ArchSummit
全球架构师峰会 2017

主办方 Geekbang. 极客邦科技 InfoQ

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

主办方 Geekbang. 极客邦科技 InfoQ

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

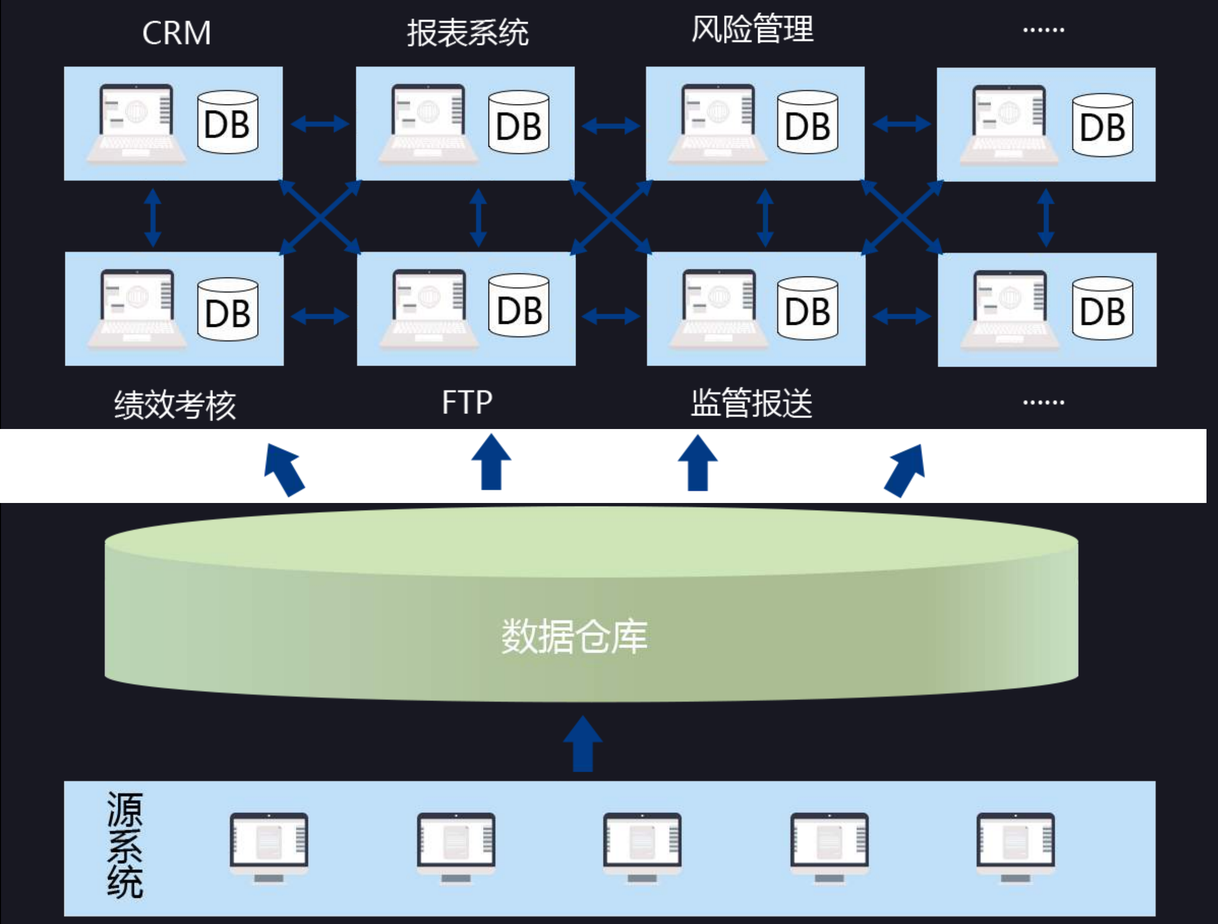
1 平台建设背景

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

03 数据治理目标难实现：企业数据模型、数据标准统计口径

04 硬件资源分散
峰值服务能力和大数据量处理能力受限

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
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

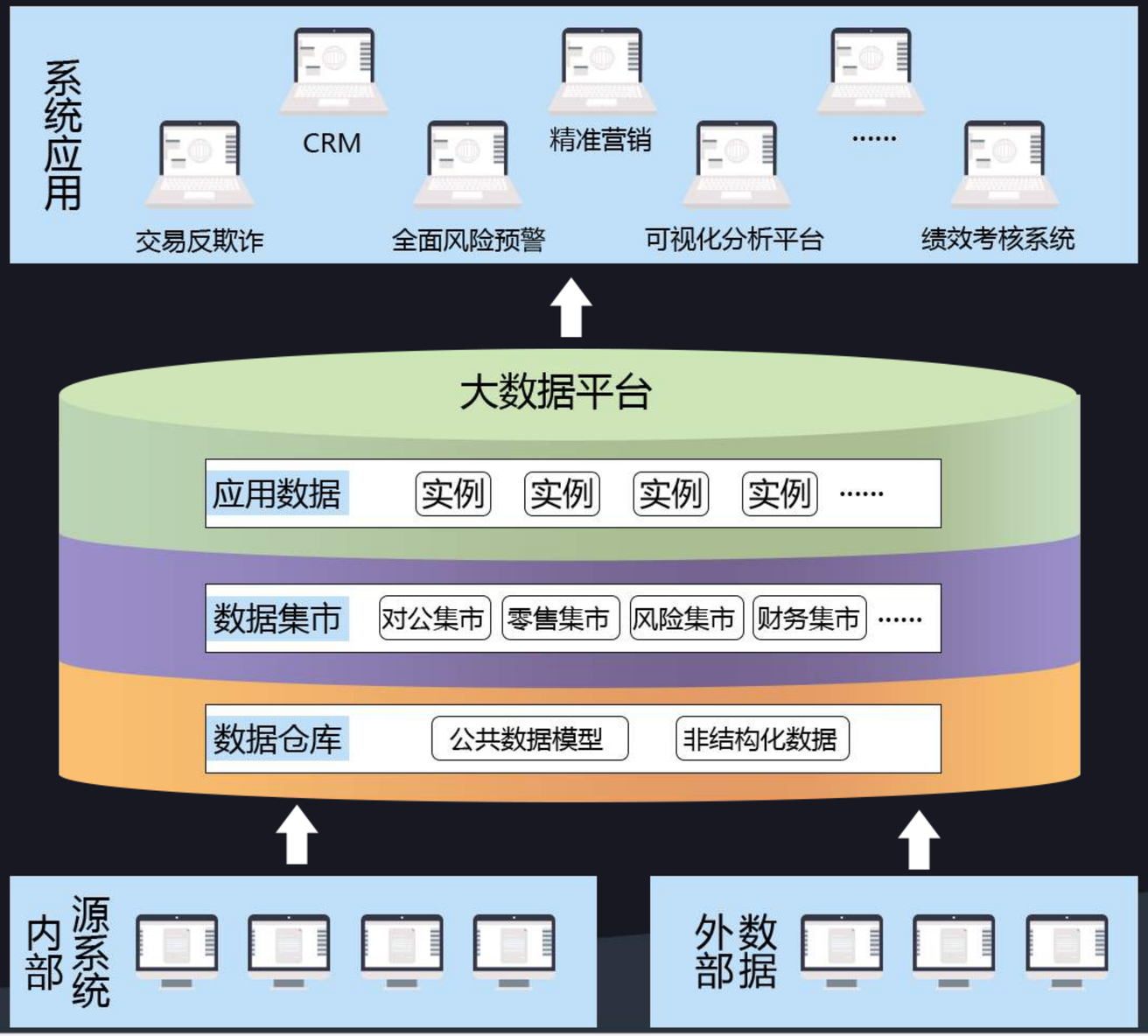
运维简单

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

01 统一规划企业信息模型

从全局视角规划设计，整合提炼多应用共性需求，构建公共数据模型层

02 一份数据，直接共享

统一调度，集中授权，直接访问，统计口径一致，实现数据标准化目标

强化数据仓库核心地位

促进数据应用开放协同

03 应用轻装上阵

削减基础数据加工任务，应用数据集市瘦身，共享融合的企业级信息视图

04 数据服务专业化

聚焦专业领域数据价值提炼，开放专业数据视图与应用服务

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

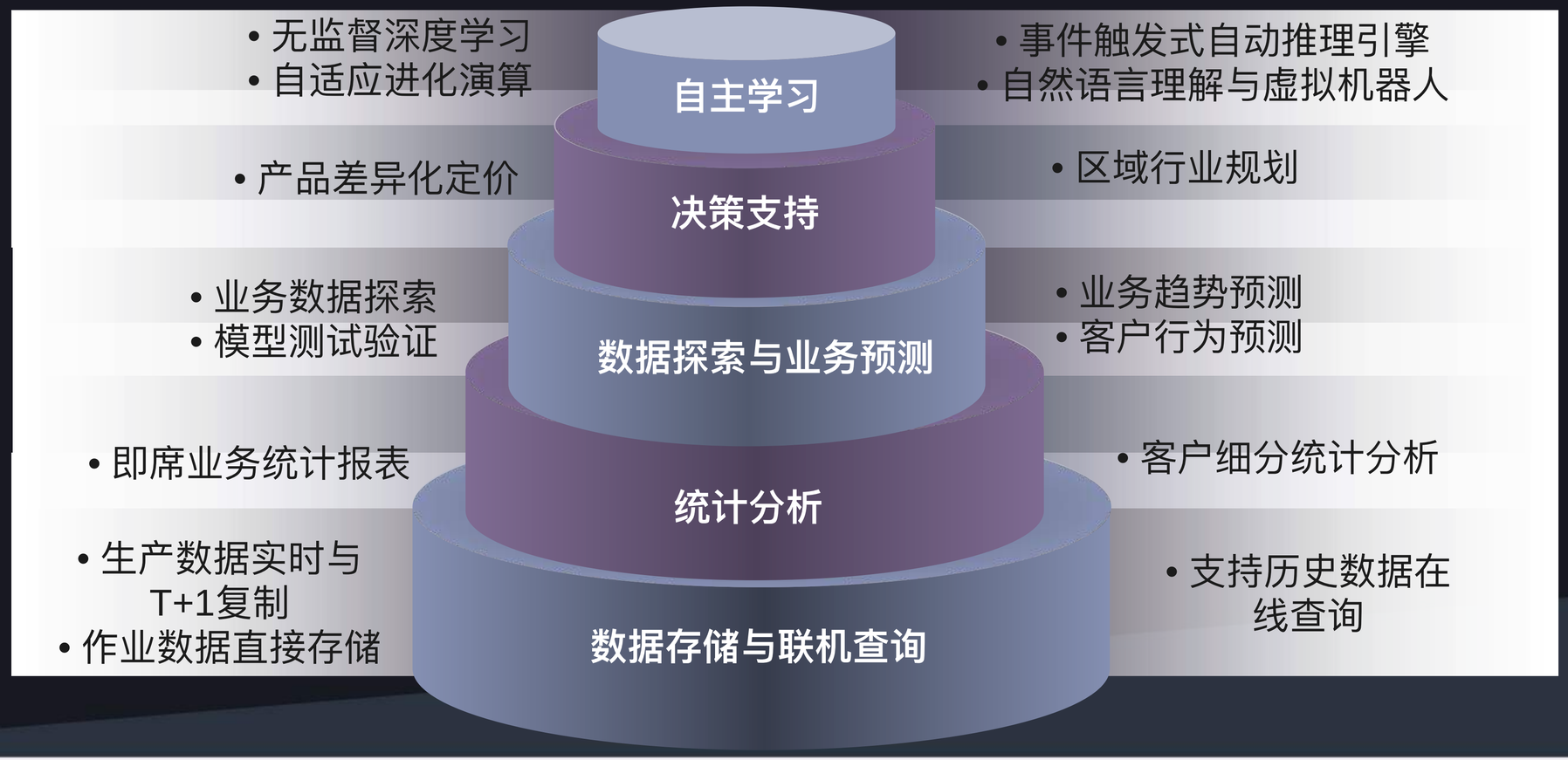
1.高并发低延迟的微服务架构
2.大数据技术集成

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

离线批处理 → 准实时 → 实时



目录

1 平台建设背景

2 数仓应用体系建设

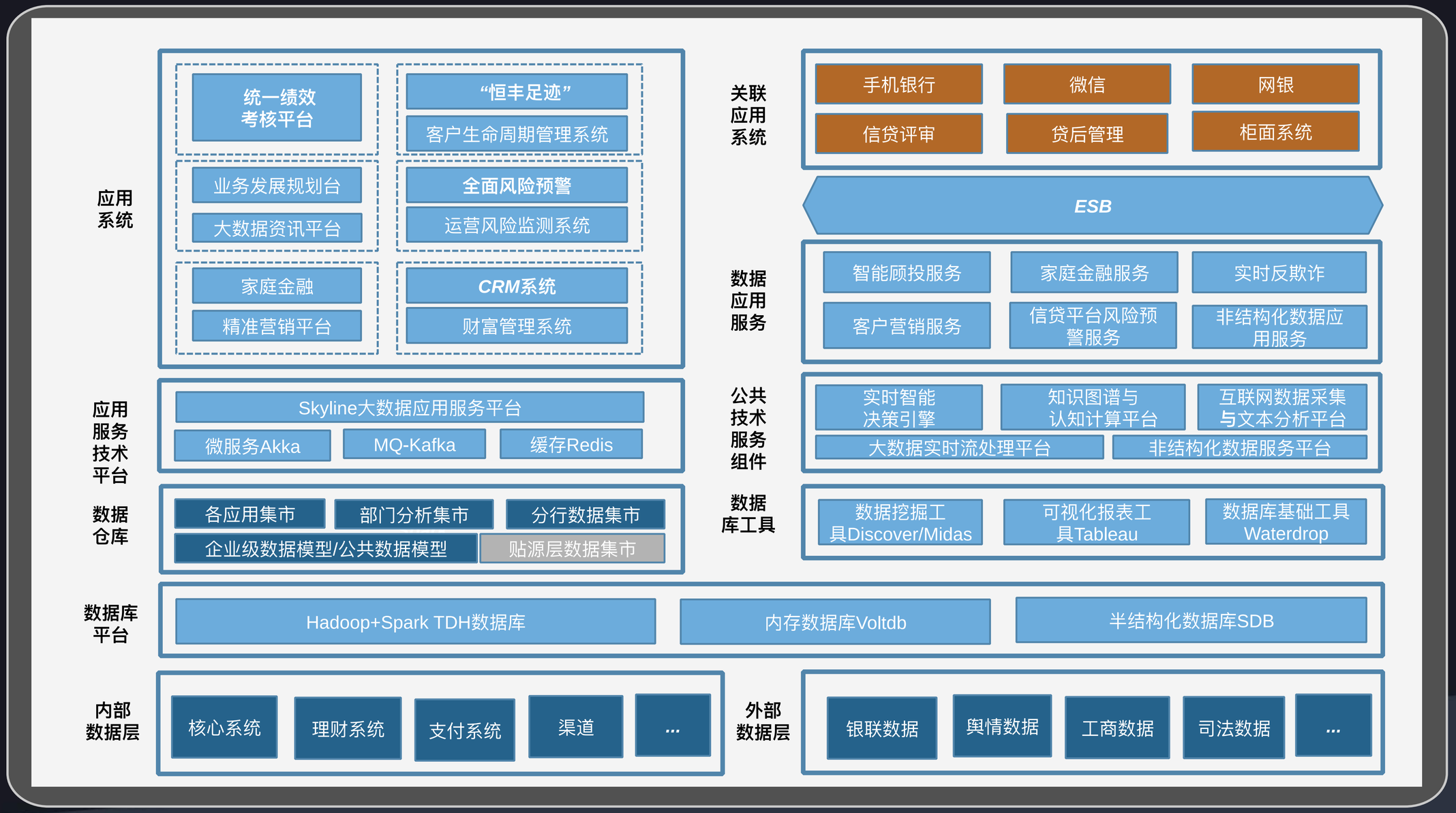
3 风控领域创新应用

4 取得成果与未来展望

数仓基础应用架构

数据运维体系
统一调度平台
大数据管理平台
综合监管集市
数据分析集市
客户关系管理集市
审计、反洗钱等其他
数据集市
历史数据服务接口
数据服务接口DSI
数据服务
历史数据平台
非结构数据区
社交媒体信息
第三方数据
...
公共数据模型层CDM
源数据历史层HDM
基础数据模型层FDM
在线数据平台
公共数据模型层CDM
源数据历史层HDM
基础数据模型层FDM
数据存储
非结构化数据接入
源数据缓冲区ODM/结构化数据接入
文件交换区FSA
数据接入
半结构化/非结构化数据
源系统结构化数据
源数据

数据应用总体技术架构


应用系统
统一绩效考核平台
“恒丰足迹”
客户生命周期管理系统
业务发展规划台
大数据资讯平台
全面风险预警
运营风险监测系统
家庭金融
精准营销平台
CRM系统
财富管理系统
关联应用系统
手机银行
微信
网银
信贷评审
贷后管理
柜面系统
ESB
数据应用服务
智能顾投服务
家庭金融服务
实时反欺诈
客户营销服务
信贷平台风险预警服务
非结构化数据应用服务
应用服务技术平台
Skyline大数据应用服务平台
微服务Akka
MQ-Kafka
缓存Redis
公共技术服务组件
实时智能决策引擎
知识图谱与认知计算平台
互联网数据采集与文本分析平台
大数据实时流处理平台
非结构化数据服务平台
数据仓库
各应用集市
部门分析集市
分行数据集市
企业级数据模型/公共数据模型
贴源层数据集市
数据库工具
数据挖掘工具Discover/Midas
可视化报表工具Tableau
数据库基础工具Waterdrop
数据库平台
Hadoop+Spark TDH数据库
内存数据库Voltdb
半结构化数据库SDB
内部数据层
核心系统
理财系统
支付系统
渠道
...
外部数据层
银联数据
舆情数据
工商数据
司法数据
...

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

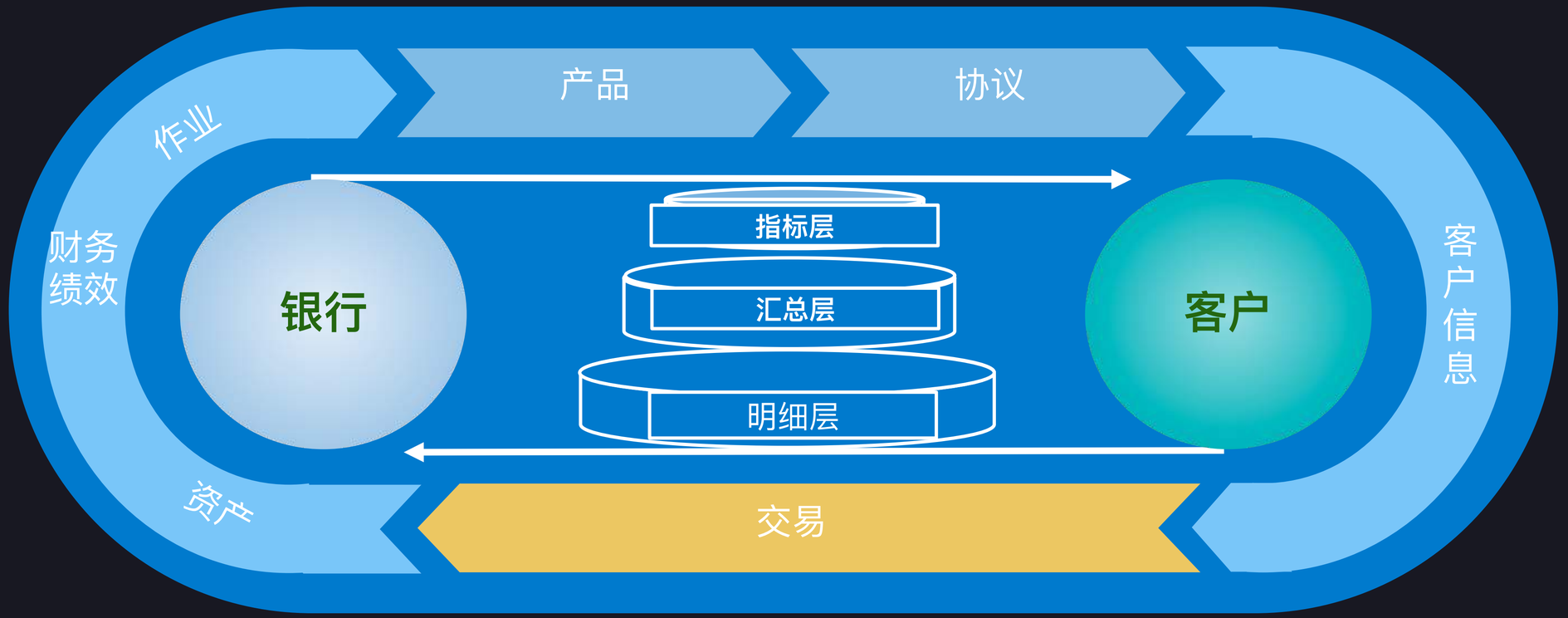
模型的稳定性、可扩展性

- 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