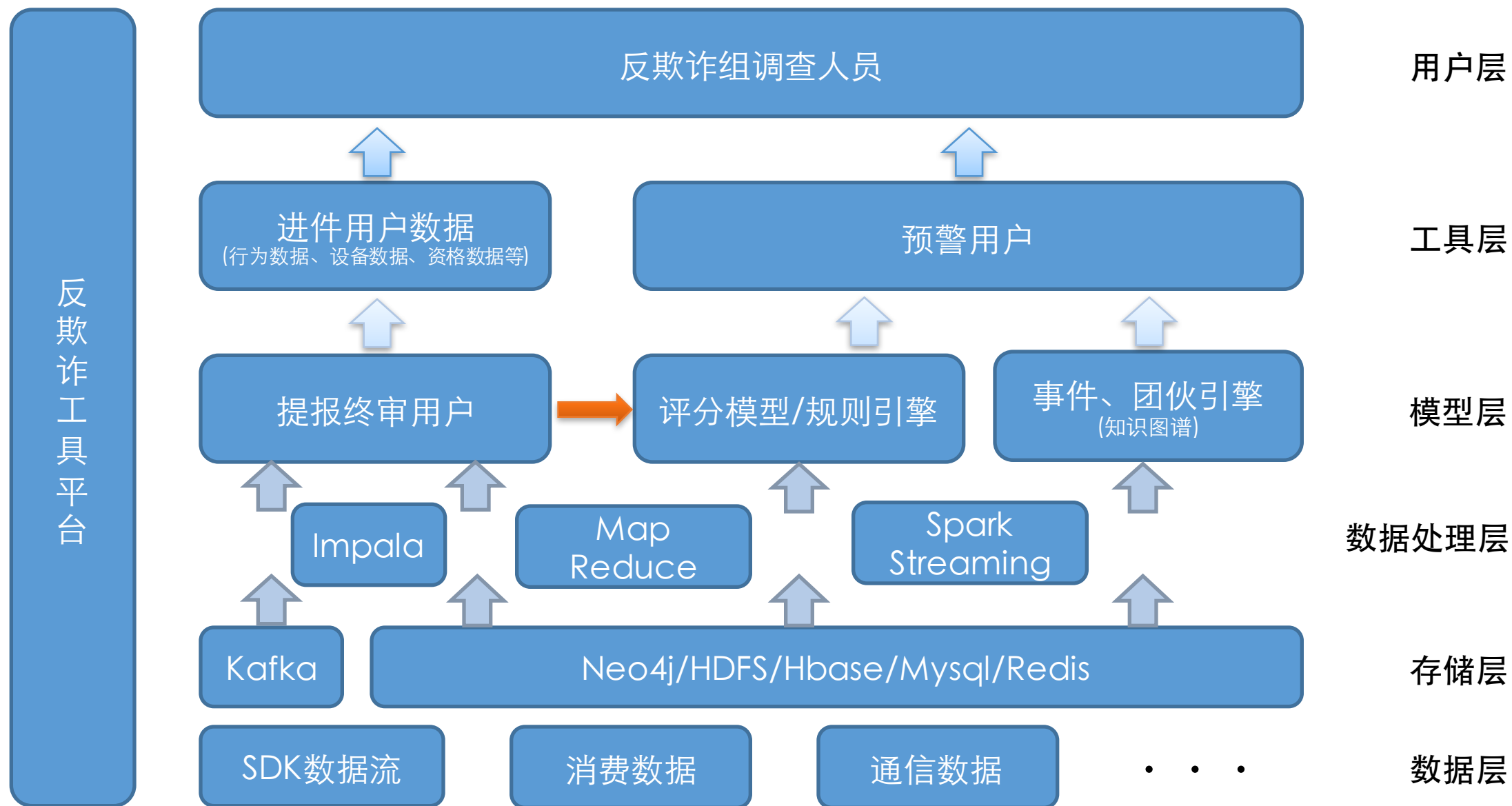
位置数据

- ✓ GPS
- ✓ IP

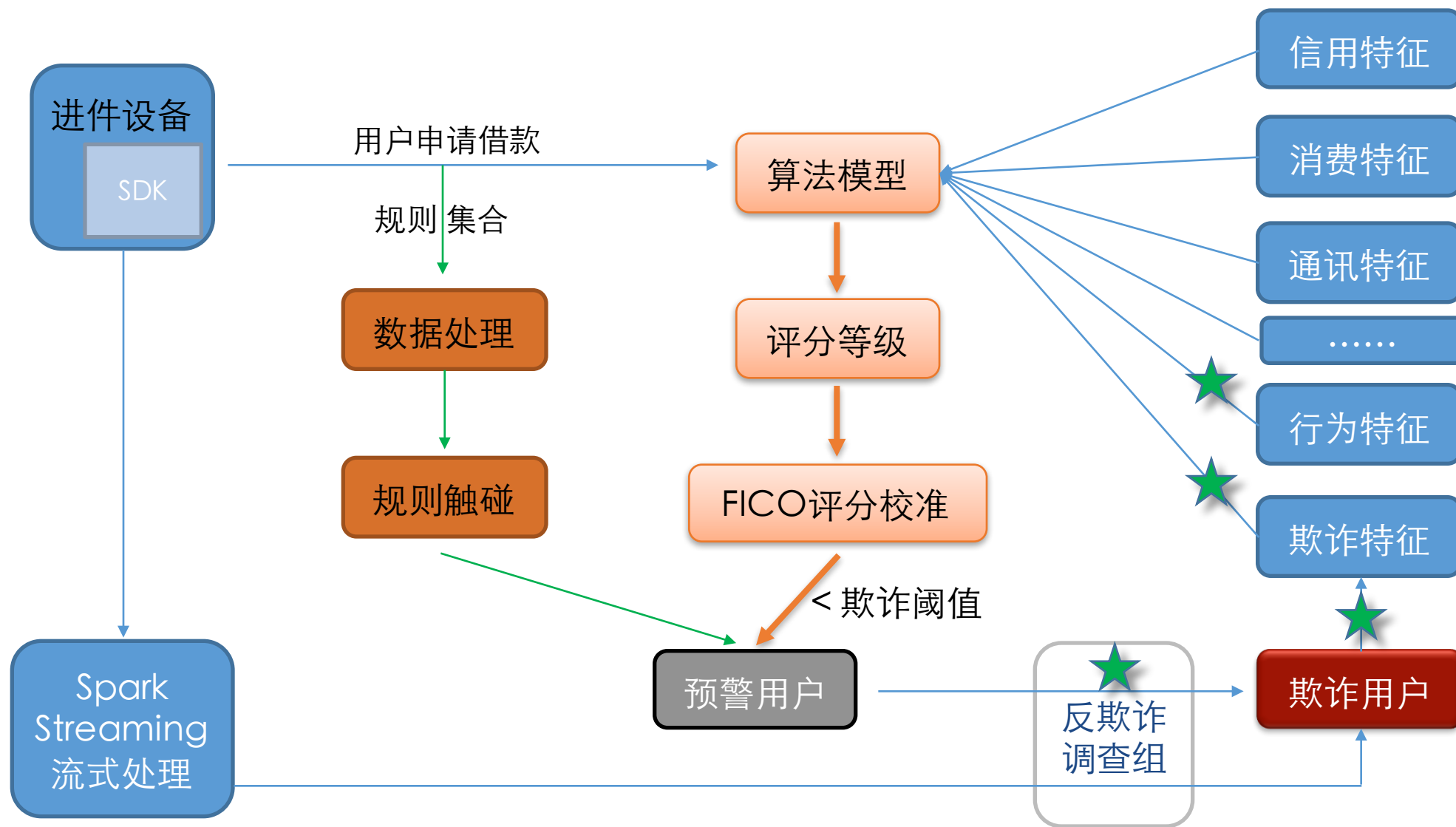
反欺诈平台架构

宜信
CreditEase

宜人贷
www.yirendai.com

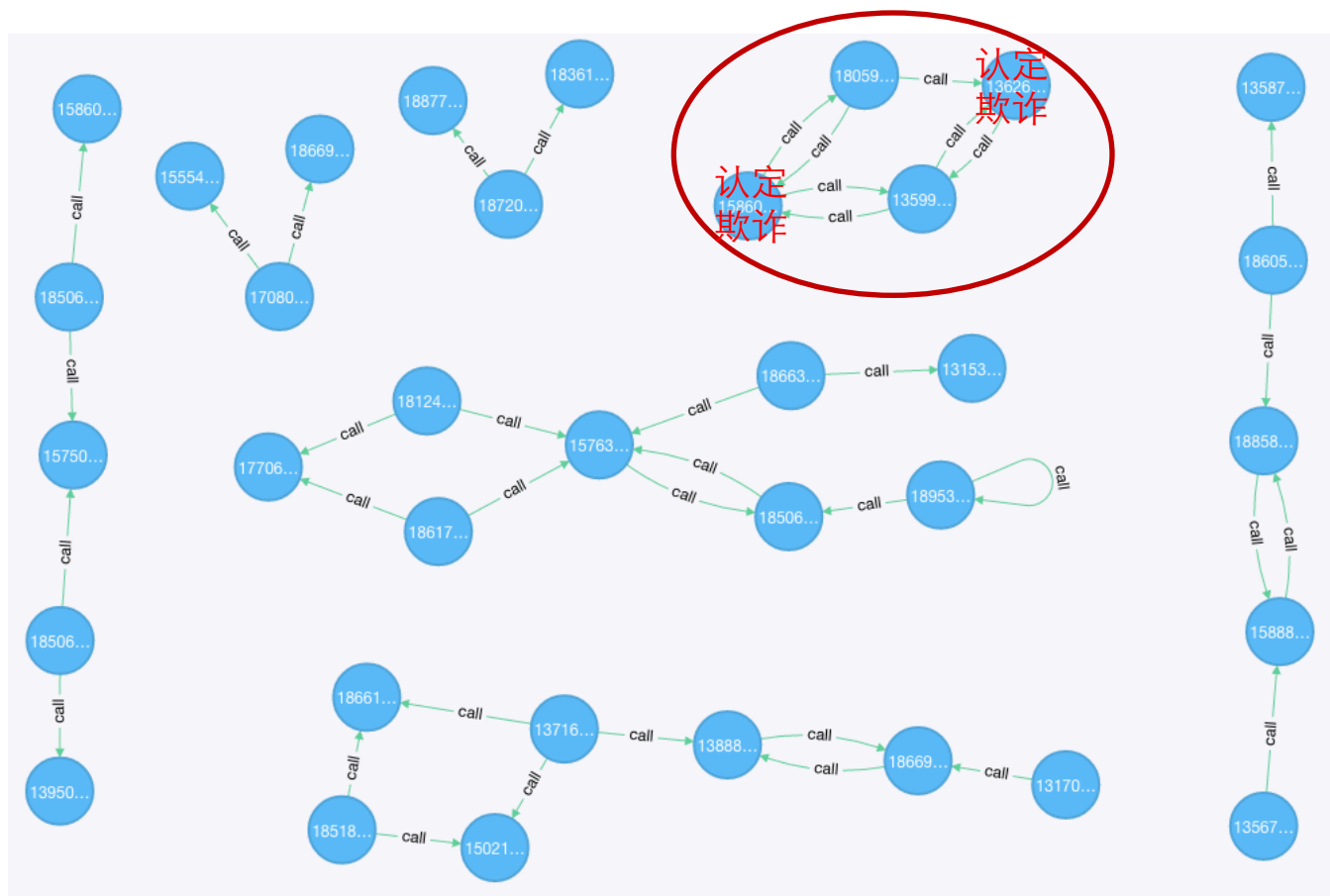


反欺诈平台工作流程

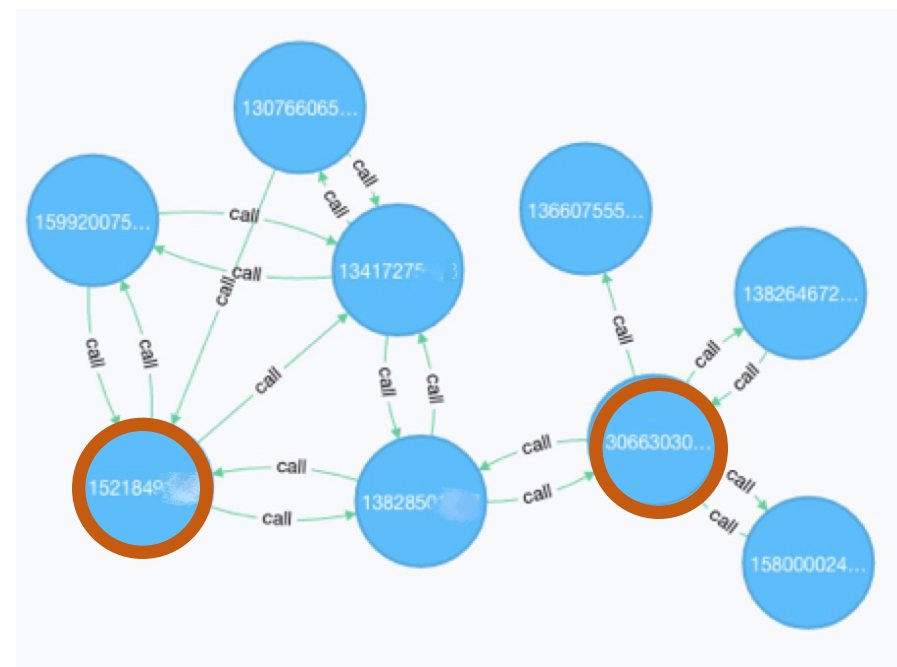


引入反欺诈调查员提升反馈效率

- 利于实时新增欺诈标注，快速反馈回模型训练（6个月->1天）
- 结合图谱挖掘，利于快速发现欺诈团伙



- 根据调查认定的欺诈用户，寻找联系紧密的用户团体



用一手行为数据和图谱信息创造商业价值

宜信
CreditEase

宜人贷
www.yirendai.com

挑战：

初步历史行为数据分析体现了显著的欺诈区分能力。怎样实时捕捉，上传，处理，和分析行为数据？

解决方案：

- 一行代码 埋点SDK
- 自动实时/准实时上传用户行为
- Flume+Kafka实时处理分析

挑战：

申请行为的数据量大，维度多，实时性要求高。怎样储存，关联，挖掘，查询数据中的欺诈倾向？

解决方案：

- Spark Streaming 流式处理
- HBase KV 查询输出
- Neo4j 集群 关联、存储、挖掘

挑战：

反欺诈调研时效性差，需要实时自动提报疑似欺诈案例，及时发现欺诈事件/团伙，来主动拦截？

解决方案：

- Go做为高效开发和运行基础
- Python连接自动提报后台
- SKLearn、GBDT、事件识别
- Cypher图谱关系挖掘



Thank You !