
GOTC 2023
全球开源技术峰会
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE
OPEN SOURCE, INTO THE FUTURE
「基础设施与软件架构」专场
从ESB, Kafka, 到DaaS
实时数据集成的技术变迁
唐建法 2023年05月28日

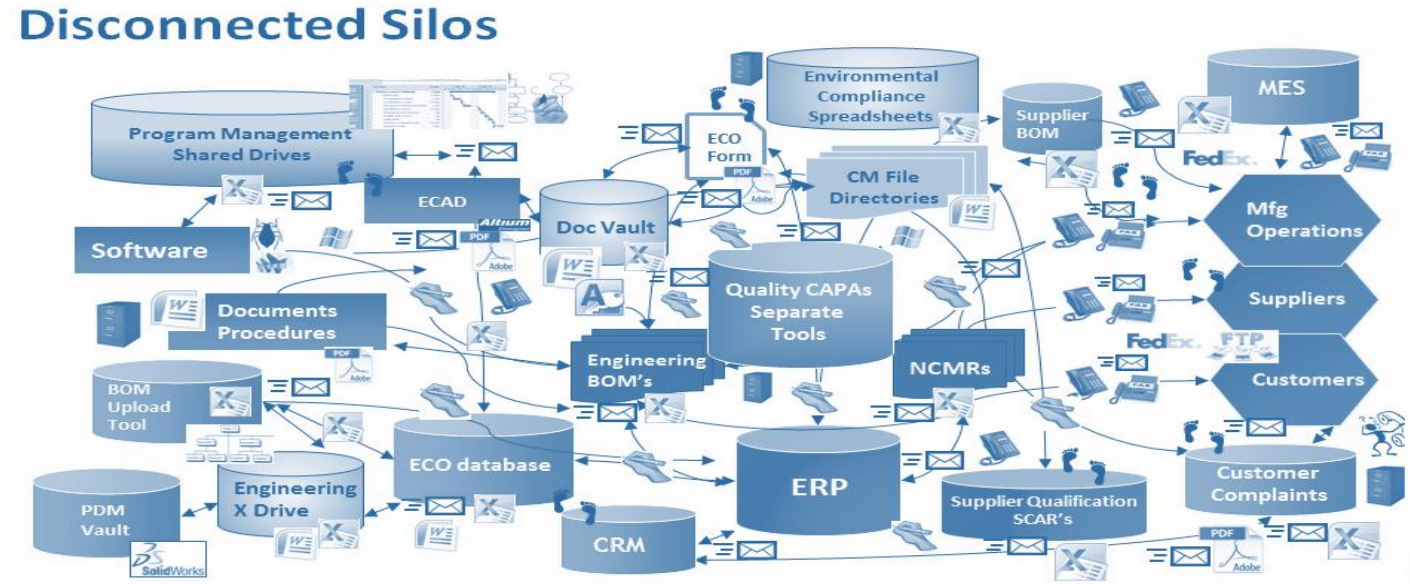
关于讲师

- TJ / 唐建法
- Tapdata 创始人
- 前MongoDB 大中华区技术总监， MonogDB 中文社区主席
- 4 年摇滚键盘手, 2 年全球背包客
- 10 年数据架构师, 3年产品创业人
- 爱好：风筝冲浪、公路车、家庭乐队、房车旅行

01

数据集成的技术变迁

企业的现状：30年的信息化形成的数据孤岛



30年

企业信息化建设

463套

大型企业拥有的系统数量

Q: 如何高效支持：

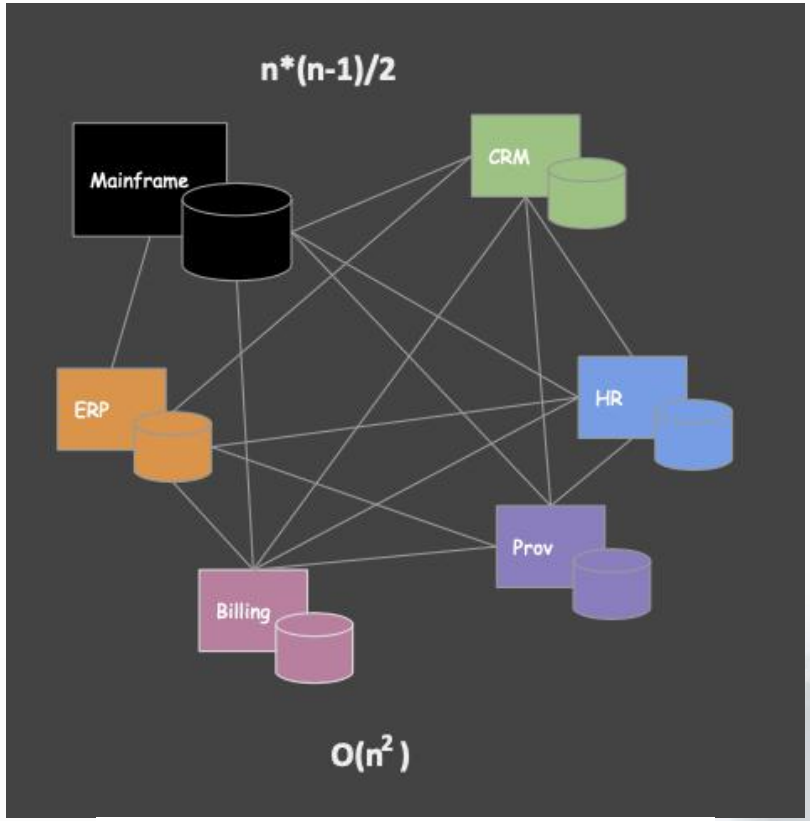
最自然的数据集成架构: Point–to–Point

Point–to–Point

- 数据消费者向数据提供者直接获取
- 通过API 或者数据抽取方式

The Problems

- 意大利面一样错综复杂，难以管理
- 难以扩展，支持多个业务系统
- 高耦合
- 重复劳动



Point to Point

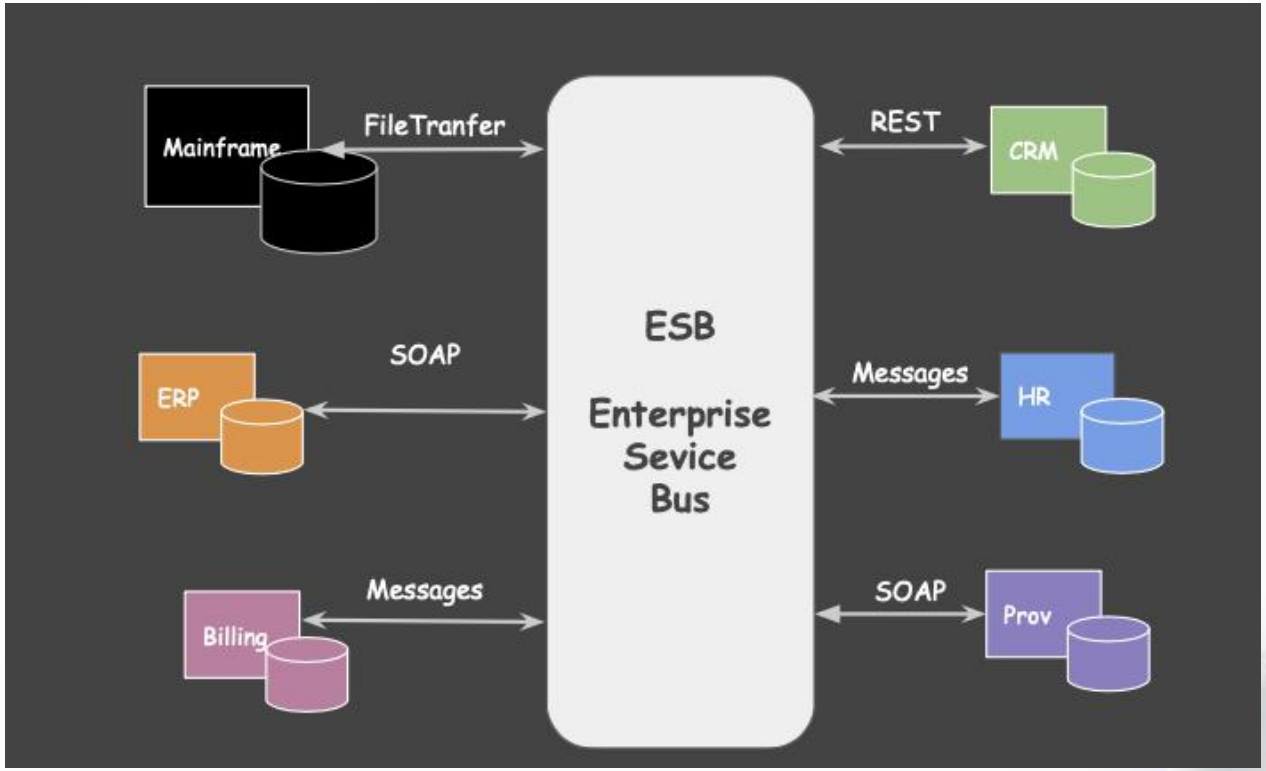
EAI / ESB:中央化的集成架构

EAI / ESB

- EAI: 通过中央模块交换数据
- ESB: 最主流的EAI 架构实现

特点和优势

- 中央化架构，提高可管理性
- 低耦合
- 可扩展性 — 增加新的系统对接
- 减少点到点的数据集成数量



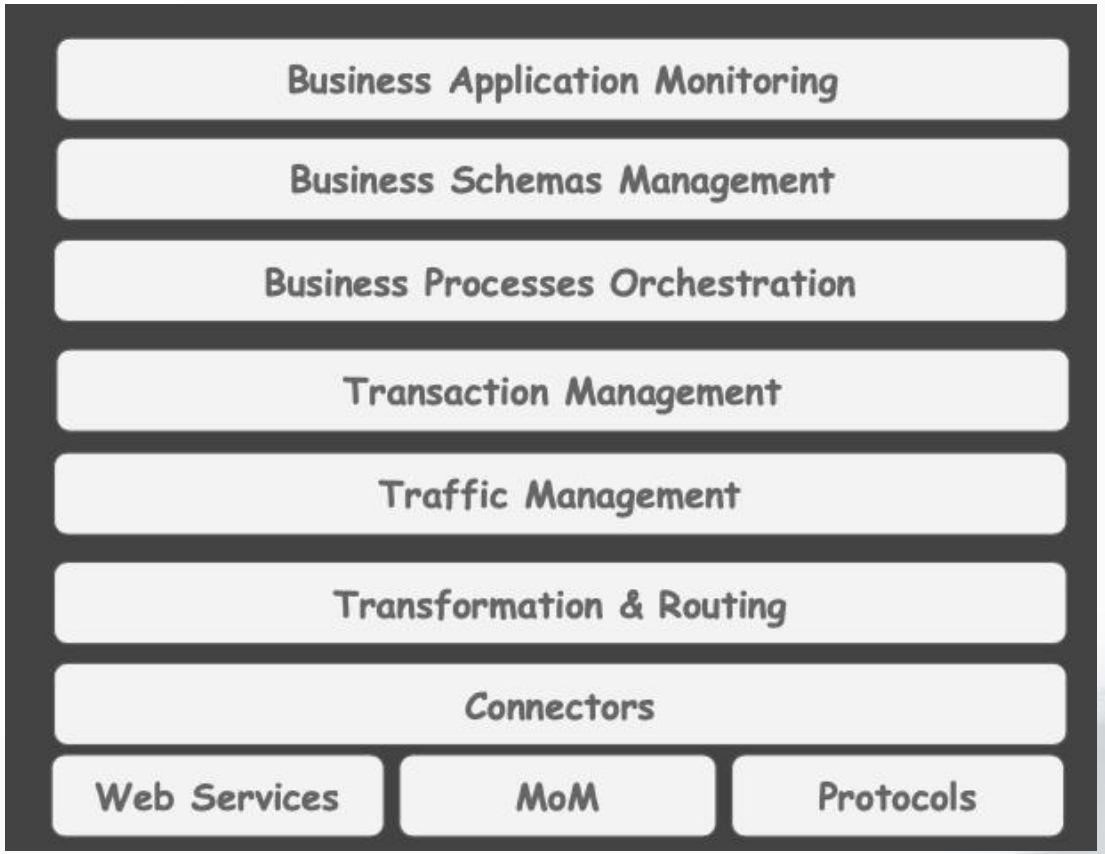
EAI / ESB

ESB: 正在被遗忘

ESB 方案的缺点

- 架构清晰，实现太复杂
- 基于SOAP的协议逻辑耦合太强
- XML ： 定义繁复，处理困难

结果： 昂贵的开发实施和维护成本



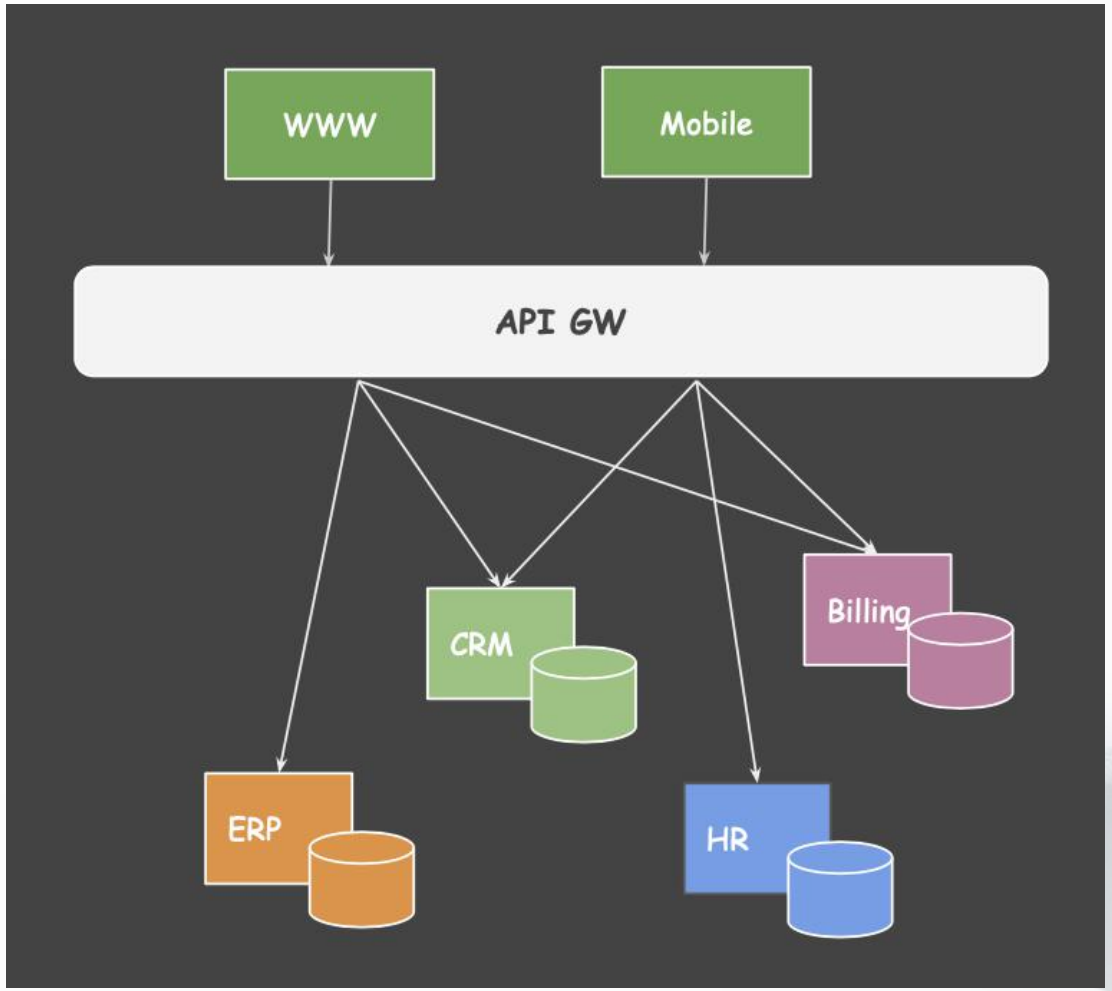
API Centric 的方案

API Gateway

- 基于RESTful / JSON
- 易读，易用
- 容易开发

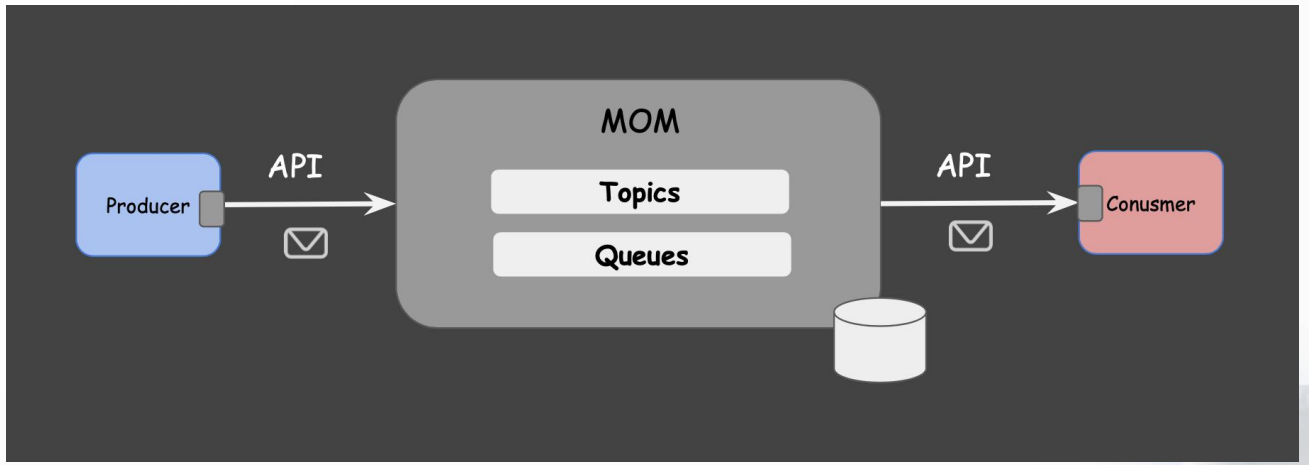
The Limitations

- 能力受限于已有系统的API设计
- 对源库有性能影响，特别是未优化SQL



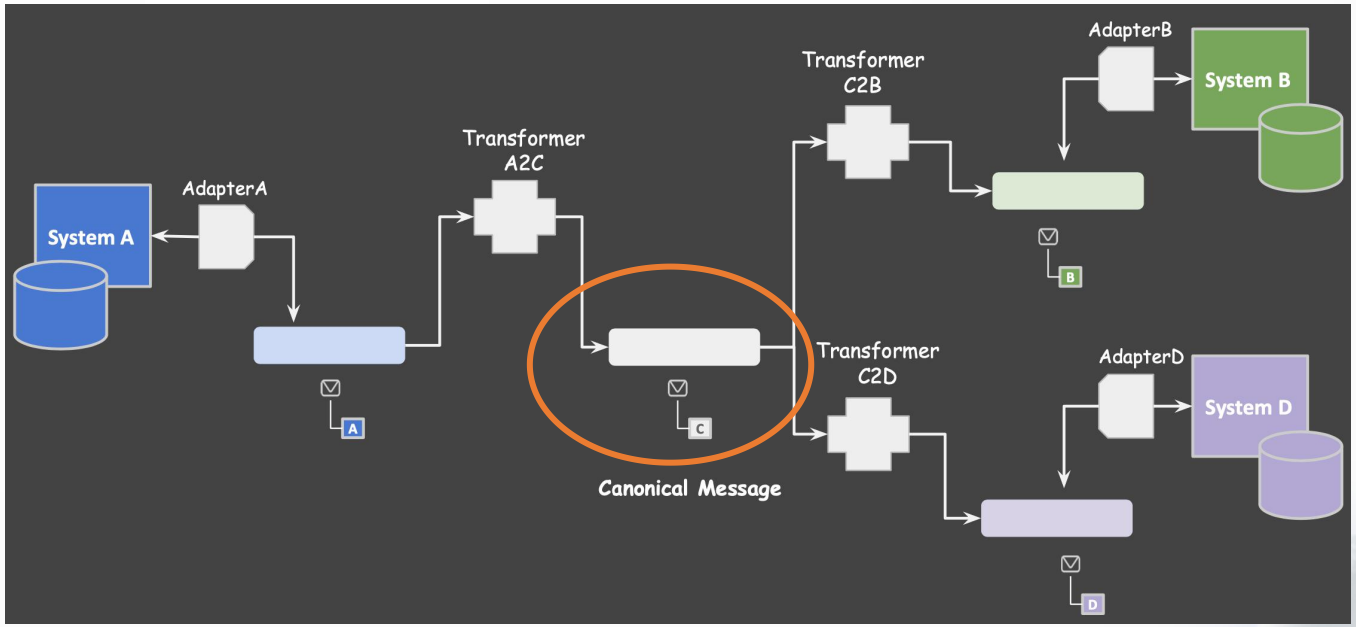
消息中间件架构： Message Oriented Middleware

- 类似ESB一样有中间件，可以
- 解耦架构 — 两端不直接相连
- 异步处理 — 提升性能
- 源和目标使用单独的Adapter，集成灵活

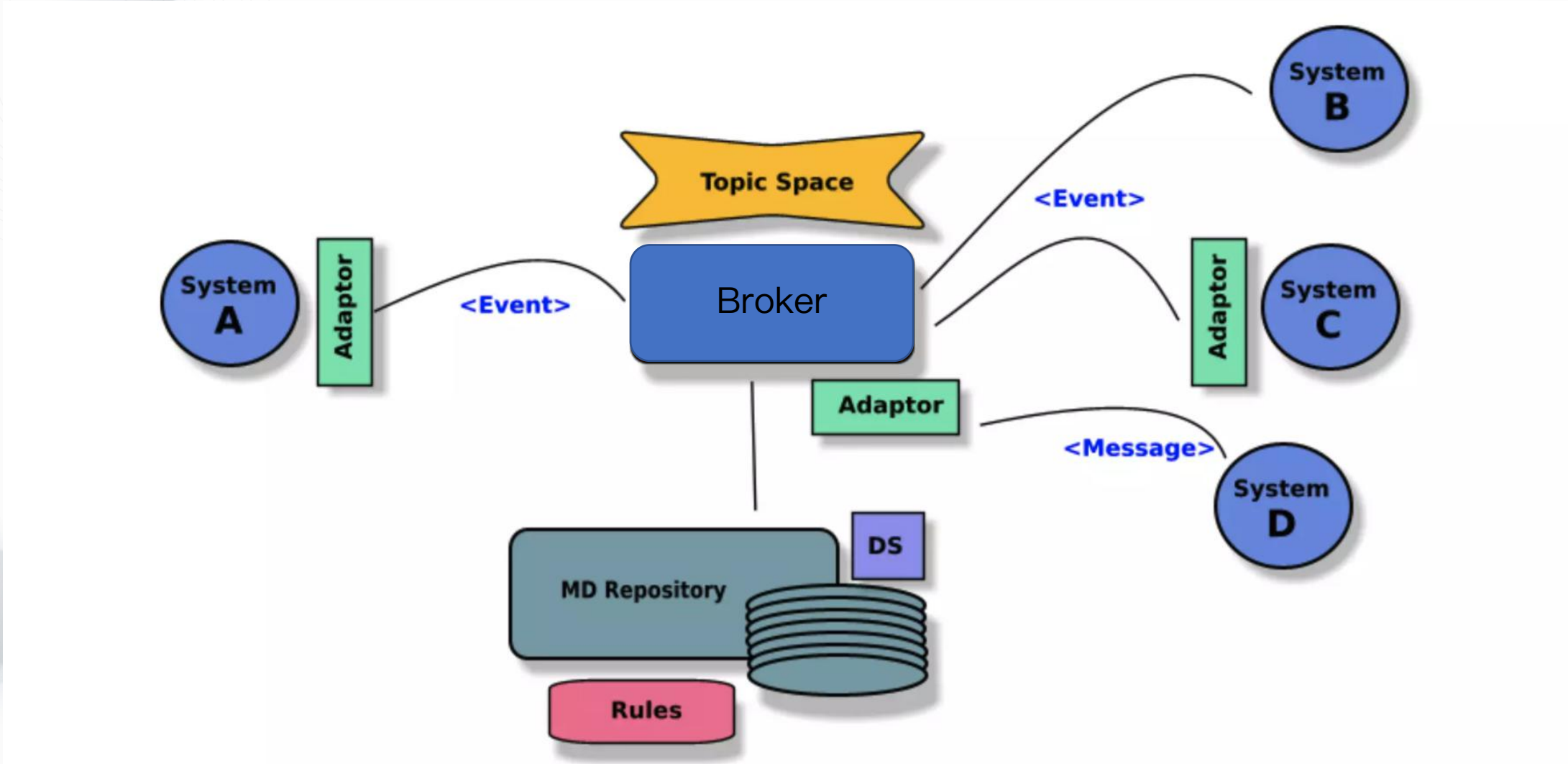


如何解耦： 使用标准的Message 格式

- 上游数据从 A 到 C （中间格式事件）
- 下游从C 到B或者C到D
- 系统之间不直接产生关系



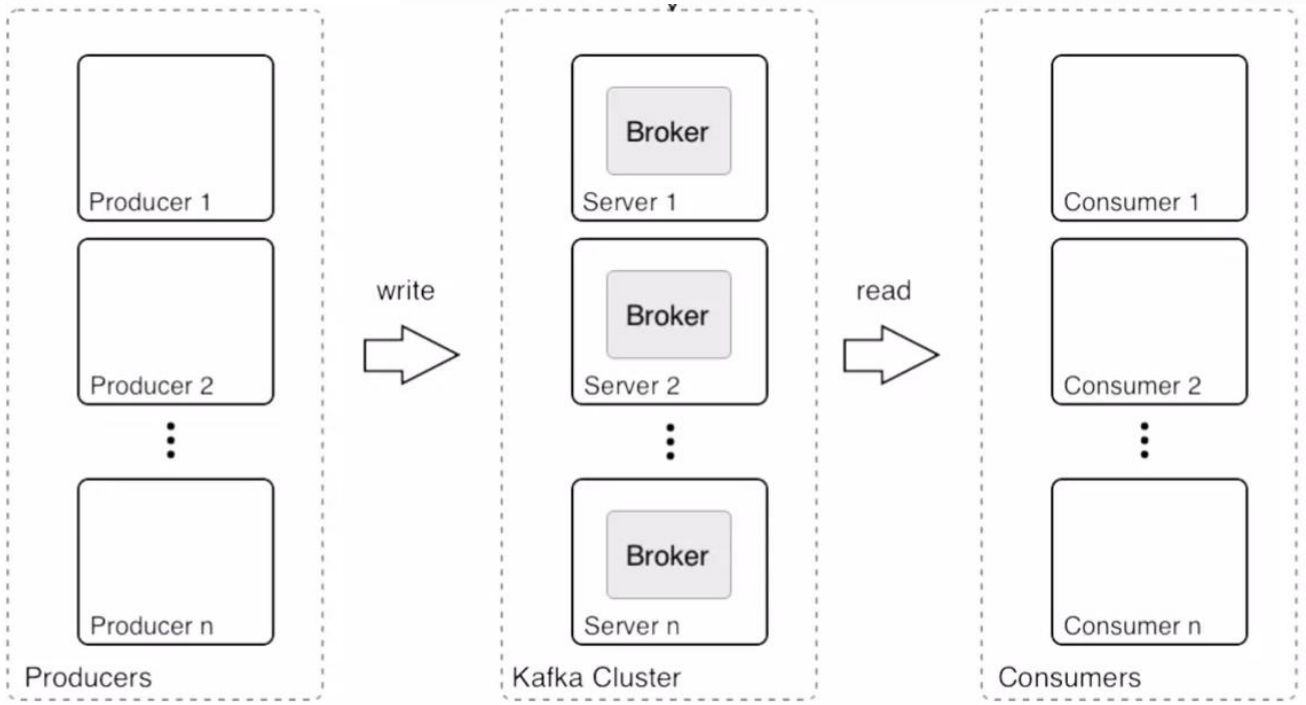
Event Driven Architecture



Kafka：最主流的MOM 实现之一

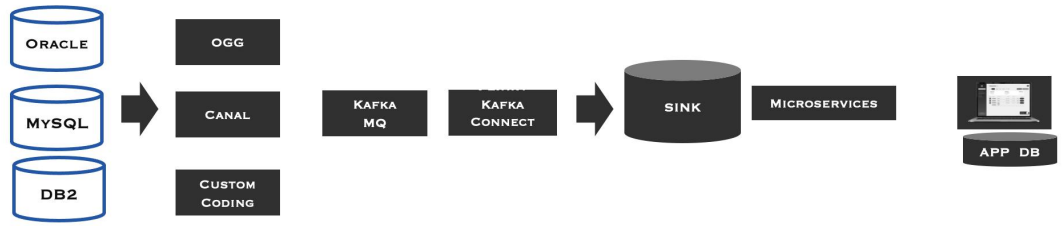
优势

- 解耦式的数据集成架构
- 分布式，高性能，可达百万QPS /秒
- 事件存储，支持重放
- 完善的开源生态：几十种语言支持



Kafka 作为数据集成方案的局限性

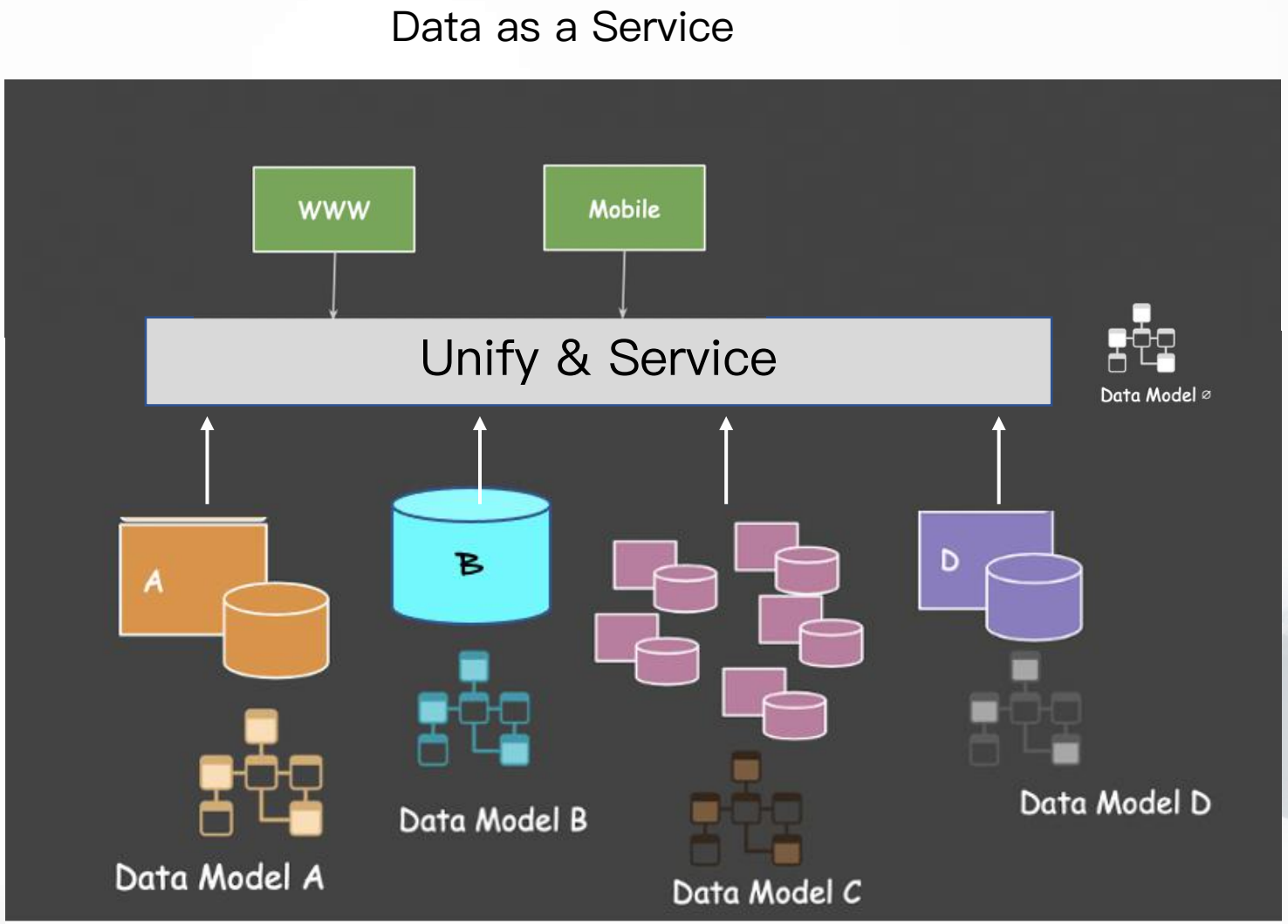
- 生产端需要解决Message 入队列的开发
- 链路较长，节点较多，对实时集成的时延较高
- 复用的是Message，消费端需要对Message 进行较多的开发来还原数据模型



- 客户端 对Offset 的管理
- 基于Message 对增删改操作的封装
- Exactly Once 的保障需要客户端配合

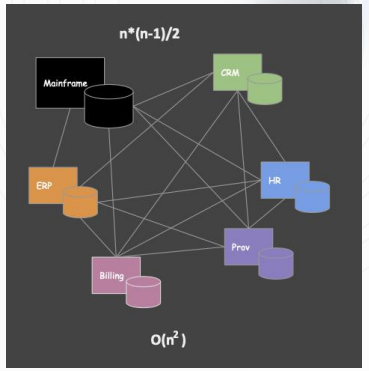
能否直接复用数据模型，而非消息？

- 将不同库的物理模型统一成一个一致的逻辑模型
- 在中间件提供统一的数据模型，而非Message
- 通过HTTP API + JSON 直接服务下游



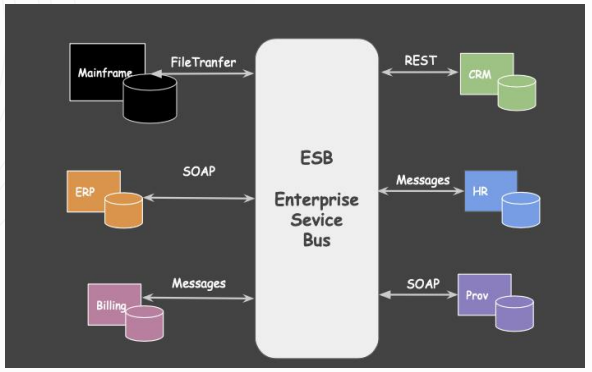
数据集成架构的优劣势

点到点集成



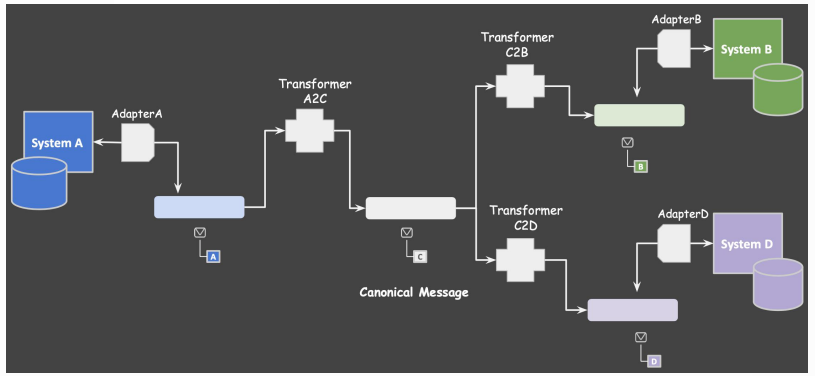
- 简单直接
- 强耦合
- 不易扩展
- 无复用

ESB 企业总线



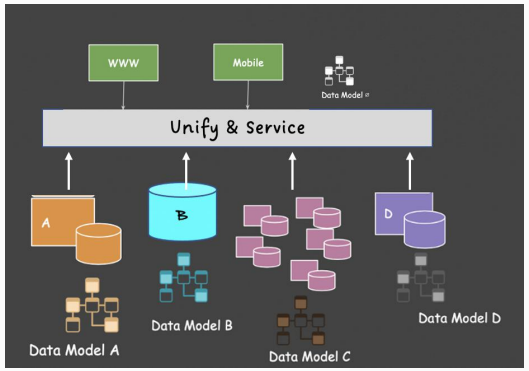
- 松藕
- 业务中央化
- 性能低，数十QPS
- 开发繁杂
- 复用性低

消息队列



- 松藕
- 消息中央化
- 异步处理，并发高性能
- 数据延迟不保证
- 开发复杂

数据即服务

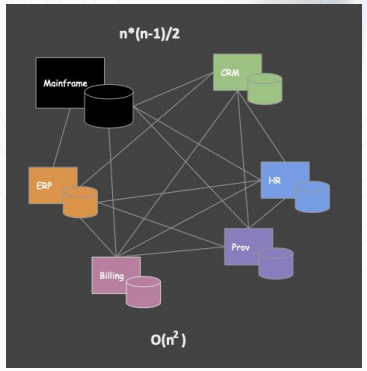


- 松藕
- 数据中央化
- 直接复用数据
- 需要额外存储
- 需要同步链路支撑

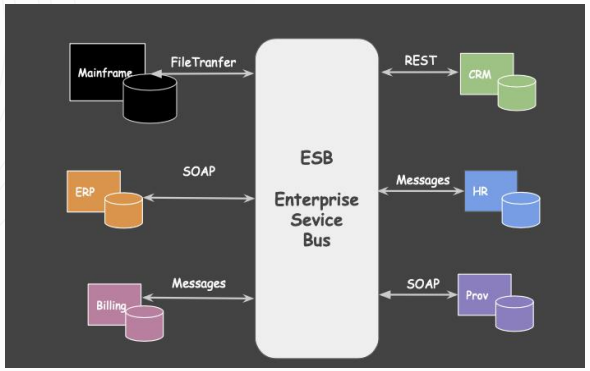
数据集成的代表产品

GOTC

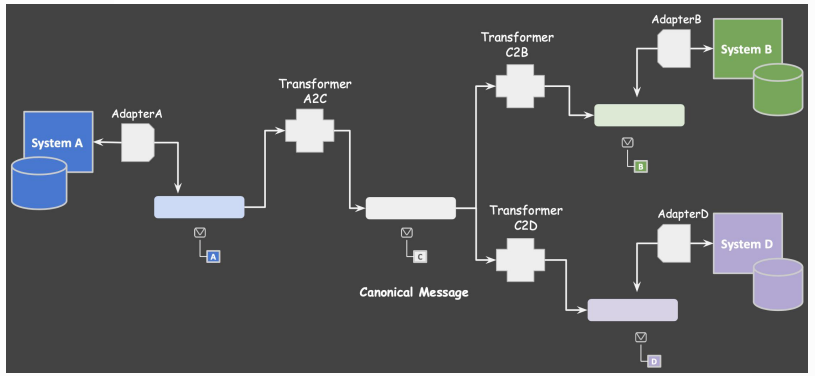
点到点集成



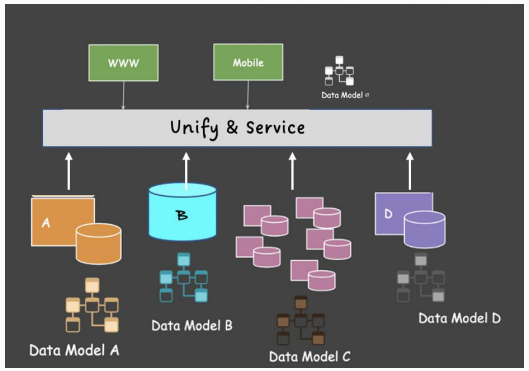
ESB 企业总线



消息队列



数据即服务



Informatica

MuleSoft

kafka

TIBCO®

pentaho

IBM WebSphere software

APACHE ACTIVEMQ®

TAPDATA

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

02

Tapdata 在实时 DaaS 数据架构上的一些实践

Tapdata： 首个基于DaaS 架构打造的开源实时数据服务平台

2019.09
成立于深圳

2019.10
千万人民币天使轮

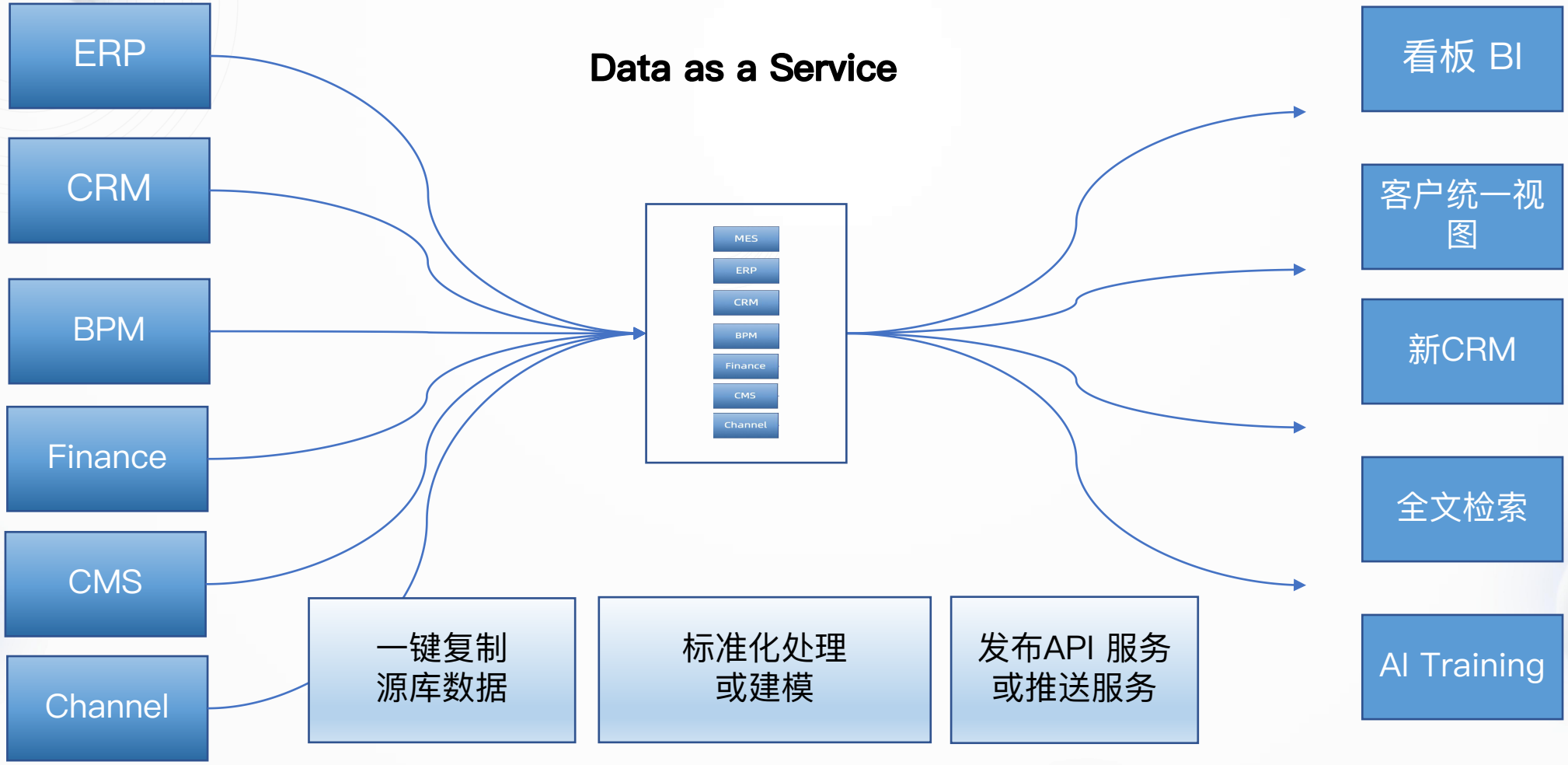
2020.09
Tapdata 1.0 GA

2021.09
数千万美元PreA轮融资

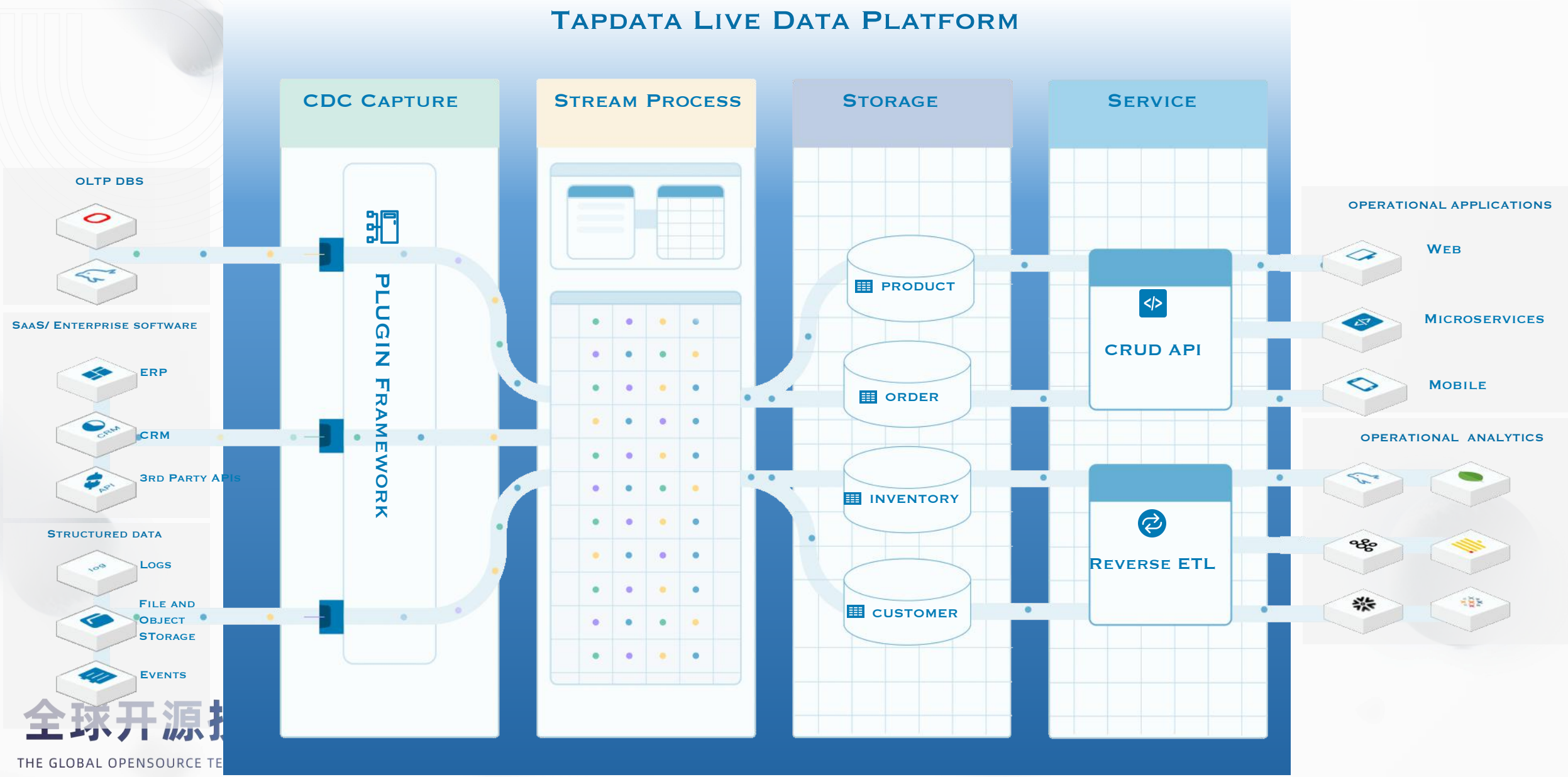
2022.07
Tapdata 2.0 GA
Tapdata 开源版发布
Tapdata 云版上线

2023.05
10，000+ 企业和云版客户

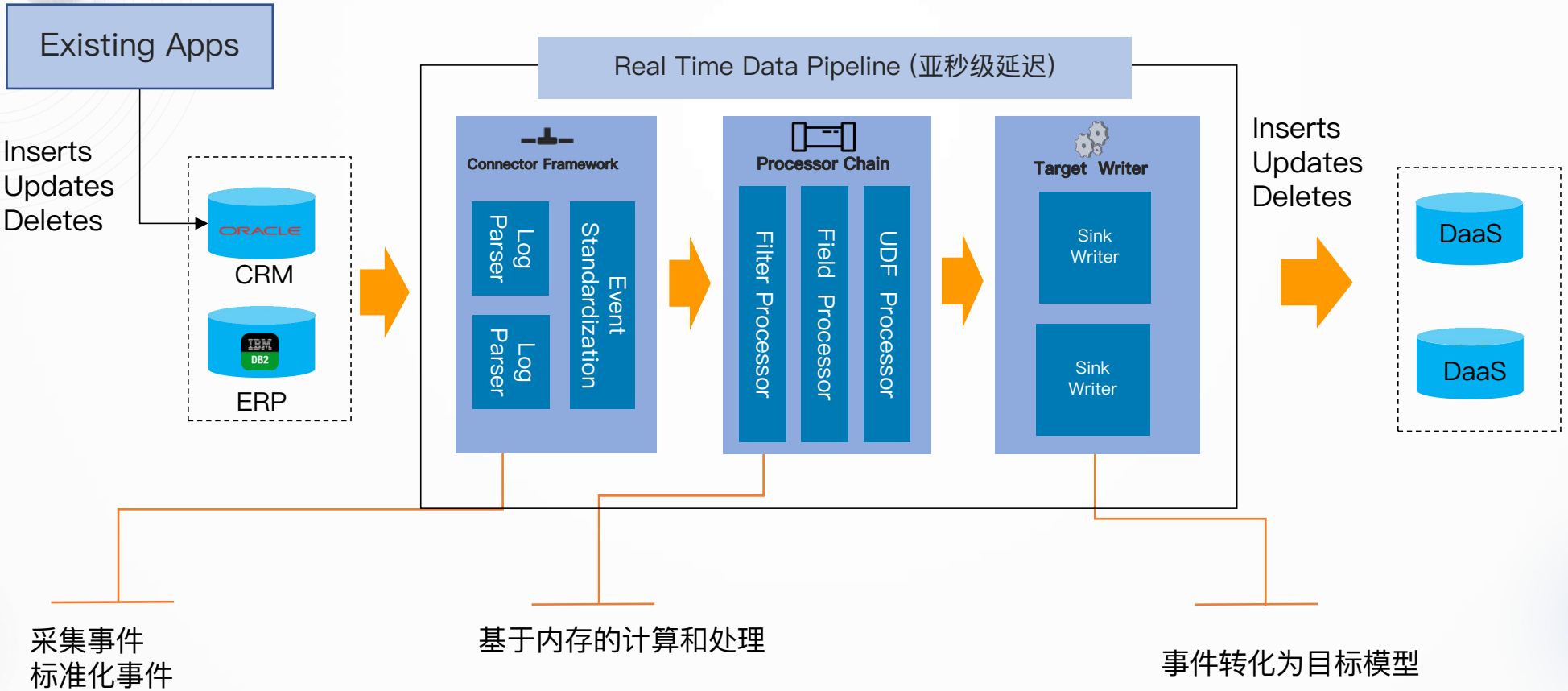
一次复制到中央化平台，通过DaaS 支持多个业务场景



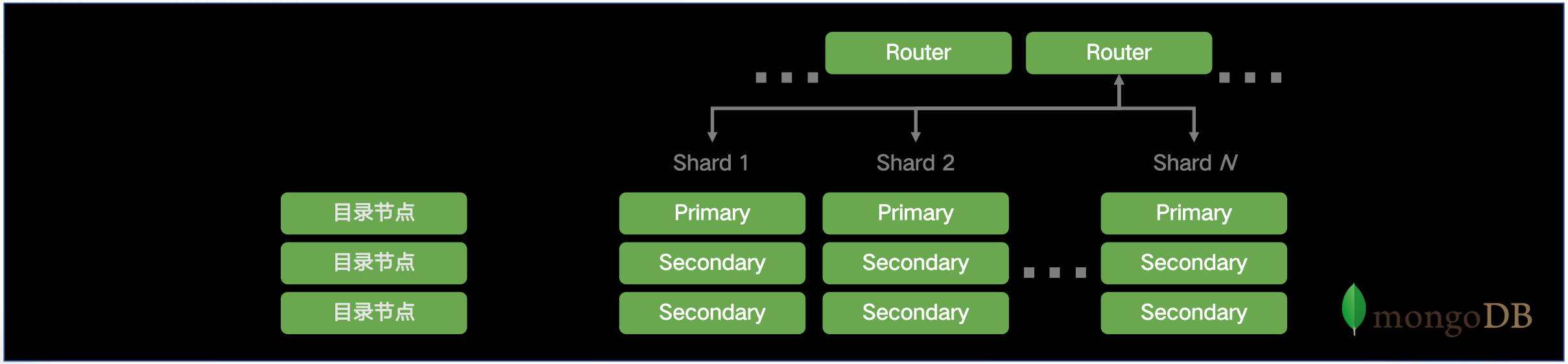
Tapdata Live Data Platform: 第一个实现 DaaS 架构的实时数据平台



实时DaaS之底层技术：基于CDC的数据采集与同步



基于分布式MongoDB 数据库的中央化存储



横向扩展能力

- TB – PB数据量支持
- 跨中心跨云部署能力

支持多种数据模型

- 模型变动灵活，易融合
- 结构化，半结构化

高性能高并发

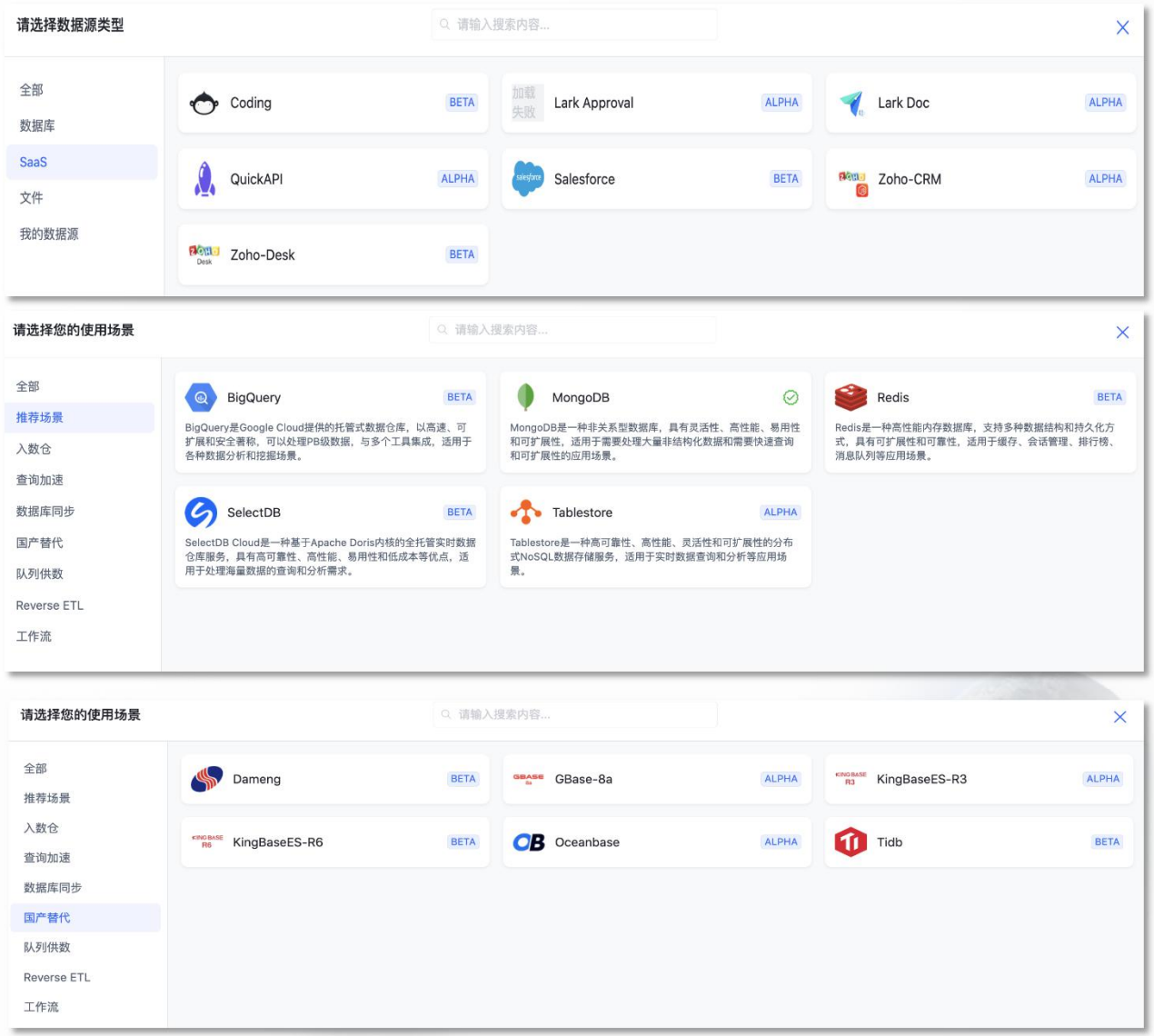
- 毫秒级响应– 比Hive快百倍
- 高并发促销场景

多工况支持

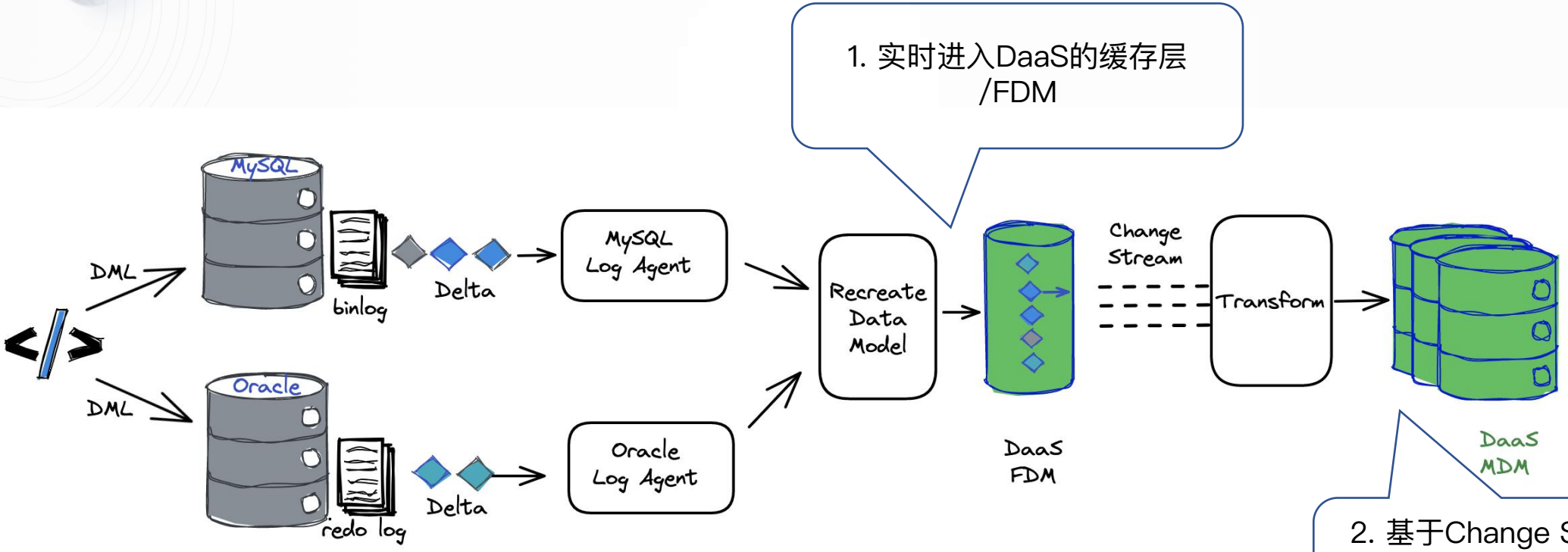
- OLTP – 即时更新
- OLAP：聚合运算

60+ 数据源， 40+ 数据目标场景支持

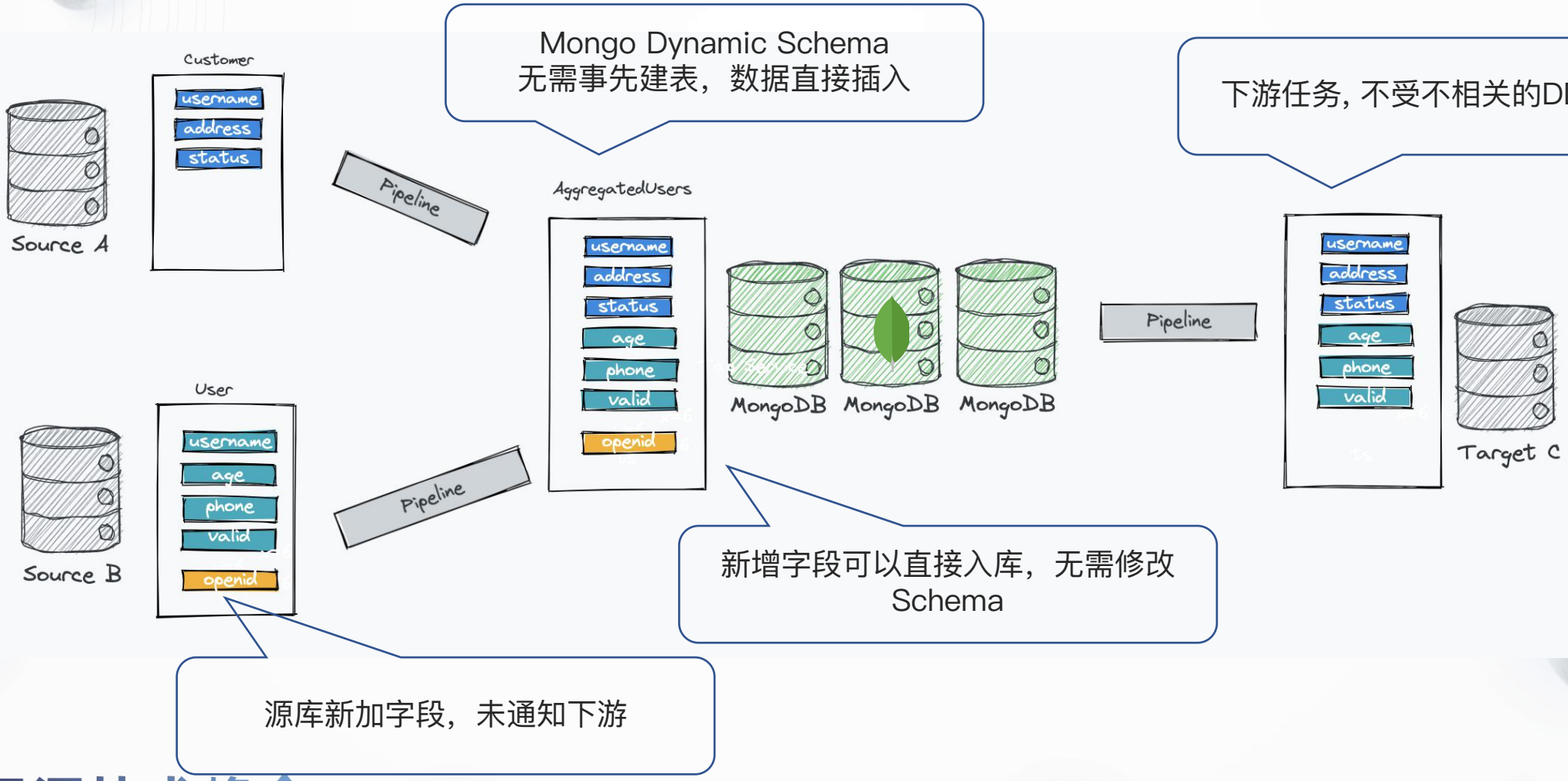
- 常见RDBMS 和NoSQL 支持，包括Oracle, DB2, SQLServer, MySQL, PG，MongoDB， Reids, ES 等
- 支持SaaS 数据源
  - 一键对接Salesforce, Zoho， 销售易等CRM，实现客户360
  - 一键对接 Amazon, Shopify 等电商，实现商品库存打通
- 云数仓 — 搭建你的现代数据栈
  - BigQuery
  - Tablestore
  - SelectDB
- 国产信创数据库对接
  - 快速迁移到达梦，南大通用，人大金仓，Oceabase等国产数据库



在DaaS 里，如何实现分层建模并且依然能够提供实时保障



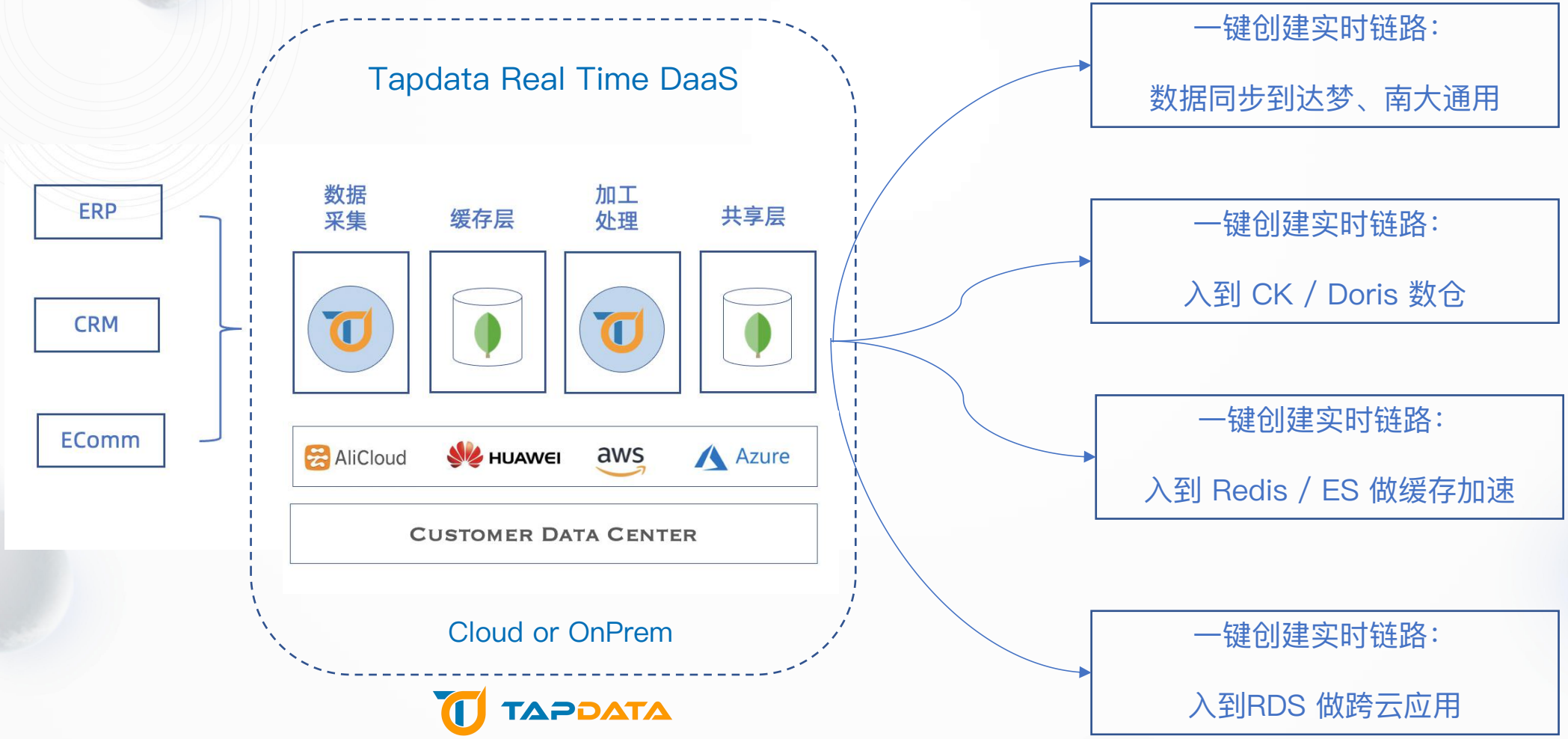
如何解决DDL对任务的影响 及多源异构数据合并的复杂性?



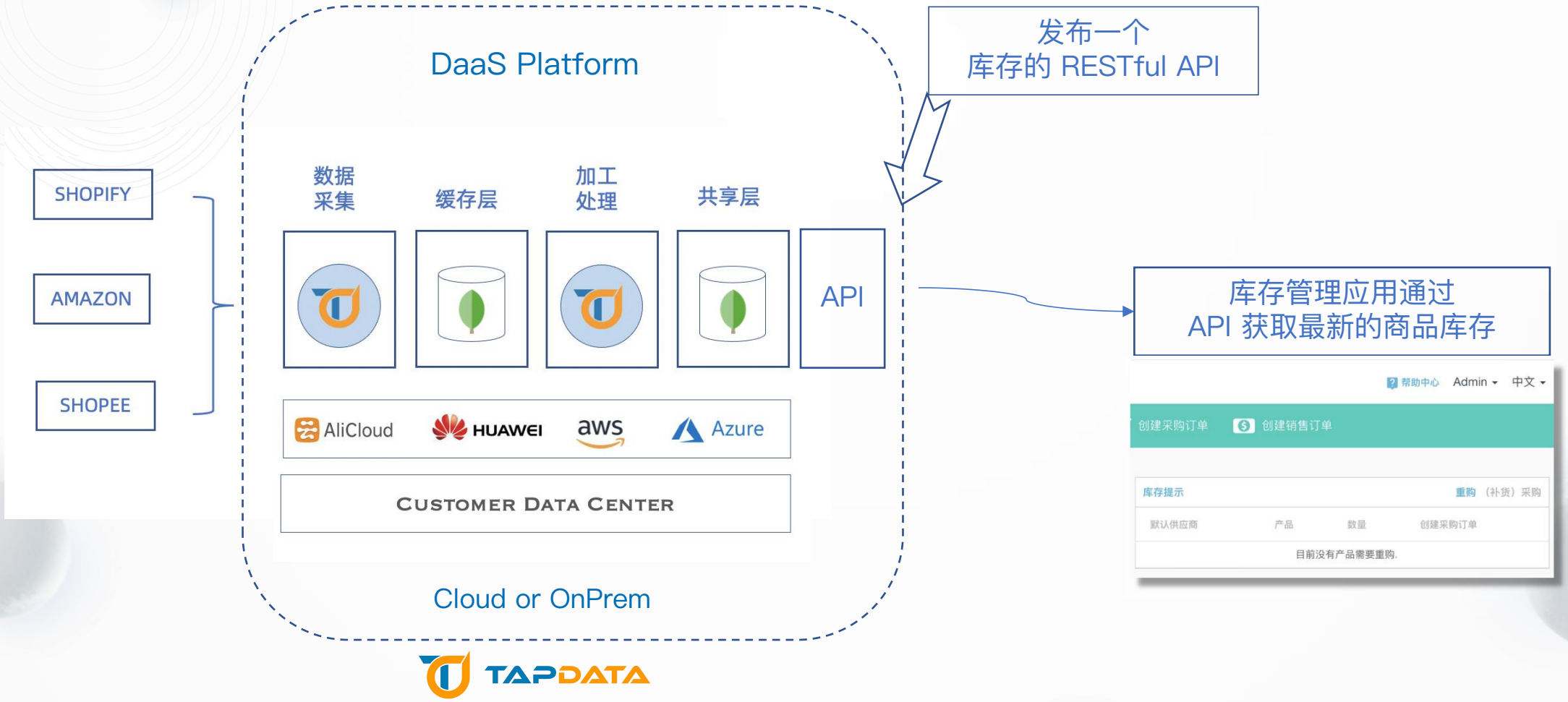
03

如何应用 DaaS

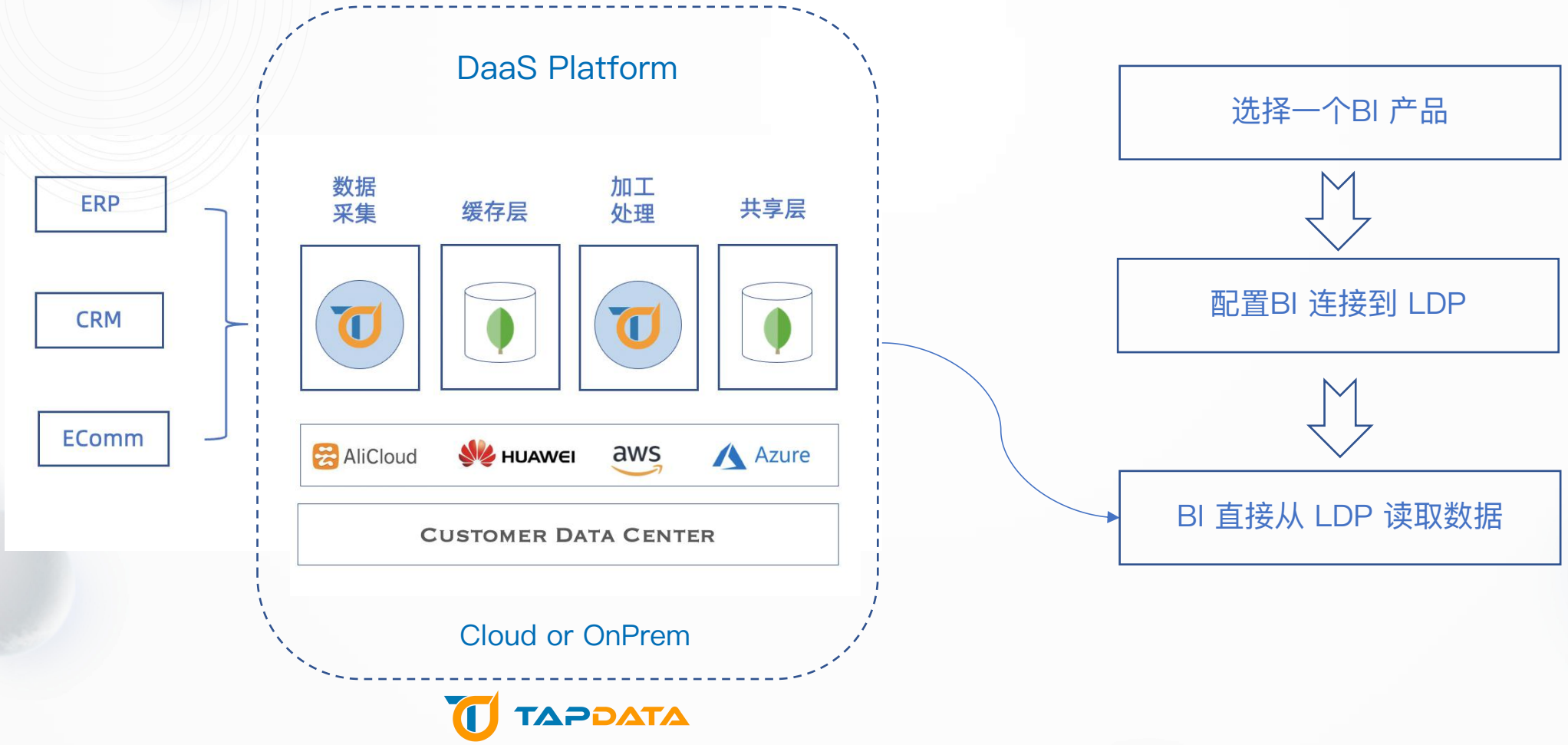
Real Time DaaS 业务场景一： 实时数据同步



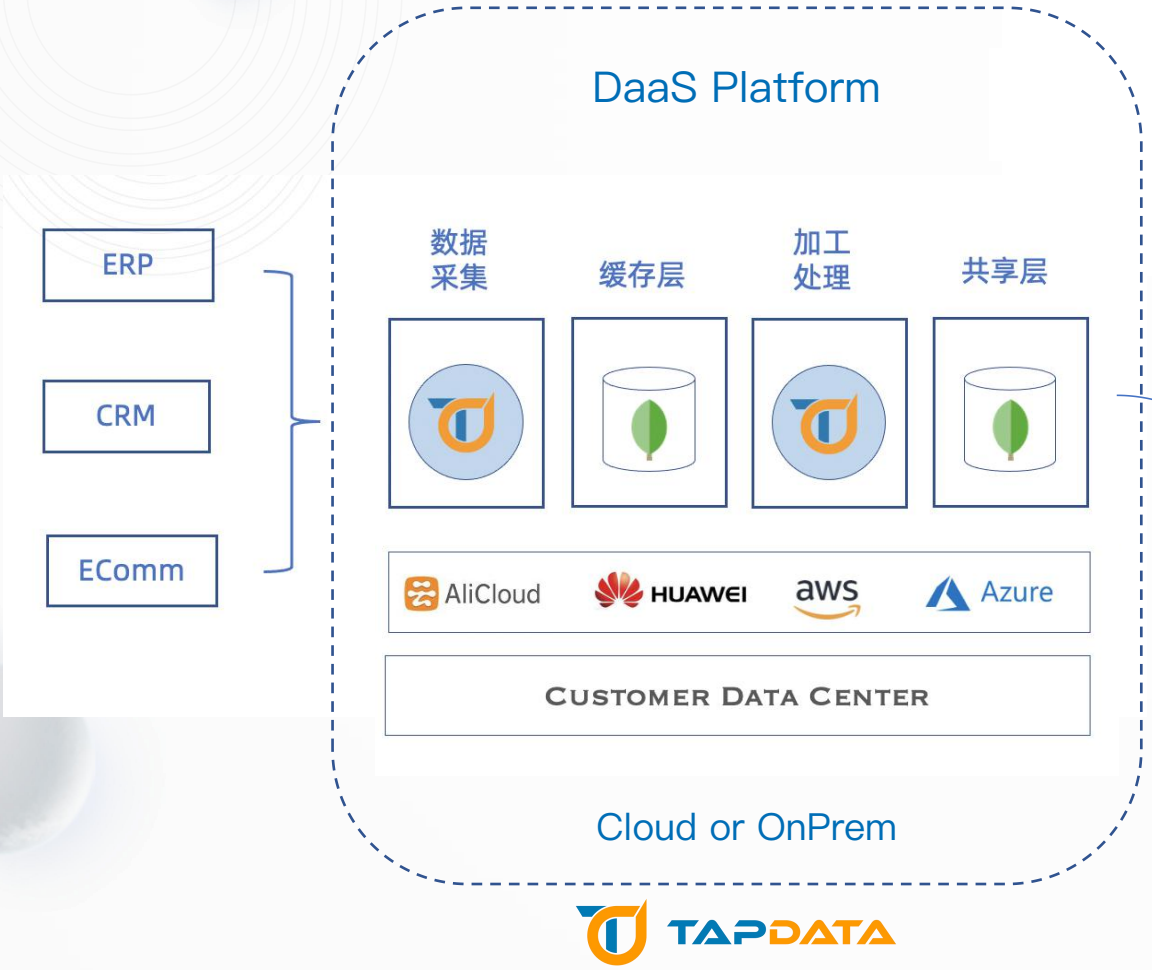
Real Time DaaS 业务场景二： 搭建客户、商品、库存360视图



Real Time DaaS 业务场景三： **快速搭建 BI 看板**



Real Time DaaS 业务场景三： 为私域大模型算法提供企业私有数据

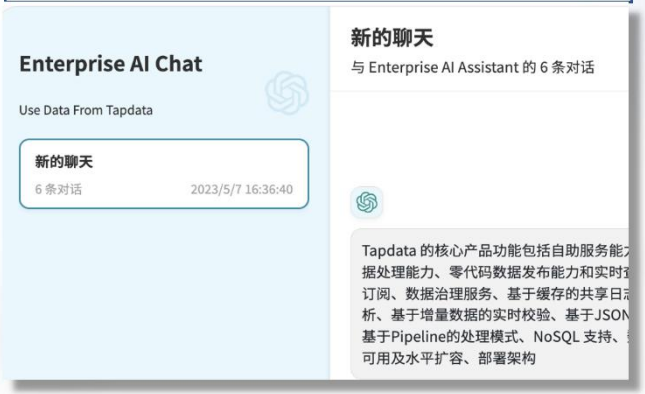


部署企业私有大模型算法平台

Tapdata 内一键创建实时链路：

数据同步到AI 平台

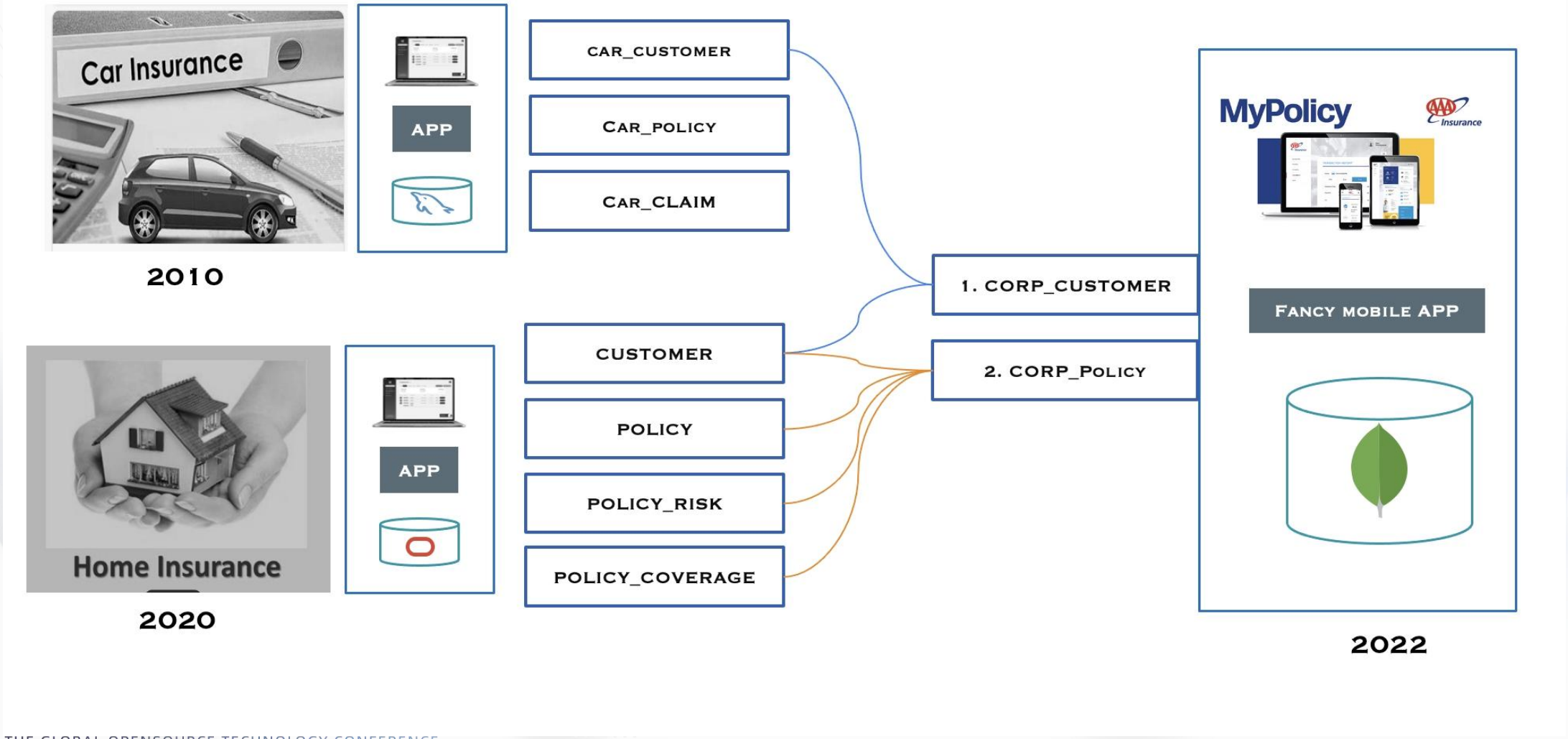
AI 提供更为实时、更为准确的答案

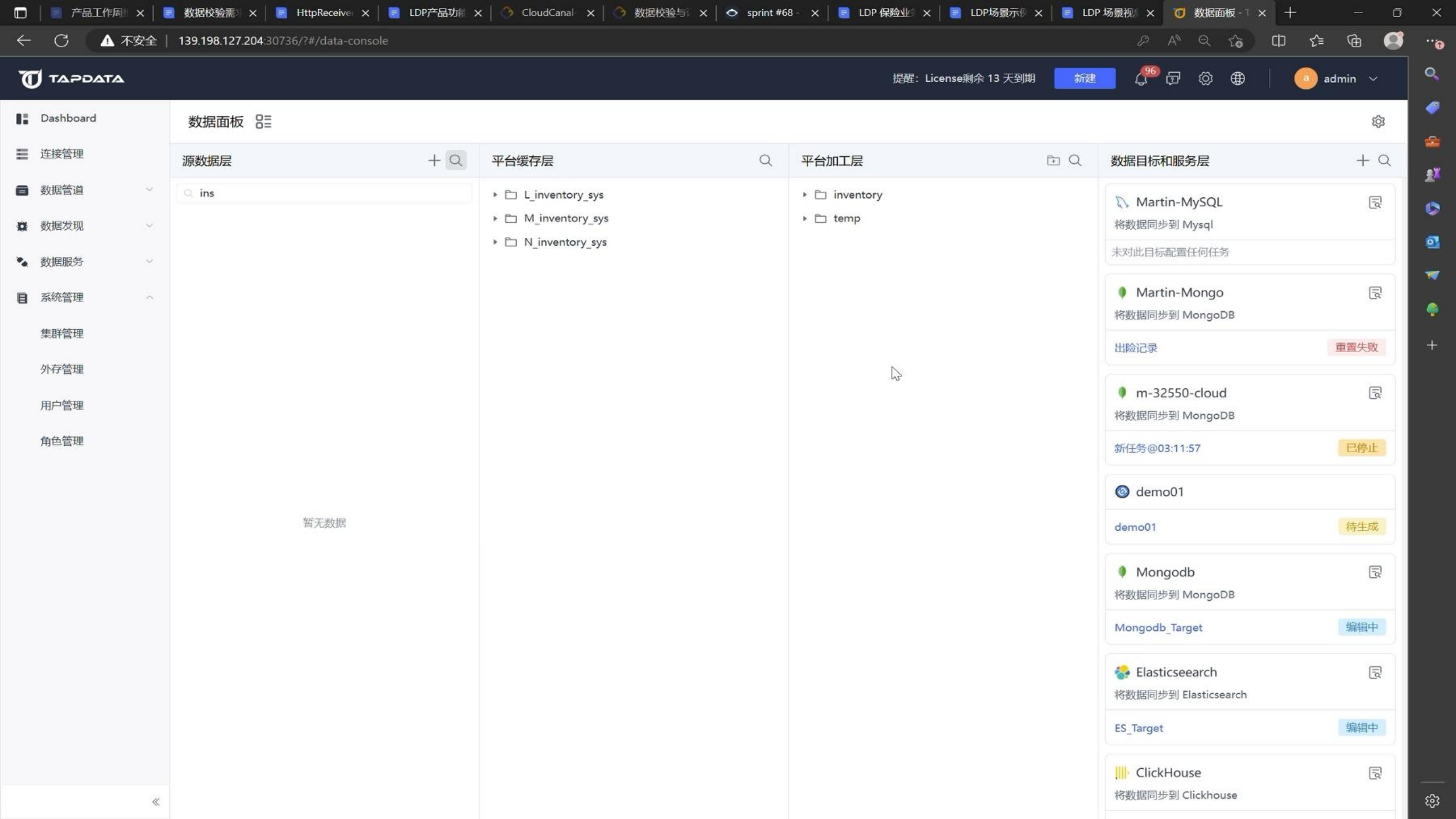


04

DaaS DEMO

Demo 场景：保险业统一客户视图




产品工作周
数据校验需
HttpReceiver
LDP产品功能
CloudCanal
数据校验与
sprint #68
LDP 保险业
LDP场景示例
LDP 场景视
数据面板 - T
不安全 | 139.198.127.204:30736/?#/data-console
TAPDATA
提醒：License剩余 13 天到期
新建
96
admin
Dashboard
连接管理
数据管道
数据发现
数据服务
系统管理
集群管理
外存管理
用户管理
角色管理
数据面板
源数据层
ins
暂无数据
平台缓存层
L_inventory_sys
M_inventory_sys
N_inventory_sys
平台加工层
inventory
temp
数据目标和服务层
Martin-MySQL
将数据同步到 Mysql
未对此目标配置任何任务
Martin-Mongo
将数据同步到 MongoDB
出险记录
重置失败
m-32550-cloud
将数据同步到 MongoDB
新任务@03:11:57
已停止
demo01
demo01
待生成
Mongodb
将数据同步到 MongoDB
Mongodb_Target
编辑中
Elasticseearch
将数据同步到 Elasticsearch
ES_Target
编辑中
ClickHouse
将数据同步到 Clickhouse

多种方式开启 DaaS 之路

云版

开箱即用

企业版

安全部署在你的
数据中心

社区版

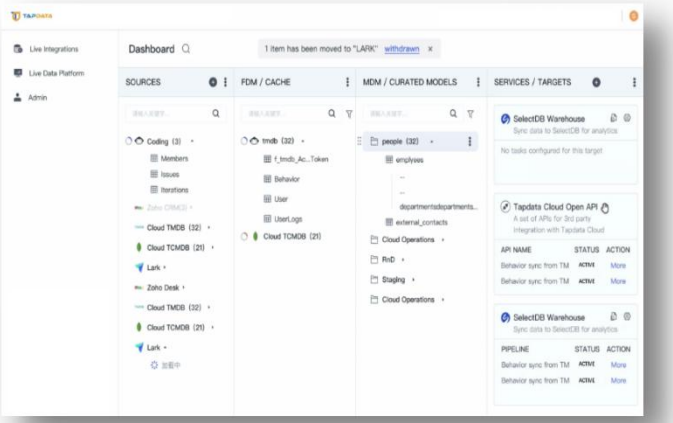
免费使用绝大部分
产品功能

最简单的方式：免费注册Tapdata Cloud 账号，马上开始

■ 免费注册

■ **选择托管模式**

- 半托管：更安全，数据不会离开你的VPC或者数据中心
- 全托管：免安装免运维，开箱即用



■ **开始免费使用**

GOTC

THANKS

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE