

【数据治理实践】第十二期：数据架构的设计与规划

前言

[上期文章](#)，我们围绕个人隐私数据保护主题，分享了数字化金融领域的个人金融信息保护背景和现状，以及如何结合业务发展要求、合规要求、客户要求进行落地的方案建议。本期，我们将介绍如何构建匹配银行整体信息科技架构、支撑数据治理工作落地展开的数据架构。

基于数据架构的重要性和适用性，我们需在深入分析银行数据架构现状的基础上，结合同业实践与银行实际需求，对数据架构的内容、配置数据治理机制等进行设计与规划。

监管要求

《银行业金融机构数据治理指引》（以下简称指引）明确提出，需要持续完善信息系统建设，并在系统层面上加强对数据采集、数据交换、数据共享的统一管理。

第十七条 银行业金融机构应当结合自身发展战略、监管要求等，制定数据战略并确保有效执行和修订。

第二十一条 银行业金融机构应当持续完善信息系统，覆盖各项业务和管理数据。信息系统应当有完备的数据字典和维护流程，并具有可拓展性。

第二十三条 银行业金融机构应当加强数据采集的统一管理，明确系统间数据交换流程和标准，实现各类数据有效共享。

数据架构的规划

相较于业务架构和应用架构，因为数据能够真实而有效地反映出信息系统支撑下的银行全面运作状况，所以数据架构在整体信息科技架构中，显得既基础又核心。对于银行数据治理来说，银行需要从整体性需求与长远性发展的角度出发，对数据架构进行针对性的设计与规划，实现数据采集、交换、共享等标准在实际业务数据流中的规范化、一致性、准确性和完整性，也只有在此基础上，才能充分挖掘数据价值，夯实数据治理体系，从业务、风险、管理和技术等方面持续而有效地支持数据管理和经营分析决策，实现银行数据驱动发展的战略。

业务：**灵活**地支持业务分析，以客户和产品的360视图分析。

风险：**全面**实现数字化风险管理，识别各个阶段和关键节点的风险。

管理：**及时**准确反映运营状态，汇总管理驾驶数据，支持监管和审计各项数据。

技术：**持续**地应用新兴技术，系统化的方式，持续有效的迭代更新。

数据架构实践

根据德勤最佳实践，数据架构的设计与规划，需要从银行整体的IT规划出发，结合业务架构、技术架构、运维规划等情况，从如下几个方面进行数据架构的构建。

设计原则

结合银行业务发展战略、数据现状调研与评估分析阶段的工作成果，遵循数据架构设计原则，紧密结合应用架构规划的成果，制定数据架构全面发展策略，从架构层面保证数据管控的有效和数据质量的提升。需要遵循的基本原则如下：

原则1. “数据解耦原则”

不同的平台应将数据剥离解耦，借此形成独立数据，这样不同的业务场景和分析工具都能共享同一份数据来达成处理或应用目的。

原则2. “无环依赖原则”

每个独立平台之间，数据的流转不应该出现循环依赖的情况。

原则3. “契约式设计”

不同平台之间、平台内部各模块之间的交互，都是基于“契约”（接口或抽象），而无需依赖于具体实现。

原则4. “组件适当性原则”

每个组件都有自己擅长的领域，要考虑数据的结构（结构化还是非结构化），处理的时效性

(毫秒级或分钟级)，以及吞吐量和访问模式等方面。

功能框架

从数据的**来源、支撑、存储、分析、应用、治理**六个部分，定位数据架构规划的功能框架。

a) **数据来源层**：包括传统的数据库，数据仓库，分布式数据库，NOSQL数据库，半结构化数据，无结构化数据，爬虫，日志系统等，是大数据平台的数据产生机构。

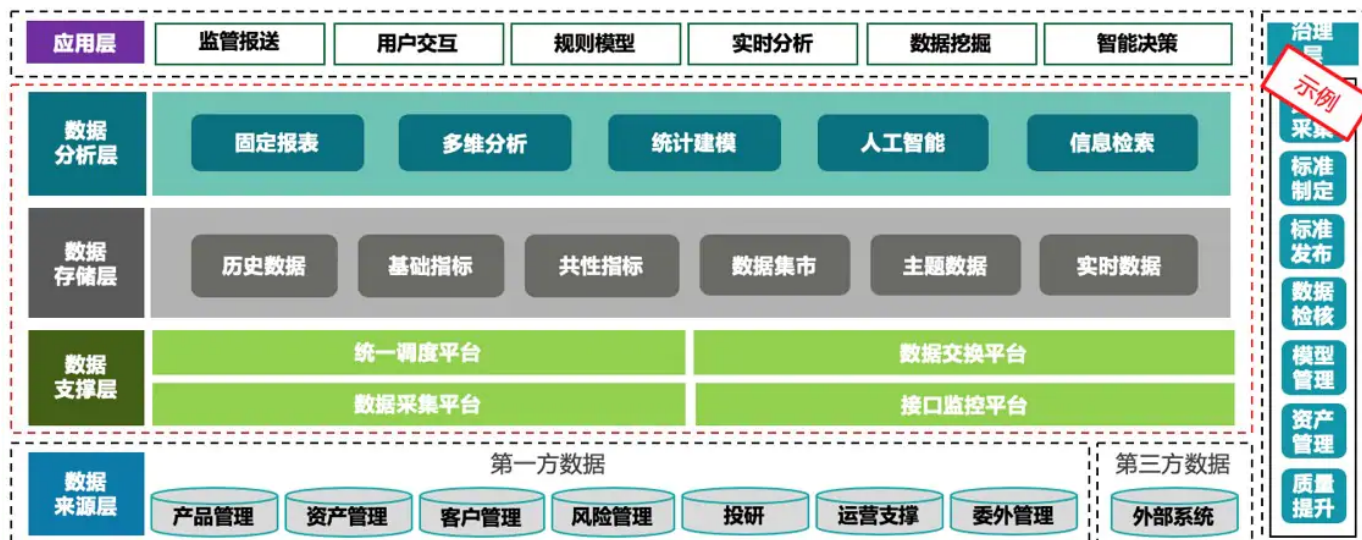
b) **数据支撑层**：包括数据清洗、数据转换、数据加工、数据关联、数据标注、数据预处理、数据加载、数据抽取等工作，该层的作用是对原始数据进行加工。

c) **数据存储层**：存储了经过清洗处理后的可用于生产系统的数据，比如元数据，业务数据库，模型数据库等，该层直接面向应用系统，要求高可靠、高并发、高精度。

d) **数据分析层**：该层实现对数据的深加工，根据业务需要，建立适用于业务的数据统计分析模型，建立大数据运行处理平台，运用数据分析、数据挖掘、深度学习等算法从生产数据集中挖掘出数据内在的价值，为业务系统提供数据和决策支持。

e) **数据应用层**：深入分析行业数据特点，梳理行业数据产品需求，建立适用于不同行业、场景的数据应用产品。

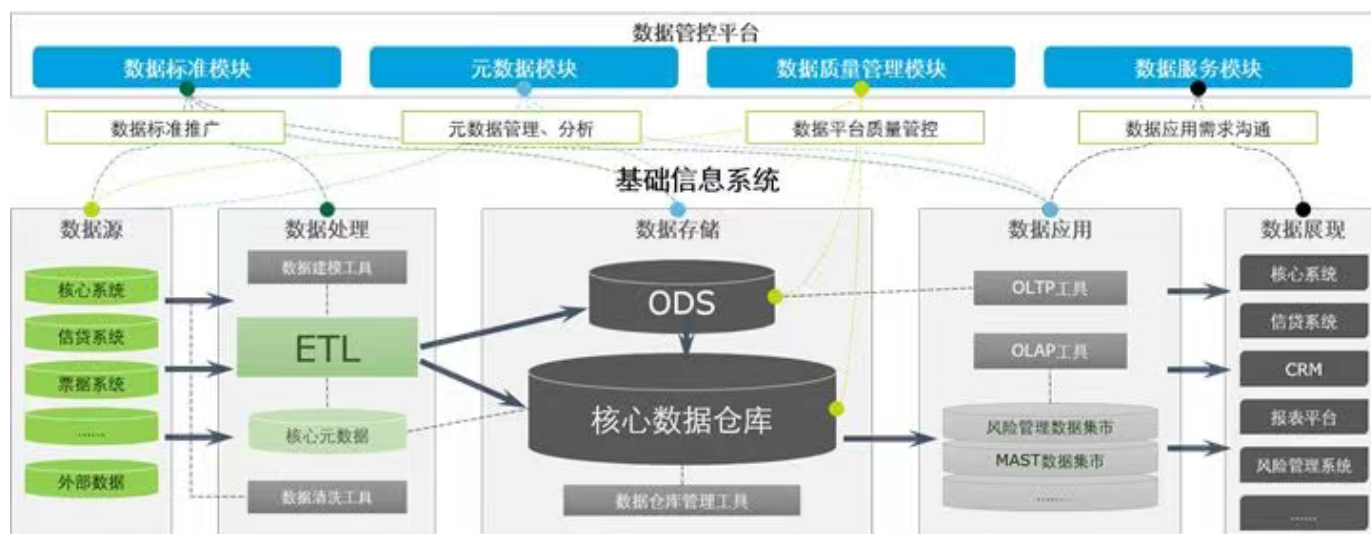
f) **数据治理层**：通过数据标准制定、标准发布、数据检核、模型管理、资产管理、质量提升等功能，明确数据规范，提升数据质量，为数据治理提供规范化支持。

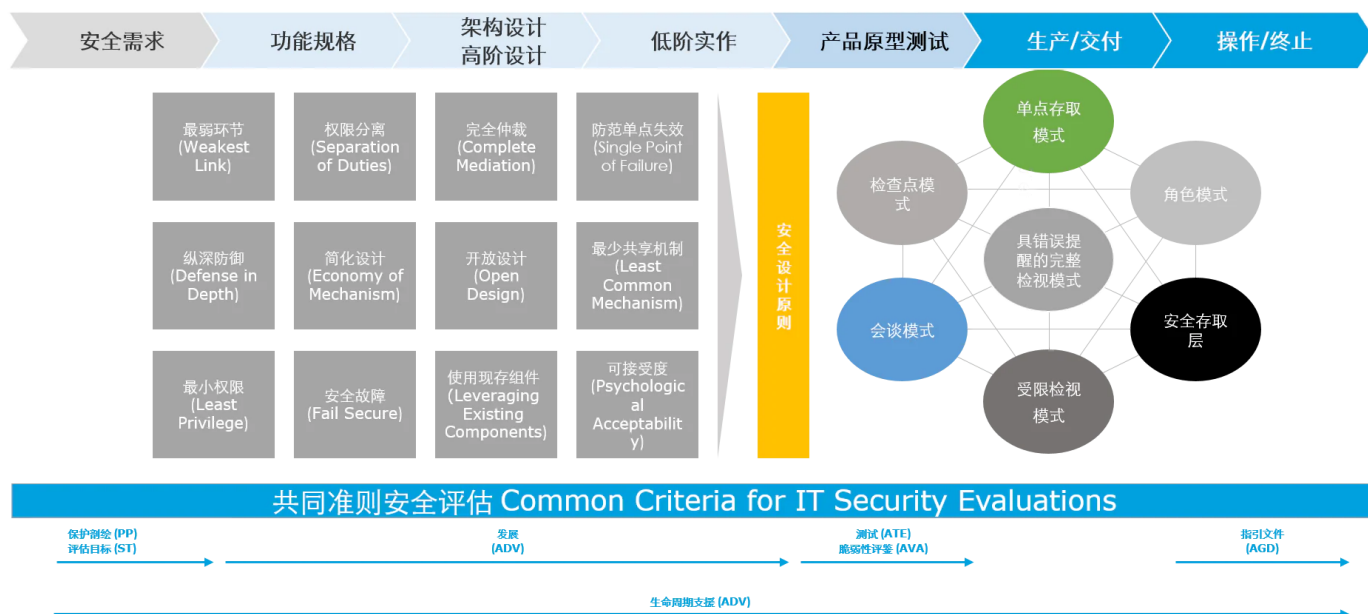


图：数据架构规划的功能框架

技术框架

基于不同的应用场景，通过设计独立的子数据架构，结合最先进的技术实践和解决方案，满足银行不断变化和更高的业务需求。



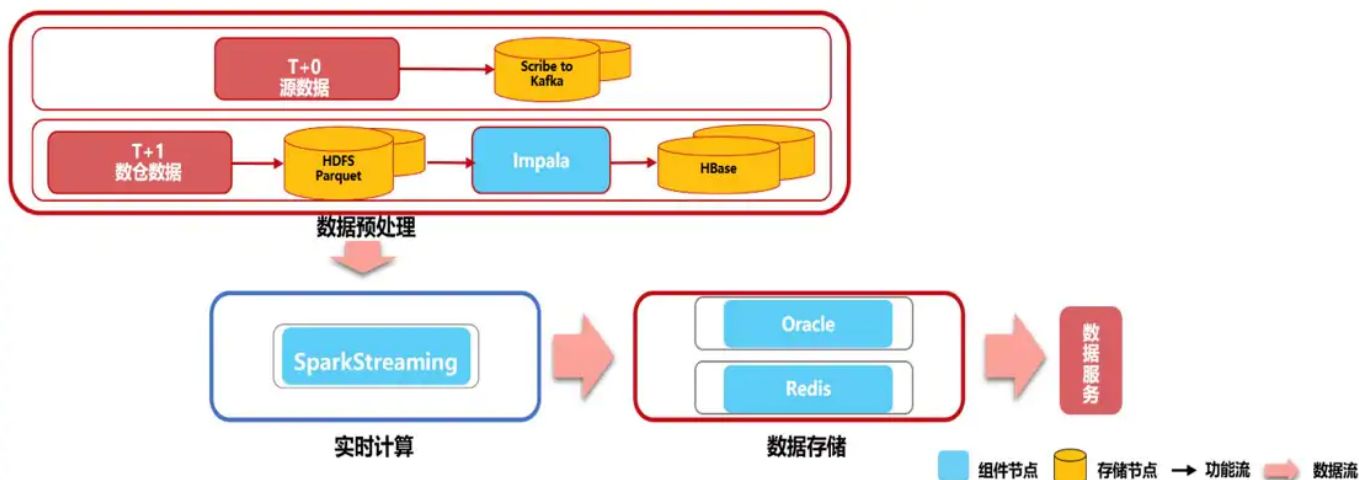


场景示例1. 线上营销业务

该场景的特征在于数据处理及结果返回时效性在200毫秒~1分钟内，主要的技术架构要素是SparkStreaming，平台开发要素是对日切场景的定义与SparkStreaming的性能优化。

解决方案：

- 通过Kafka将实时数据流（T+0）引入SparkStreaming实时计算框架中
- 通过HDFS、Impala、HBase处理后的数仓数据（T+1）引入SparkStreaming计算框架中

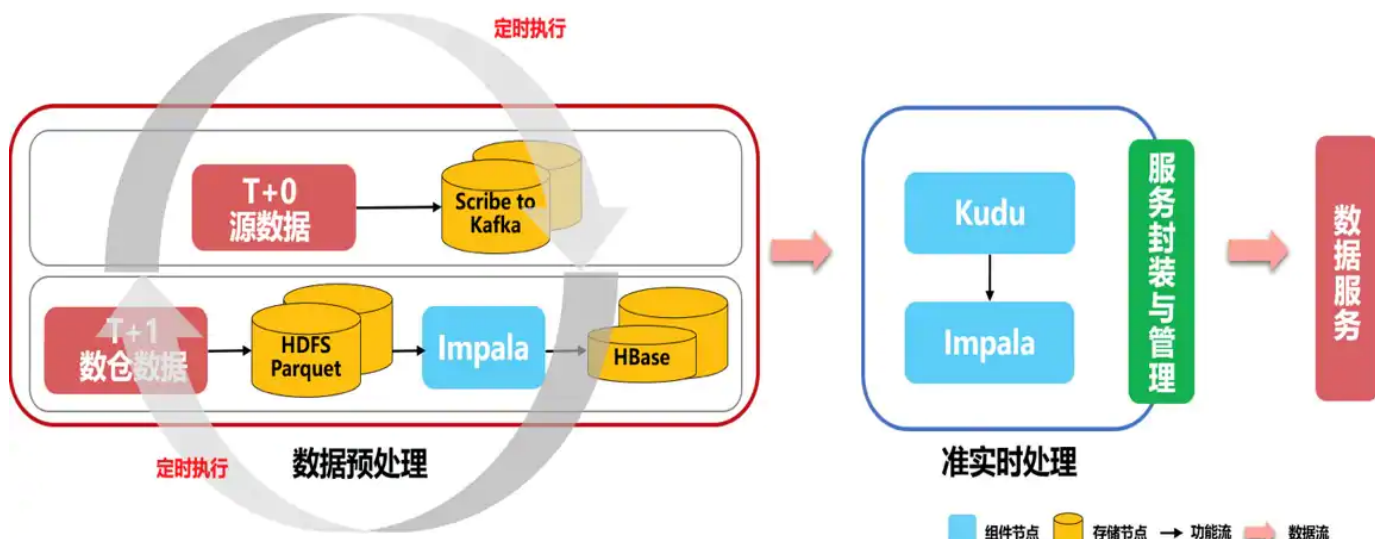


场景示例2. 用户账户变更统计

该场景的特征在于数据处理及结果返回时效性在1分钟~5分钟内，主要的技术架构要素是Kudu/Impala，平台开发要素是对微批、日切场景的定义与Kudu的性能优化。

解决方案：

- a. 通过Kafka将实时数据流（T+0）引入实时数仓Kudu中，并通过Impala提供数据服务；
- b. 通过HDFS、Impala、HBase处理后的数仓数据（T+1）引入实时数仓Kudu，并通过Impala处理提供数据服务



结语

建立符合生态发展需求并支持数字化业务的新型数据架构，有助于夯实数据治理的落地与优化，提升数据治理工作的规范化 and 高效化，确保数据统一管理和高效运行，使数据价值转换成核心竞争力。在数据战略已确定、组织架构已搭建、各项数据标准、指标已规范梳理后，IT领域的数字治理工作如何配套开展，数据治理体系如何落实和融入到信息系统建设的方方面面，将是信息科技部门的人员所最为关注的。继数据架构之后，在后续文章中，我们将对数据管控平台相关实践进行阐述分享。