

基于大数据技术重构数据仓库应用 的探索和实践

赵宏伟

恒丰银行



QCon

全球软件开发大会

成为软件技术专家 的必经之路

[北京站] 2018

2018年4月20-22日 北京·国际会议中心

7折

购票中, 每张立减2040元

团购享受更多优惠



识别二维码了解更多



极客时间

重拾极客精神·提升技术认知

下载极客时间App

获取有声IT新闻、技术产品专栏，每日更新



扫一扫下载极客时间App

AiCon

全球人工智能与机器学习技术大会

助力人工智能落地

2018.1.13 - 1.14 北京国际会议中心



扫描关注大会官网

SPEAKER INTRODUCE



赵宏伟

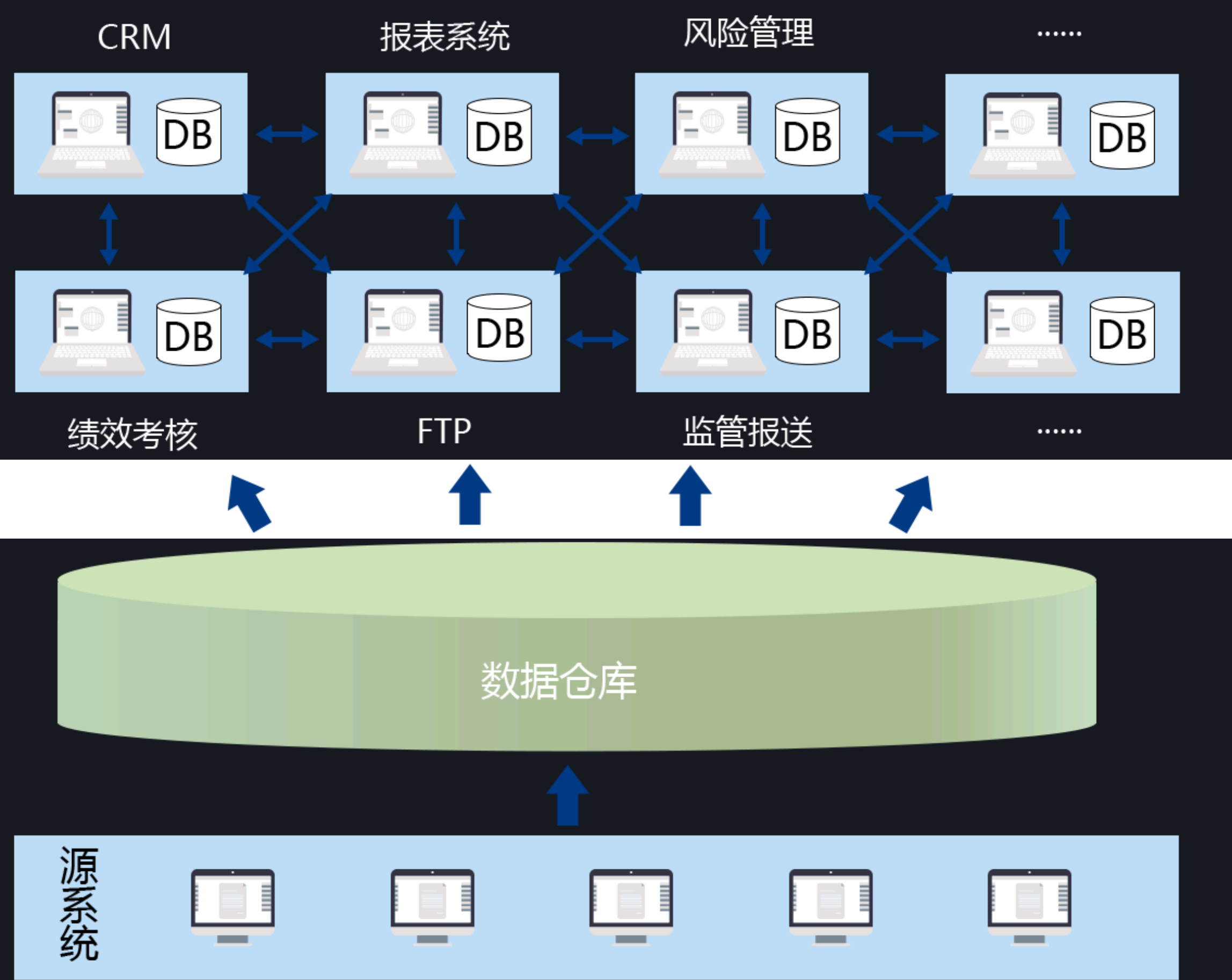
恒丰银行科技开发部高级项目经理

拥有十年系统建设经验，参与设计、实施过众多数据仓库、数据集市、数据分析类、营销管理类系统的建设工作，拥有丰富的数据建模、产品设计、系统架构设计和实施经验。现阶段负责恒丰银行大数据平台、营销管理条线系统的建设工作。

目录

- 1 平台建设背景
- 2 数仓应用体系建设
- 3 风控领域创新应用
- 4 取得成果与未来展望

烟囱式应用现状存在的问题



01 应用野蛮生长
数仓地位尴尬

02 跨应用数据共享困难，大量的数据冗余

04 硬件资源分散
峰值服务能力和大数据量处理能力受限

03 数据治理目标难实现：企业数据模型、数据标准统计口径

大数据技术助力构建大规模数据处理平台

高性价比

1

- 1.相同计算与容错能力，基础环境软硬件成本只需原来的1/3-1/5，大幅降低项目预算
- 2.分布式并行计算技术解决传统数据库架构海量数据的加工能力难题

弹性伸缩

2

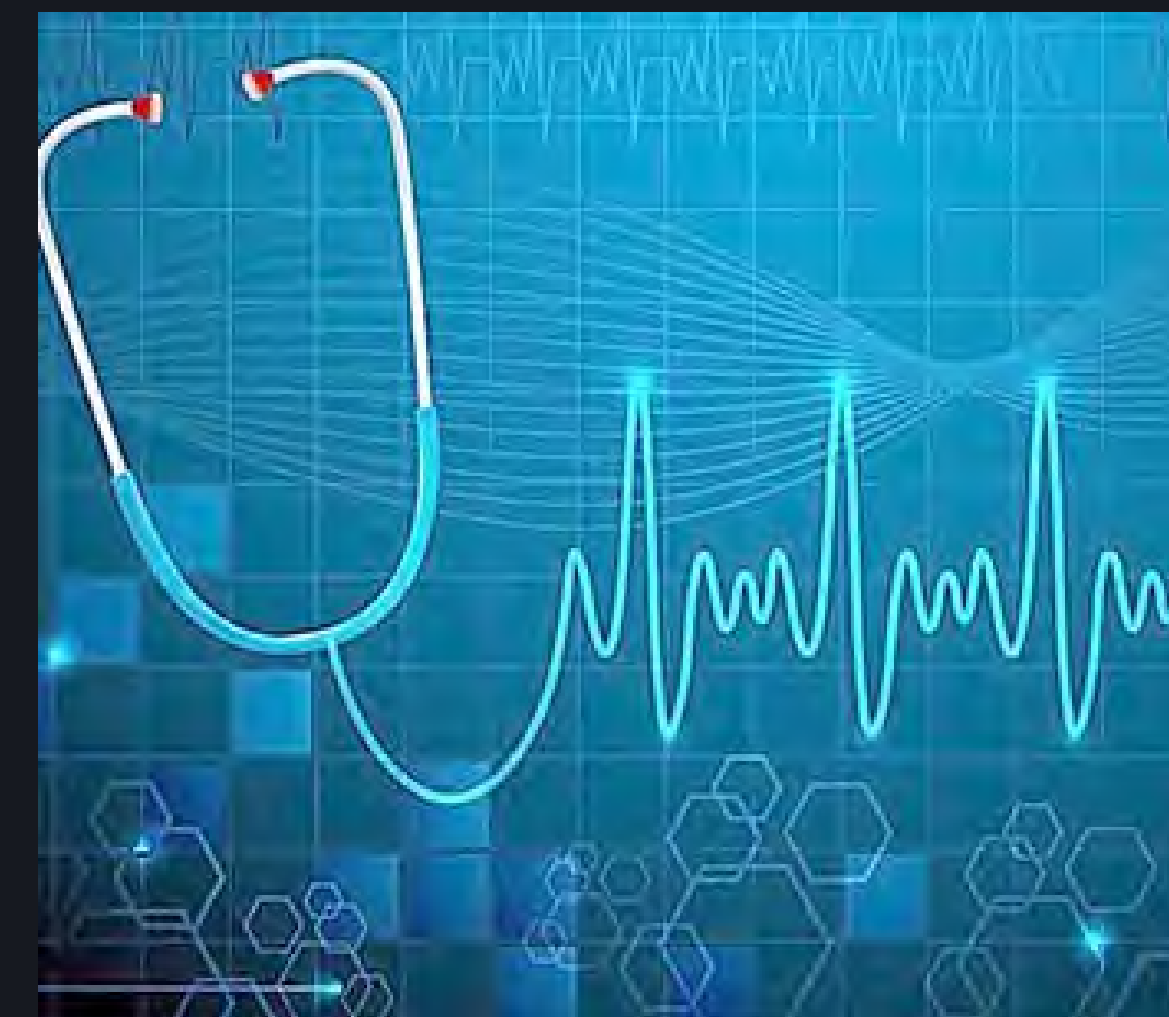
- 1.构建大规模计算与存储资源池，用一个平台承载以数据仓库为核心的大部分数据应用；SSD介质加速随机读写速度
- 2.在线横向扩容，实现资源弹性分配与隔离，快速满足应用需求变化

开放便利

3

- 1.开源技术促进平台快速升级迭代，提升核心技术自主掌控能力
- 2.整合存储、加工、实时流处理、机器学习等多样化能力，降低集成难度

基础数据平台技术选型



性能优异

1. 兼顾大数据批量处理和小样本数据精确查询统计的性能需求
2. 优化的数据存储与访问技术，支持索引、分区、行列混合存储
3. 尽量避免GC引发的性能抖动，避免大数据量广播
4. 计算资源有效管控

容易开发

1. 支持SQL2003标准和存储过程，原有应用迁移成本低
2. 编程接口与开源主流兼容，支持ODBC/JDBC标准接口

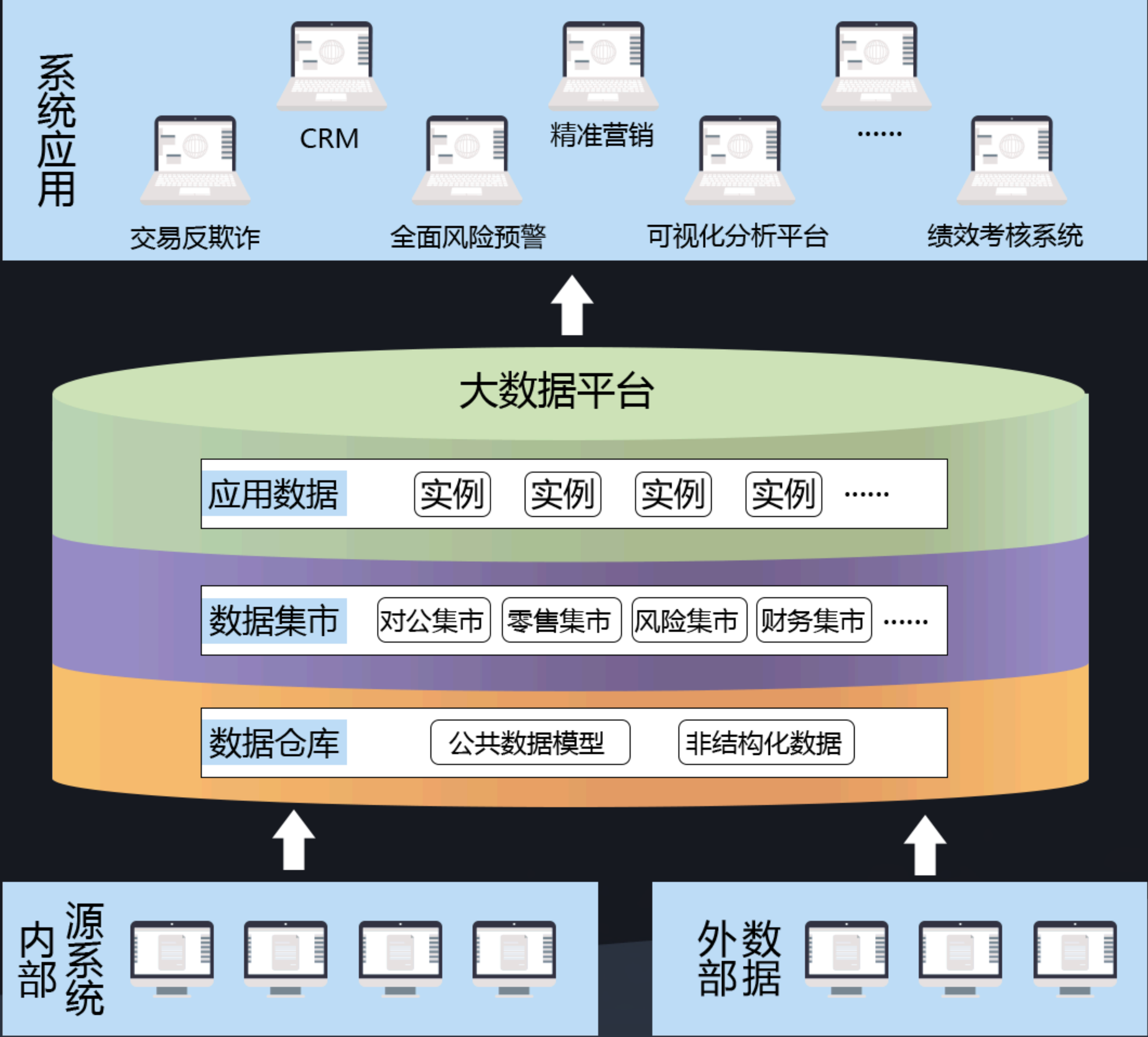
运维简单

1. 高度容错，无单点故障
2. 完善的运维监控管理，开放的监控输出接口
3. 支持在线弹性扩容

支持到位

1. 厂商掌握核心技术
2. 有经验的运维支持团队
3. 响应及时的产品开放团队

构建统一的数据管理平台释放软件开发生产力



构建统一的数据管理平台释放软件开发生产力

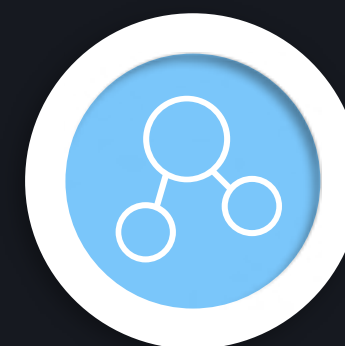


拓展数据仓库新能力



业务数据开放能力

开放高时效性的业务主题应用集市,提供自主数据探索与业务建模的可视化工具



多样化数据整合能力

内外部结构化与非结构化数据的整合加工与共享, 构建更全面的客户信息视图



准实时数据应用能力

全渠道数据实时整合, 实现客户营销、风险管理、业务分析等领域的实时智能应用

技术平台规划

- 1. 海量数据高效采集、存储、加工
- 2. 数据标准化治理、数据生命期管理
- 3. 多租户数据服务资源管理

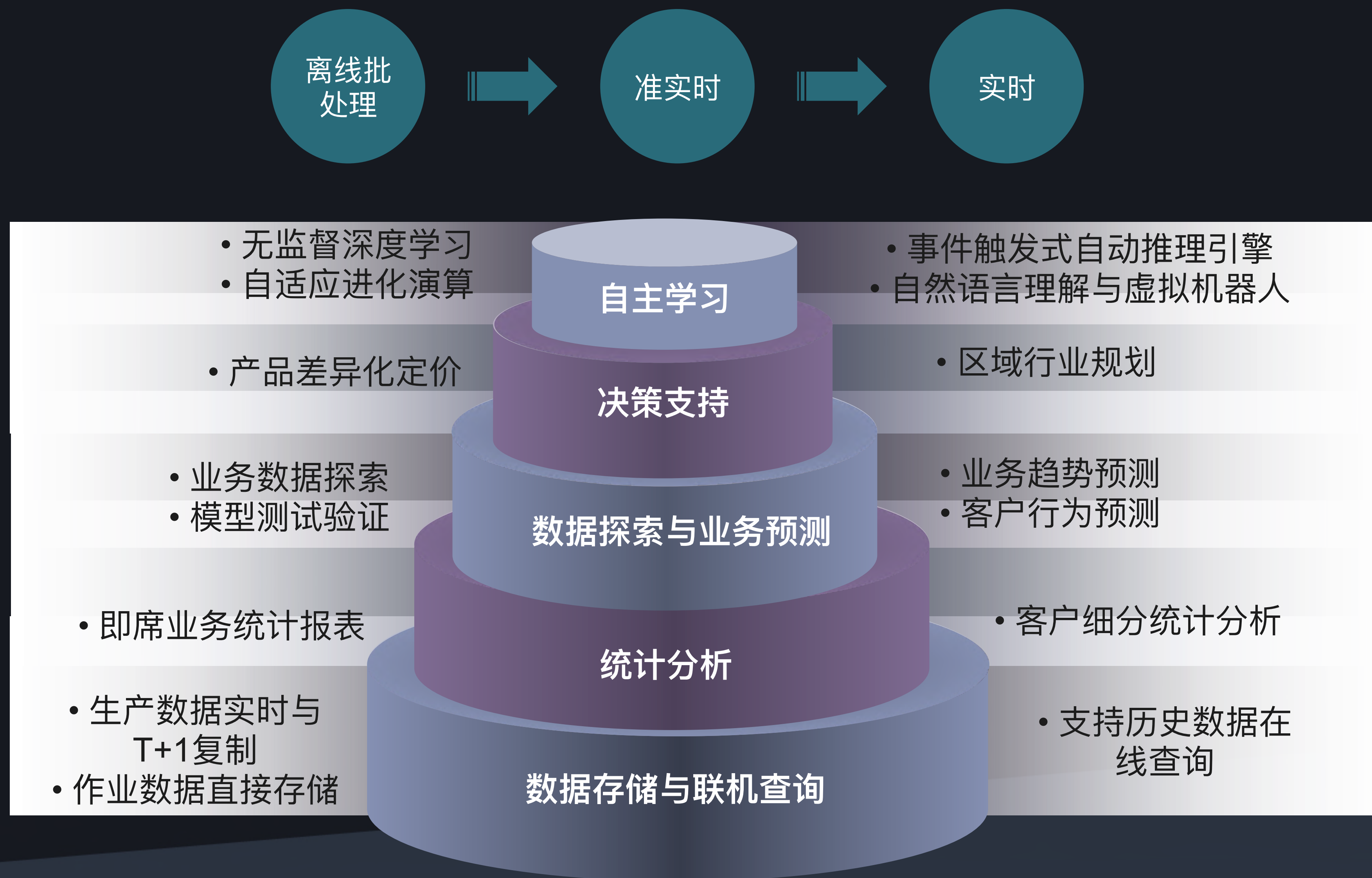


- 1. 高并发低延迟的微服务架构
- 2. 大数据技术集成

- 1. 业务数据可视化
- 2. 交互式数据分析
- 3. 图形化的挖掘建模工具

- 1. CPU/GPU混合并行计算架构
- 2. 并行挖掘算法和深度学习框架
- 3. 并行处理语言、实时流与图计算

构建企业级数据应用能力



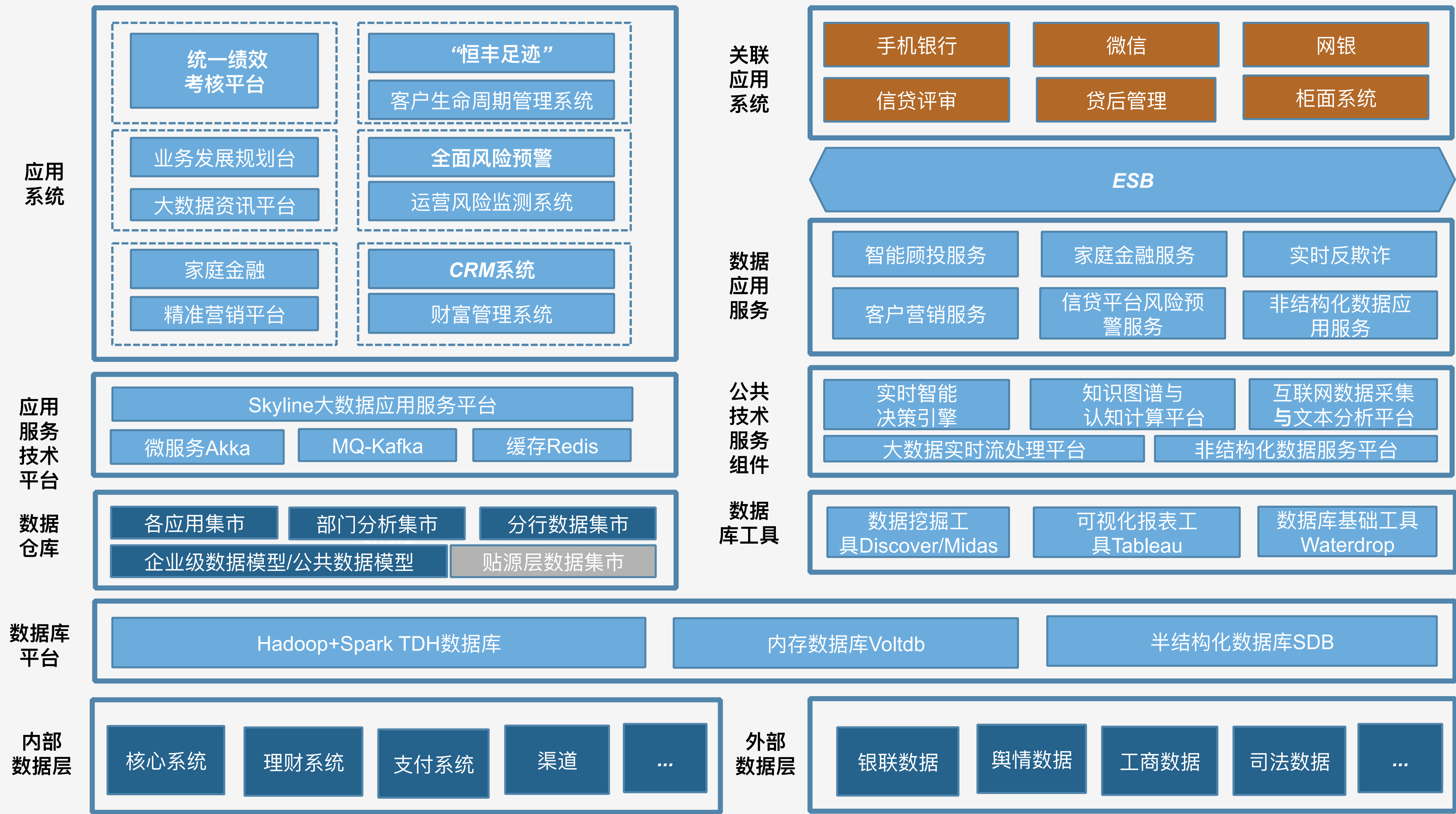
目录

- 1 平台建设背景
- 2 数仓应用体系建设
- 3 风控领域创新应用
- 4 取得成果与未来展望

数仓基础应用架构



数据应用总体技术架构



公共数据模型层设计

采用维度建模为主

- 面向主题
- 覆盖银行分析决策的各个方面
- 满足维度模型的高效性和易理解性

数据的整合性

- Single Source, Single View
- 数据共享平台
- 为各种分析应用提供单一的、整合的数据来源
- 统一的数据定义标准和编码规则

数据的准确性

- 数据具有可逆性，能够真实地反映原始数据的面貌
- 数据具有可回溯性，能够准确地反映历史数据情况

数据的完整性

- 能涵盖银行现有的业务范畴以及数据范围
- 重要实体、重要关系、重要分析维度属性均保持完整

模型设计原则

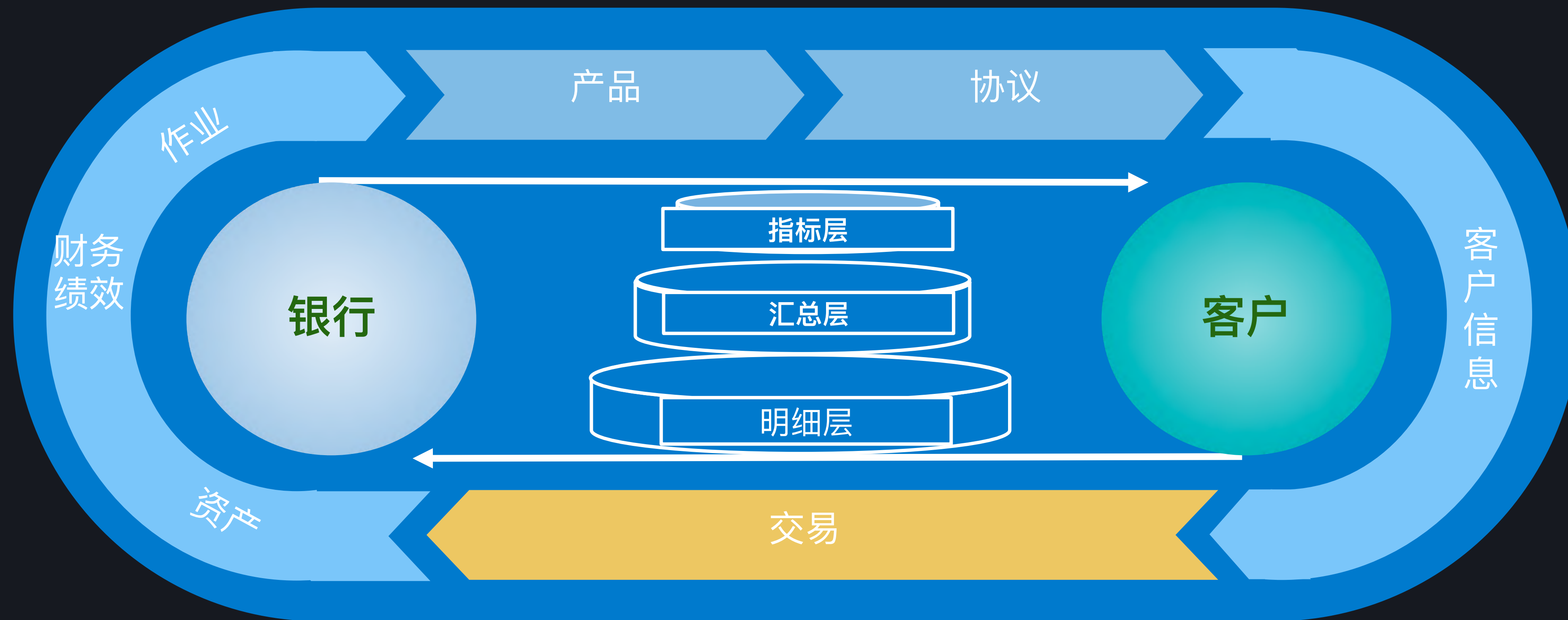
模型的稳定性、可扩展性

- 结构上应该是稳定的、灵活的、可扩展的
- 足够的灵活性才能适应复杂业务情况以及业务的变化
- 高抽象化的模型能便于扩展

模型的可用性

- 便于最终用户理解
- 统一的规范、规则定义、业务语言
- 层次、关系清晰
- 数据无二义性
- 文档完备

公共数据模型主题成果



财务绩效	资产	交易	产品	协议	客户信息
主要是科目总账和统计科目信息	主要是银行持有资产和押品信息	主要是交易、传票以及特定业务、渠道的交易	主要是通用产品信息以及特定产品信息	主要是容器账户、结算账户、授信业务和国结业务	主要是客户基本信息、客户关联信息和客户分类信息

实施过程遇到的问题与对策

跨节点的数据网络传输带来的IO成本

问题：IO成为性能瓶颈

对策：客户号存在并作为分片键，并且在表关联操作中将客户号相等的计算条件作为必要条件；
元数据表尽可能设计为复制表；
避免在分片键上出现空值，导致数据分布过度倾斜



避免复杂的SQL编写

问题：编译器难以判断嵌套SQL在每个节点上的初始结果集是否可以驻留直接使用，结果往往需要汇聚后再广播给每个节点，增加大量的网络IO成本

对策：尽可能把过滤条件放到嵌套SQL内部执行，减少中间结果集大小，降低数据广播带来的处理延迟



分布式计算任务带来的调度协调成本

问题：看似较小成本的SQL实际执行成本（时间）比传统数据库要高很多

对策：包括规避存储过程内游标使用，尽可能用聚合合成复合SQL语句或拆分成若干批量数据更新处理步骤

技术支持工具

问题：跨集群数据自动复制，系统监控、SQL性能分析诊断等技术工具尚不完善。

对策：与平台厂商合作，自开发技术工具作为补充

基础数据平台技术优化工作



➤ 数据权限管理

- 1.多分支机构行级权限管控
- 2.列级权限，用户级数据脱敏定义

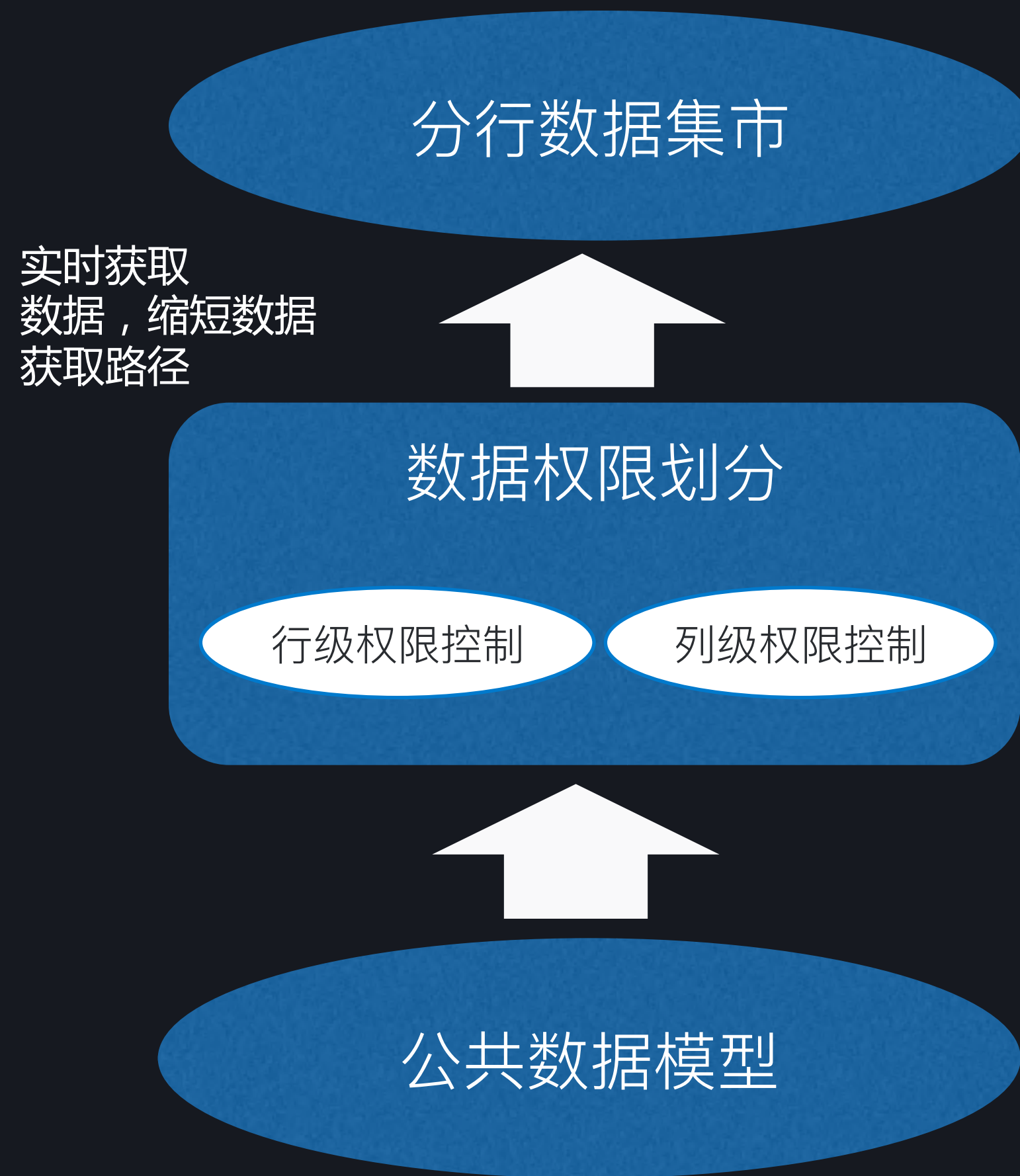
➤ 服务质量管控

- 1.任务级资源管控避免不良设计影响整体性能
- 2.SQL执行成本、执行时间、排队时间等多维质量管控策略

➤ 实时监控预警

- 1.针对实例和组件不同层级的实时监控搭建
- 2.实时采集组件可用性、资源占用情况、任务排队数、平台事件等信息
- 3.配置智能告警规则

数据权限管控



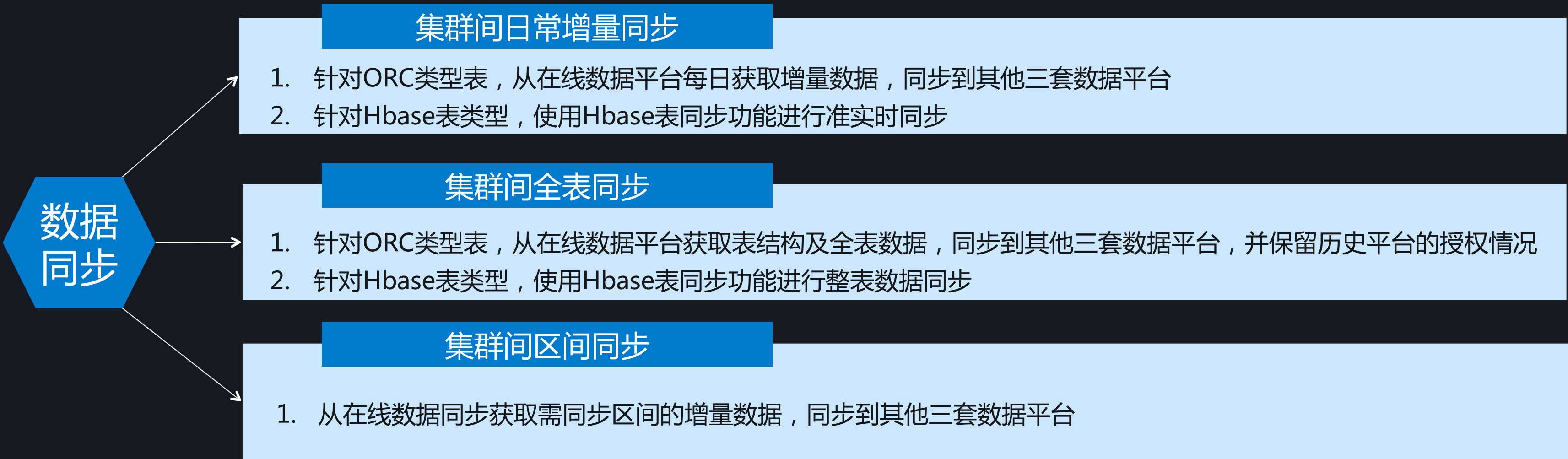
行级权限

- 行级权限实现数据的行级授权
- 分行用户只能查询该分行数据
- 基础模型统一口径加工

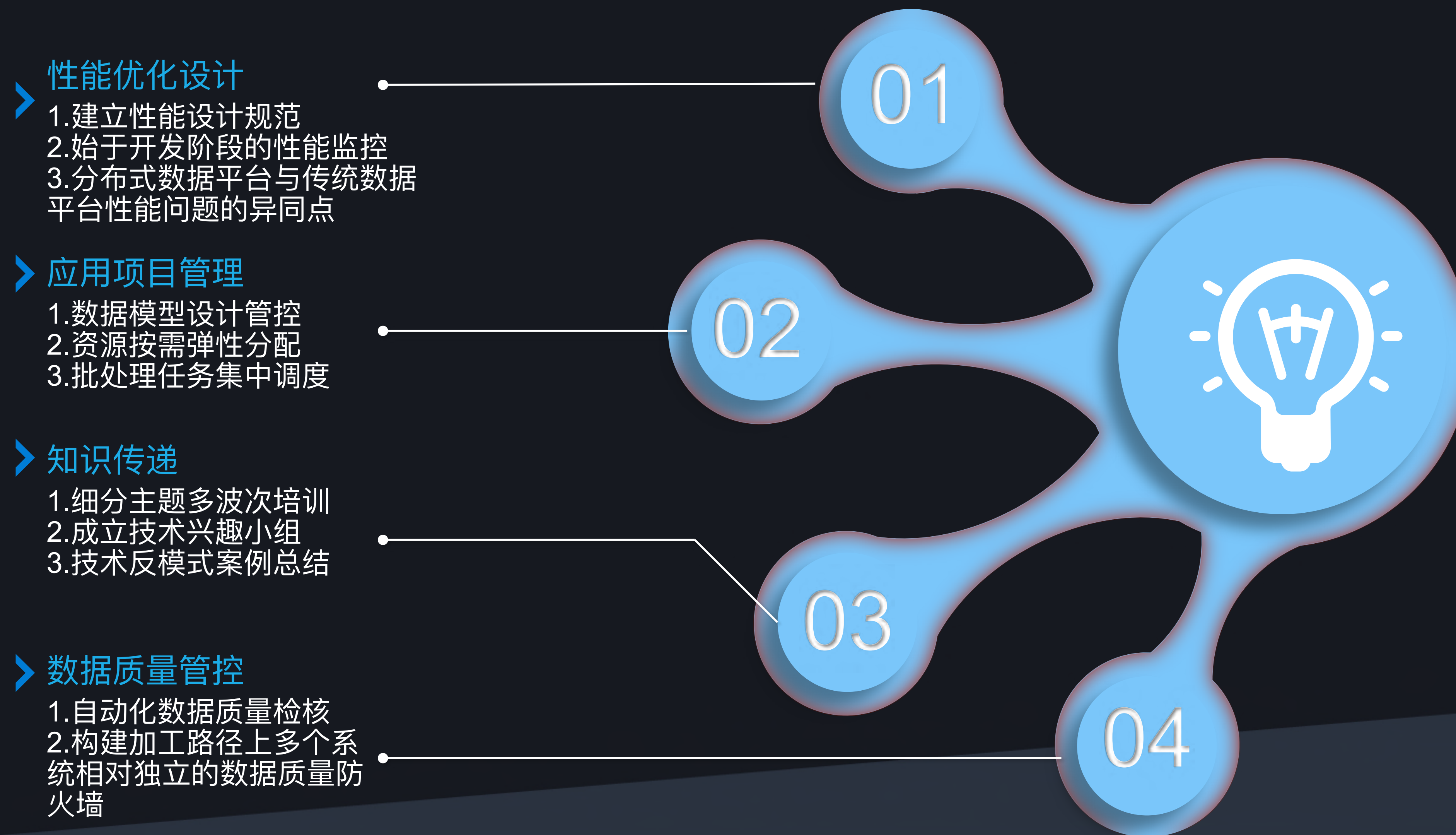
列级权限

- 列级权限实现数据脱敏
- 针对不同用户设定不同的查询结果

集群灾备以及数据同步



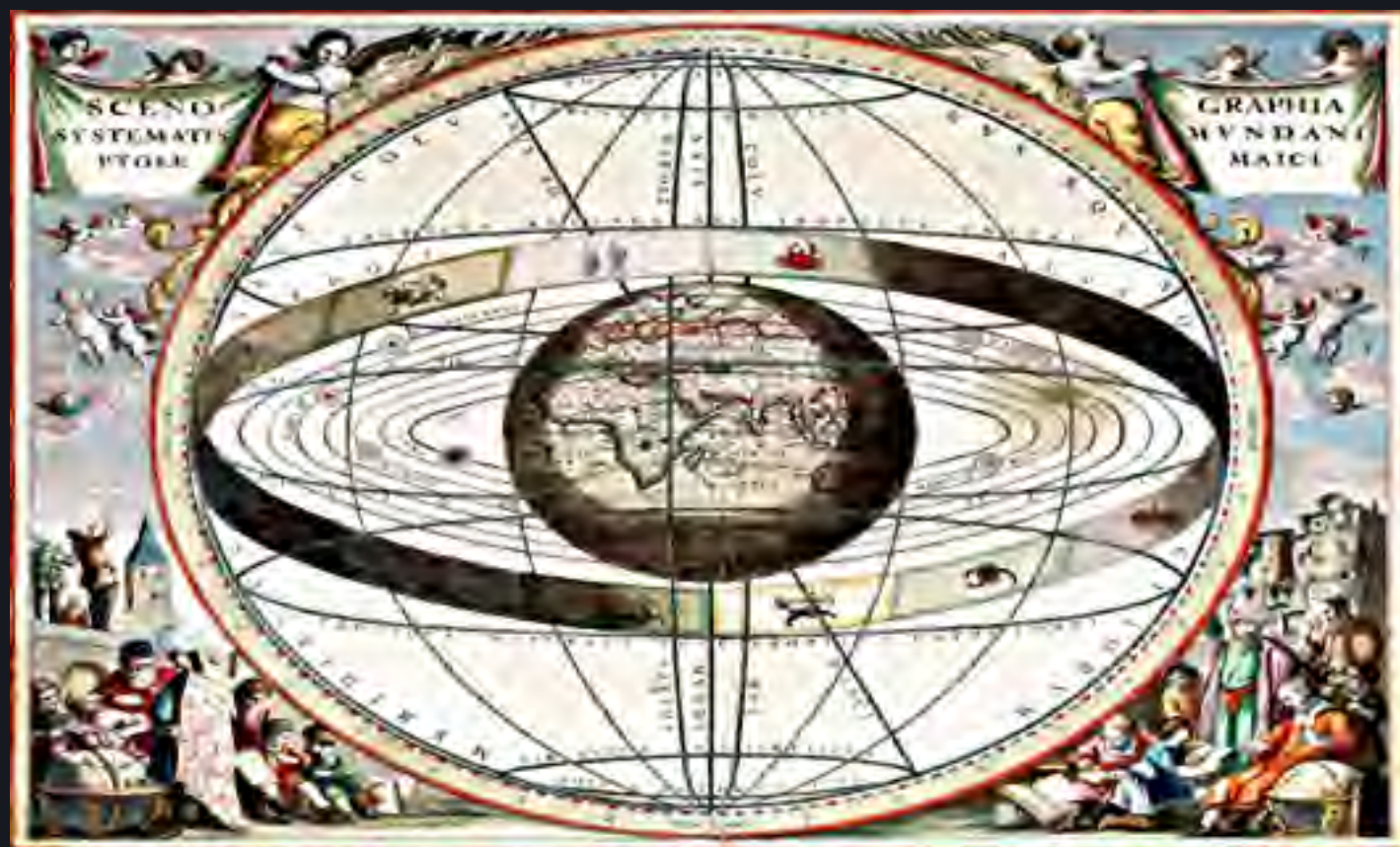
技术实施关键点



目录

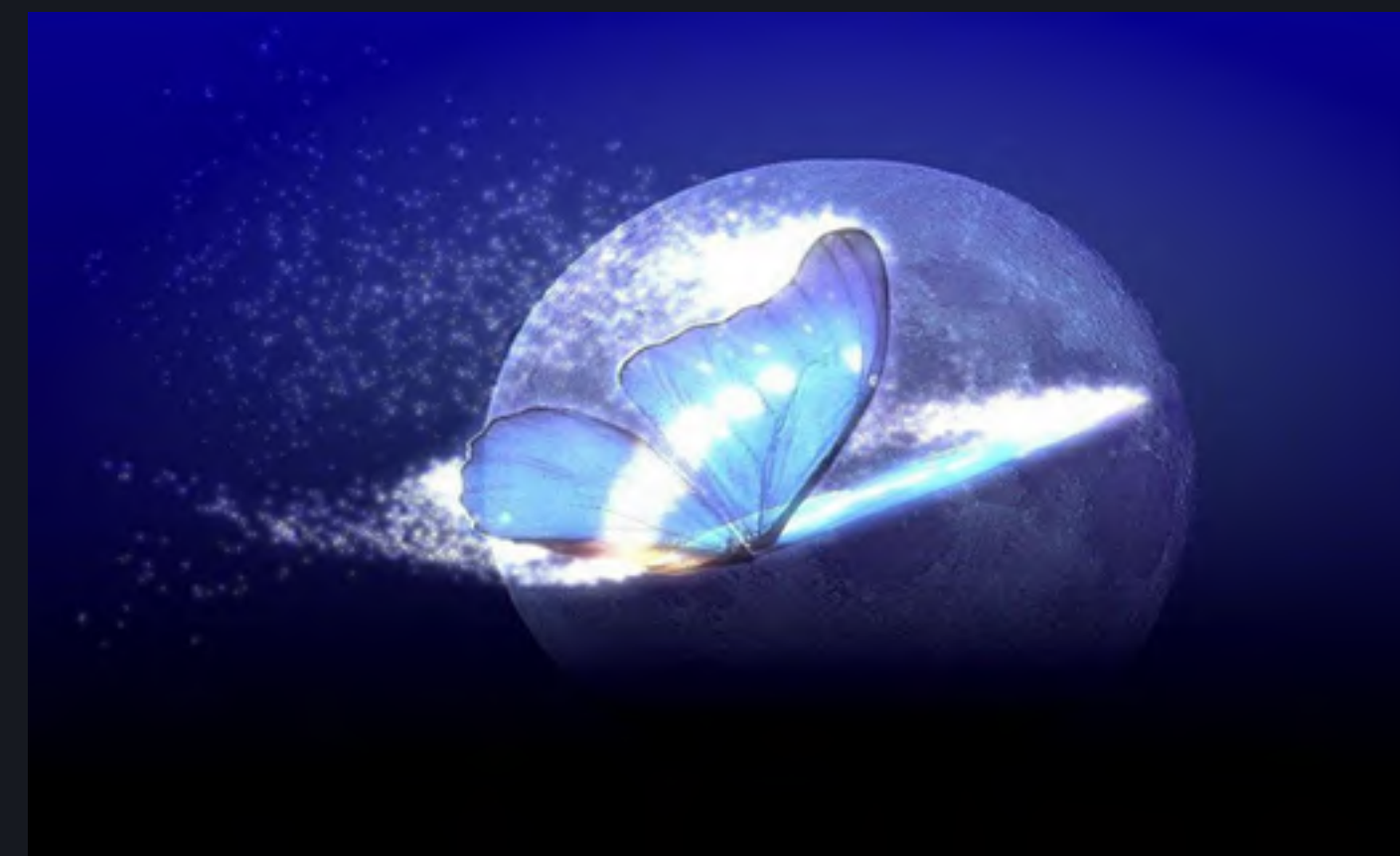
- 1 平台建设背景
- 2 数仓应用体系建设
- 3 风控领域创新应用
- 4 取得成果与未来展望

思路-风险管理的两种思考体系



传统思维方法

- 1.需要准确财报数据才能判断企业经营状况
- 2.选取合适的变量和模型，通过对历史数据的拟合验证，可以预测客户是否将出现信用违约
- 3.评审与贷后环节对人的判断能力非常依赖
- 4.外部环境复杂多变，系统性风险难以防范



大数据思维方法

- 1.大部分客户财报数据不准确，是否能从公开信息变化辅助判断客户经营是否正常
- 2.信用违约的原因很复杂，没有足够多数据支撑模型精准预测，但多个关联信息可以推测违约概率是否在变大
- 3.多个环节引入并更多的客观数据可减少人为判断风险
- 4.可以建立数字化监控体系来对系统性风险感知预警

风险管理应用规划

行业数字地图：

- 1、自动遴选、组织数据，提升规划效能
- 2、量化分析提升授信决策效率

3、感知行业趋势变化，敏捷应对系统性风险

客户交易风险管理：

- 1、流处理技术实现实时欺诈风险识别
- 2、自主学习和优化的反欺诈模型
- 3、识别和阻断事中风险

柜员操作风险管理：

- 1、智能决策引擎实时产生预警工单
- 2、自动化风险审查作业模式
- 3、提升内审作业效率，防范道德风险

业务协作、风控前置：

- 1、便捷采集申请信息，提升流程效率
- 2、自动交叉验证外部数据
- 3、减少人为判断，规避人员道德风险

全面的信用风险预警服务：

- 1、多维风险视图提升决策效率
- 2、知识图谱技术分析风险传导，提升风险感知能力
- 3、智能决策引擎汇聚专家智慧，降低人员要求
- 4、信用欺诈检测有效防范道德风险



风险管理全生命周期系统支持

业务规划

客户引入

业务办理

客户存续

业务规划平台：

1. 区域行业发展趋势分析
2. 行业龙头企业动态跟踪
3. 潜在客户智能检索
4. 授信政策制定

移动信贷应用：

1. 目标客户信息采集
2. 贷前风险排查
3. 额度测算

信用风险预警：

1. 企业风险视图
2. 简易评分卡
3. 风险预警信号

信用风险预警：

1. 客户风险预警
2. 评分卡
3. 风险决策树
4. 模型与策略
5. 信用欺诈检测

运营风险监测：

1. 柜员操作风险
- ## 交易反欺诈：
1. 客户交易风险

信用风险预警：

1. 客户风险预警
2. 组合风险预警
3. 信用违约预测
4. 风险影响分析
5. 风险缓释

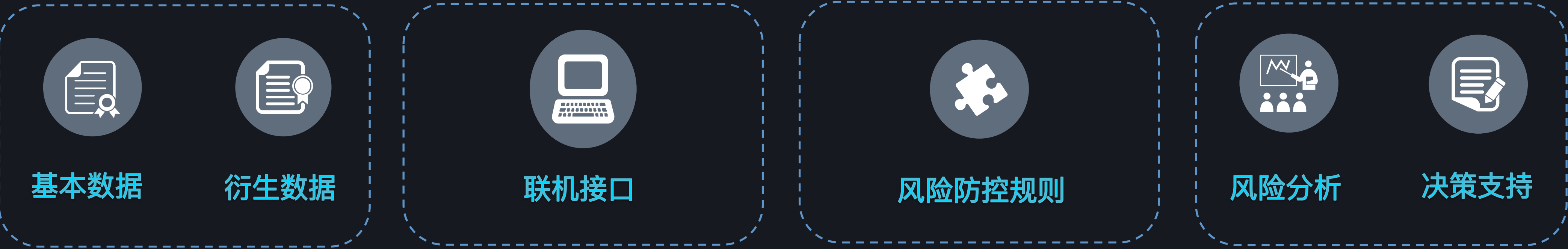
移动信贷应用：

贷后检查

业务规划平台

授信政策重评估

风险预警系统-业务能力规划



基础数据支撑服务 统一联机接口服务 统一风险防控规则配置服务 综合风险分析、决策支持



数据价值挖掘——群体信用违约预测模型

要解决的问题

识别股权、交易、担保网络的系统性风险
规避循环担保、过度授信
量化企业违约风险



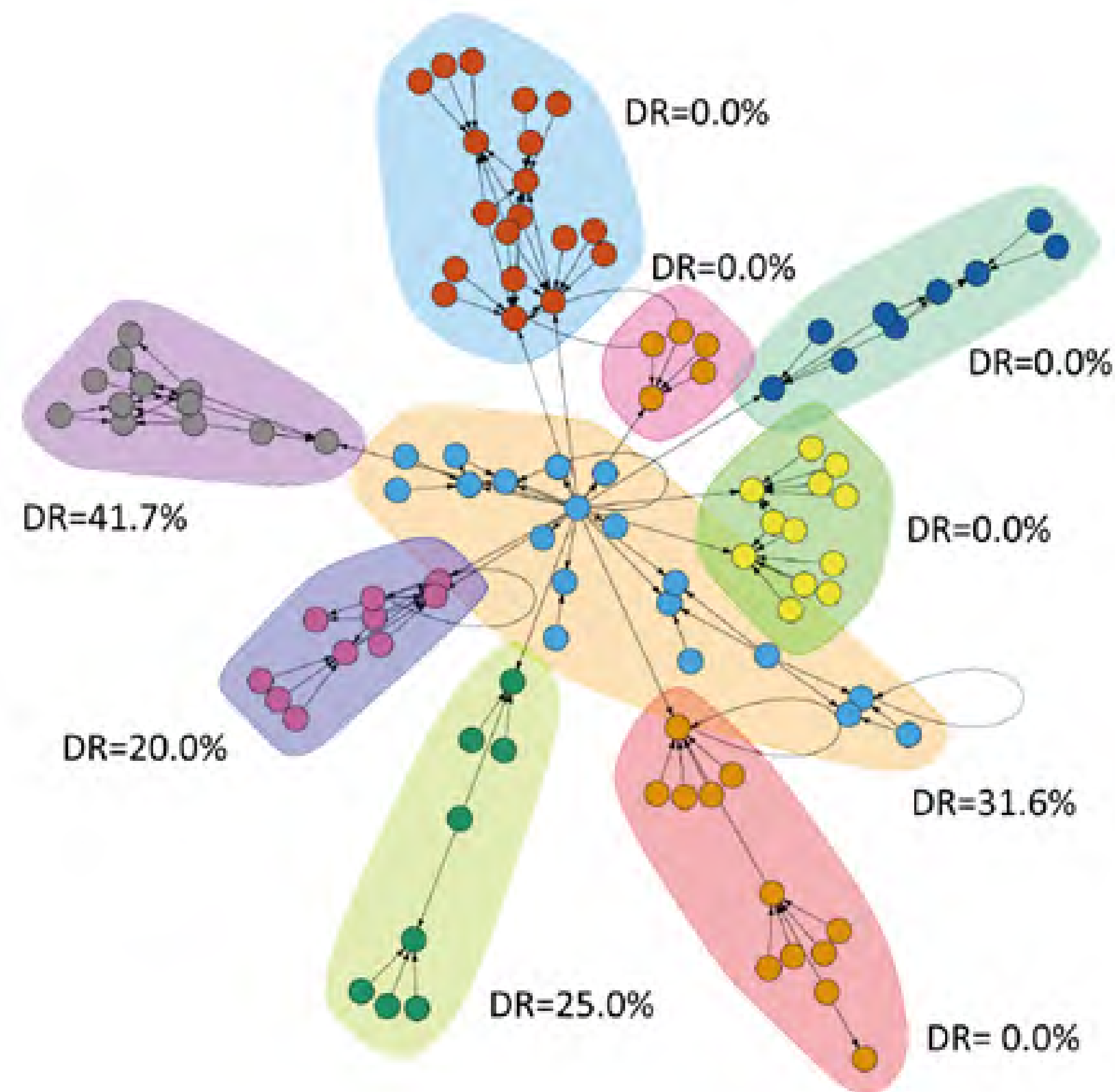
技术实现

客户属性、信用行为、关系图谱、社区特征
复杂网络技术挖掘违约风险影响分子
关系链图特征、客户行为特征建模



模型输出

关系网络可视化风险监控
高风险预警客户名单输出
授信评审阻断策略因子



交易反欺诈系统

对全渠道的客户交易行为进行全方位监控，最大程度避免客户的经济损失。

覆盖全
电子渠
道

实时
侦测

灵活的
规则配
置

一站式
管理平
台

风险事
件持续
跟踪

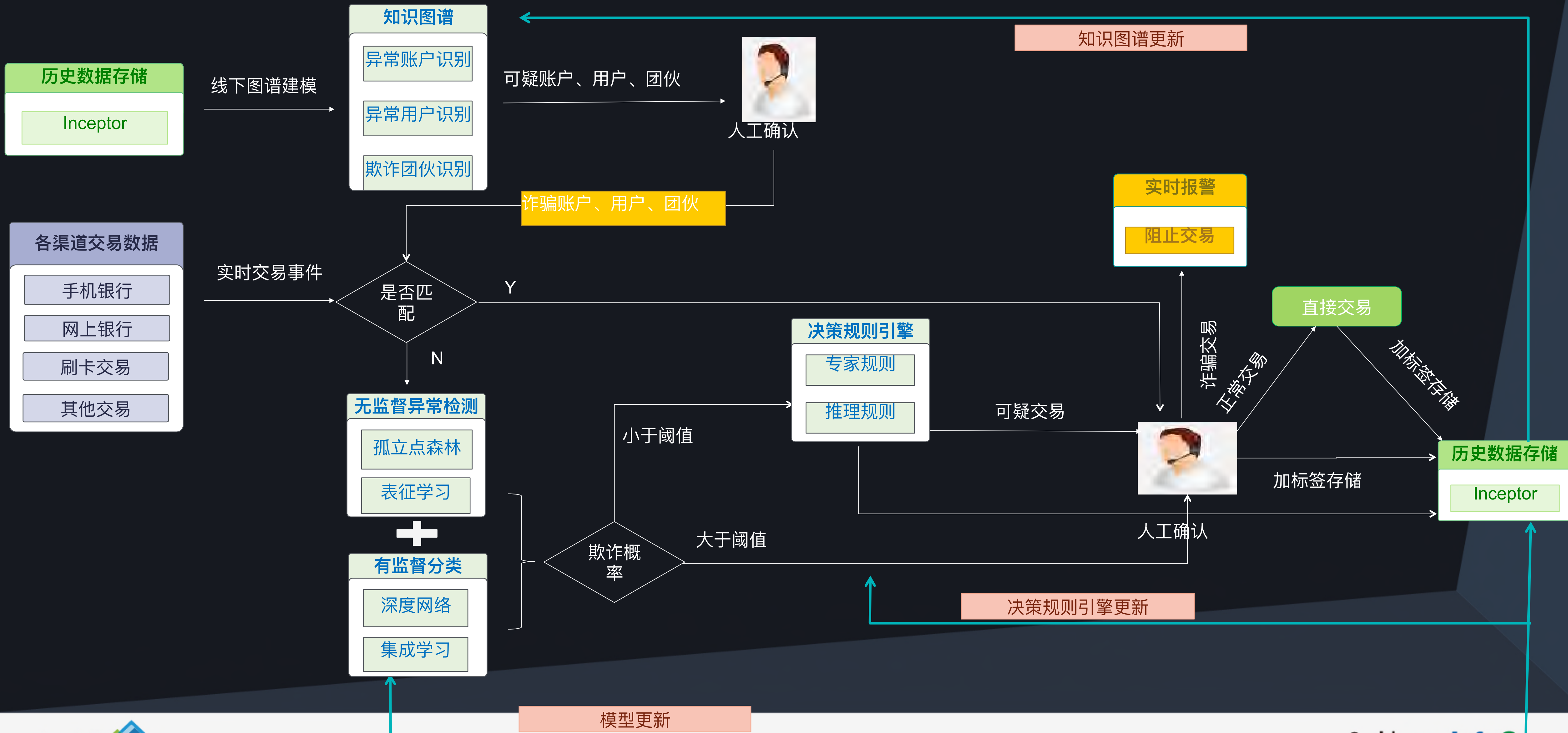
实时采集客户交易数据，支持实时风险侦测，对于欺诈交易实施事中阻断，对于可疑事件进行事后提醒。

对可疑事件进行持续跟踪，建立风险等级评定体系。

建立风险特征模型库，支持灵活且快速部署新的业务规则，能及时有效地防范各类风险。

对接客服系统和短信平台，系统内实现一键触达客户；且支持报表统计。

数据价值挖掘——交易反欺诈应用流图



实施过程的困难与挑战

外部数据成本过高

缺乏高度整合的公开市场数据

政府、国营事业单位数据时效性差

业务团队风险理念转变

高层强力支持

风控流程和操作管理办法的适应调整

智能技术成熟应用需要时间

学术成果的转换成本、试错成本

数据科学人才缺乏、培养需要时间

行业知识图谱构建缺乏业务专家

外部信息源、产业标准化信息

实体逻辑关系、风控专家规则




目录

- 1 平台建设背景
- 2 数仓应用体系建设
- 3 风控领域创新应用
- 4 取得成果与未来展望

原数仓

 数据量规模 6T

 模型处理个数 1500


 处理效率 13个小时

VS

大数据平台

 数据量规模 400多T

 模型处理个数 2300

 处理效率 2个小时

单个模型效率对比：

□ 处理效率是指，从接入核心系统数据算起到模型数据处理完成的日终时间。

以存款账户表为例：

- 数据处理逻辑相近，数据量一致
- 原数仓存款模型 --平均耗时：3小时30分钟
- 大数据存款平台模型 --平均耗时：39分钟

取得成果-提升全行数字化支撑能力



实时的大数据平台能力

- 01.13个部门集市、18个分行集市、26个应用集市管理380TB数据，日实时处理200万交易数据
- 02.对外发布110个服务接口，月均调用近200万次
- 03.对接26个外部数据源，月采集企业舆情80万条每天聚合8大行业资讯、23类市场指数、200多份投研报告

创新应用助力业务发展

- 01.32个创新应用，发布1100个业务功能
- 02.全行2500个用户，月均使用6万余次
- 03.335张业务可视化报表，每次支撑6000次统计分析
- 04.月推送实时业务提醒60万条，累积推荐潜在客户5万户

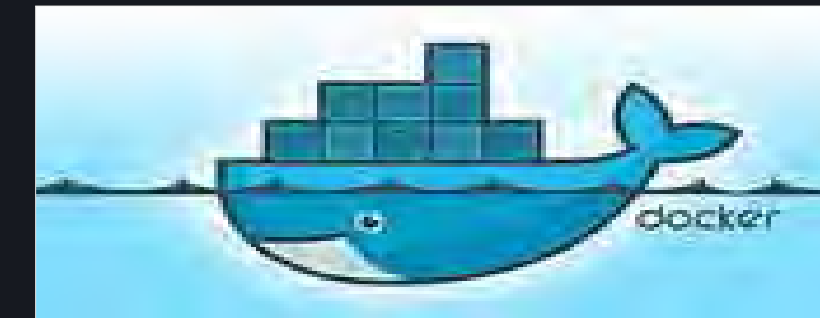
全天候的风险监测体系

- 01.实时跟踪30万行业标杆客户
- 02.300多个信用风险预警规则，年触发风险预警信号8000次
- 03.各类平台贷风险服务接口23个，月均调用5万余次
- 04.196个运营风险监测模型，月均生成工单4500笔

工作展望

1

基于Docker 容器技术，建设面向多租户的大数据平台，实现更细粒度的资源管控与调度



2

研发面向业务团队、可定制的实时智能业务决策引擎，满足实时营销、实时风险管理、实时交易反欺诈等多种场景需求。



3

融合行内外多样化数据，深度提炼数据价值，优化业务领域模型，为业务一线和客户提供更多场景的智能化数据服务



4

建立行业知识图谱技术平台，满足客户价值评估，行业风险传导路径，重大事件影响分析等业务需求



THANK YOU

如有需求，欢迎至 [\[讲师交流会议室 \]](#) 与我们的讲师进一步交流

