腾讯云 | DataFun.

云上数据集成的挑战和实践

腾讯云 高级工程师

目录
CONTENT

01 云上数据集成的挑战
企业核心诉求梳理

02 云上数据集成平台设计
方案和平台分析

03 云上数据集成产品落地
DataInlong产品简介

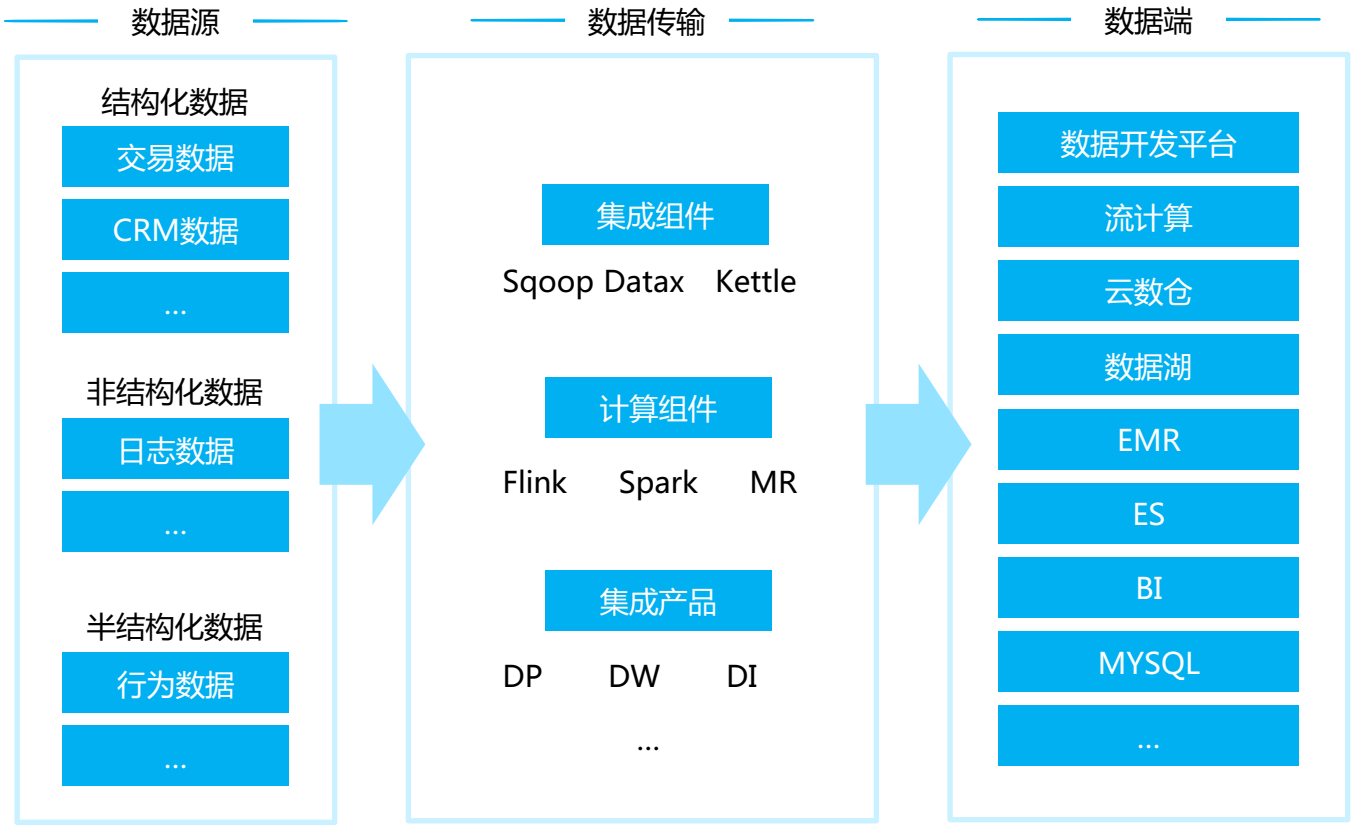
04 云上数据集成的实践
客户经典案例介绍

腾讯云 | DataFun.

01 云上数据集成的挑战

企业核心诉求梳理

数据集成业务概述



数据集成常用方案介绍

引擎	数据源种类	吞吐	稳定性	开源社区	场景
Spark	基于API扩展	高	资源依赖复杂，环境适应性差	顶级社区	大批量 实时同步 数据转换
Flink	社区有部分，支持行扩展	高	环境适应性、数据源亲和性差，跨集群访问配置复杂	顶级社区	大批量 实时同步 数据转换
Sqoop	支持的数据源种类偏少	高	基于MR框架，运行在hadoop集群，比较稳定	顶级社区	大批量 实时同步 数据转换

Flink： 部分场景吞吐优势，时延低，周期性数仓等场景下灵活性不足、资源利用率低。
Sqoop：支持数据源种类少，基于MR框架，稳定性高。
Spark：吞吐极高，参数配置复杂，稳定性差。

企业对数据集成的诉求

场景化

大数据上云
湖/仓搭建
线上服务系统

海量化

TB级每小时
百万条每分钟
天级别持续运行

多样化

数据源种类多
单个链路的数据源多
数据源的数据格式多

服务稳定

容错性强
准确率高
故障自动恢复

运维简单

屏蔽底层资源
可拖拽配置
数据传输可视化

快速搭建

开箱即用
插件灵活扩展
无缝衔接生态

云上数据集成面临的挑战

技术支撑

- 百万亿级数据传输能力
- 轻量+海量（标准）架构，队列组件灵活插拔
- 毫秒级数据时延
- 自带极致性能、高稳定数据队列
- 全链路数据审计
- 资源弹性伸缩能力

场景全面

- 30余种异构数据同步
- 支持离线、实时场景数据ETL
- 支持主动上报、日志变更拉取等数据采集方式
- 全、增量、全增量融同步策略

开箱即用

- 数据链路可视化拖拽配置
- 全链路数据指标图形化展示
- 支持按业务所需定义数据链路与字段映射
- 资源全托管免运维

生态一体

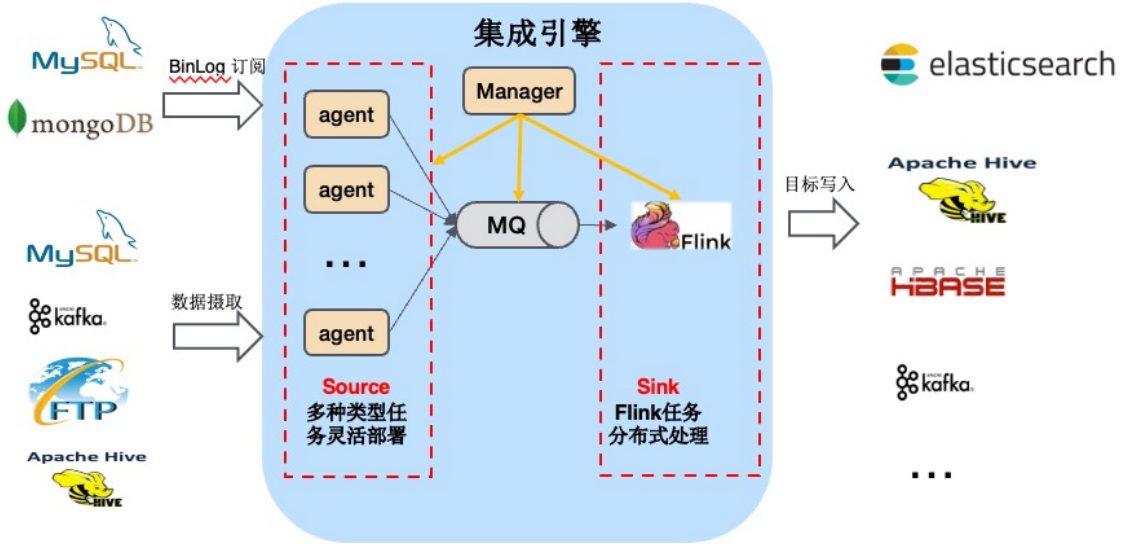
- 支持云上工具产品
- 支持云原生大数据平台
- 支持云原生数据湖
- 支持云原生数仓
- 支持云数据库、云中间件等

腾讯云 | DataFun.

02 云上数据集成平台设计

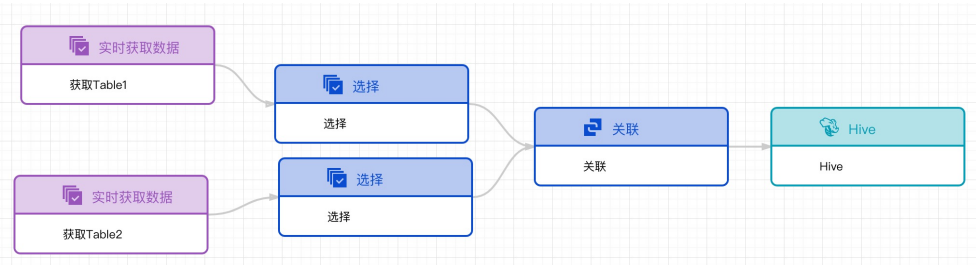
方案和平台实现

全场景数据集成解决方案设计



关键设计

- 多Agent支持
- 批流一体，配置统一
- 读写端解耦



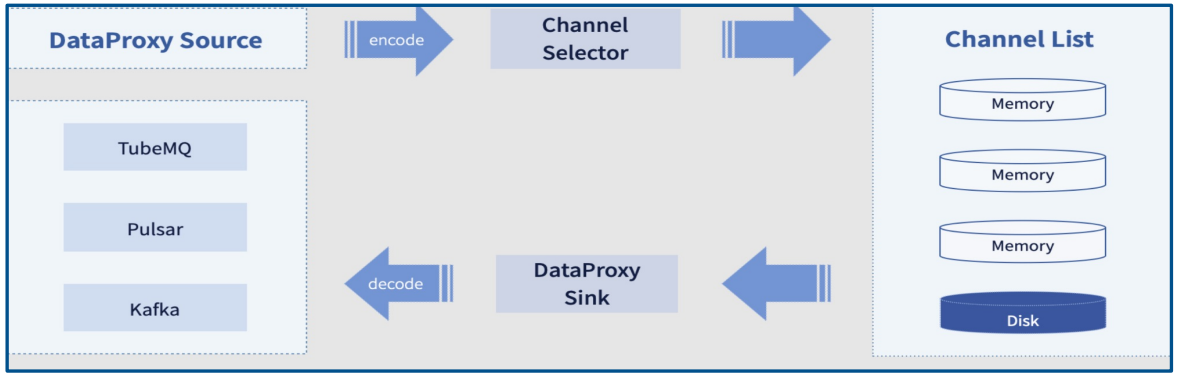
业界类似的解决方案：

AWS ： kinesis+firehouse（多产品配合）
华为 ： DIS（配套解决方案）
阿里 ： DataHub（配套解决方案）

高性能低延迟队列方案分析

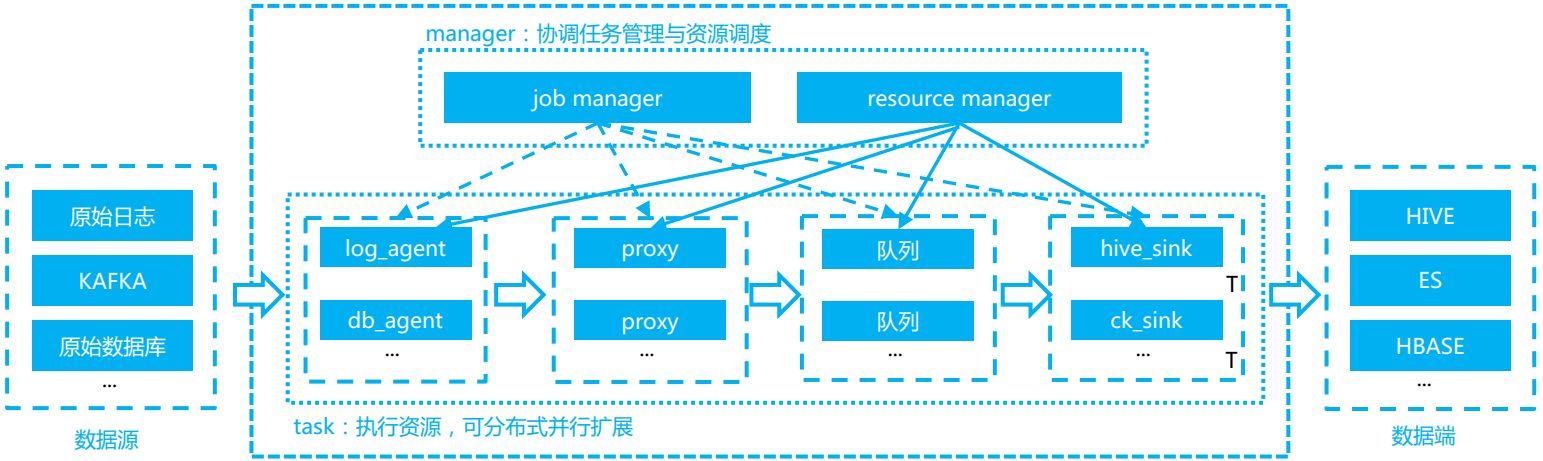
- 毫秒级时延：基于毫秒级低时延消息队列，端到端数据同步秒级时延
- 高性能：支持切换不同的缓存队列，基于存算分离架构的MQ在海量吞吐场景下具备更好的性能和稳定性

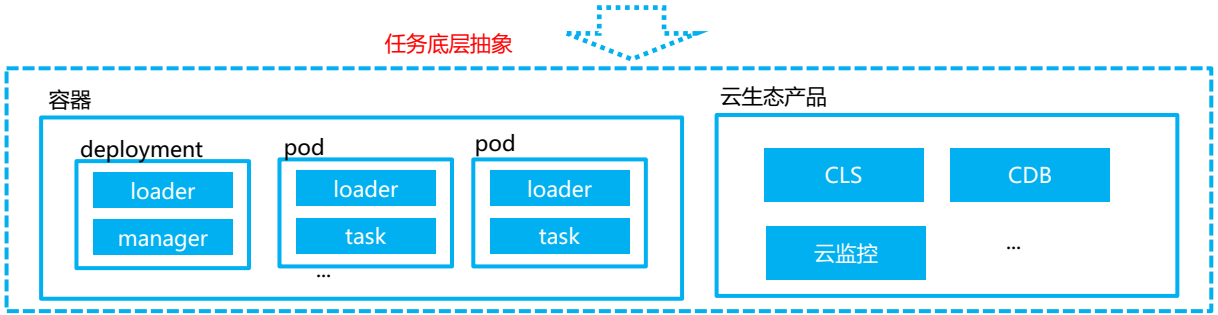
Comparison	TubeMQ	Kafka	Pulsar
Latency	Very low , 10ms	Low , 250ms	Very low , 10ms
TPS	High , 14W+/s	Normal , 10W+/s	High , 14W+/s
Filter consume	Supports client filter or server filter	Supports client filter	Supports client filter
Data	No copies	Multiple copies	Multiple copies
Reliability	Relies on RAID 10	Low	High, autorecovery
Stability	High, running in Tencent for almost 8 years with 33 trillions of message per day	Unstable when topics grows	HIgh
Client language	supports Java or C++	1 client (Official support)	7 kinds of client
CAP Model	AP	AP or CP	CP or AP



- inlong在数据采集和消息队列间增加DataProxy ，用于连接收敛、路由、数据压缩和协议转换
- 消息队列异常出现发送失败时，DataProxy会将消息缓存到本地磁盘进行容灾转发

数据集成弹性平台实现





设计原则

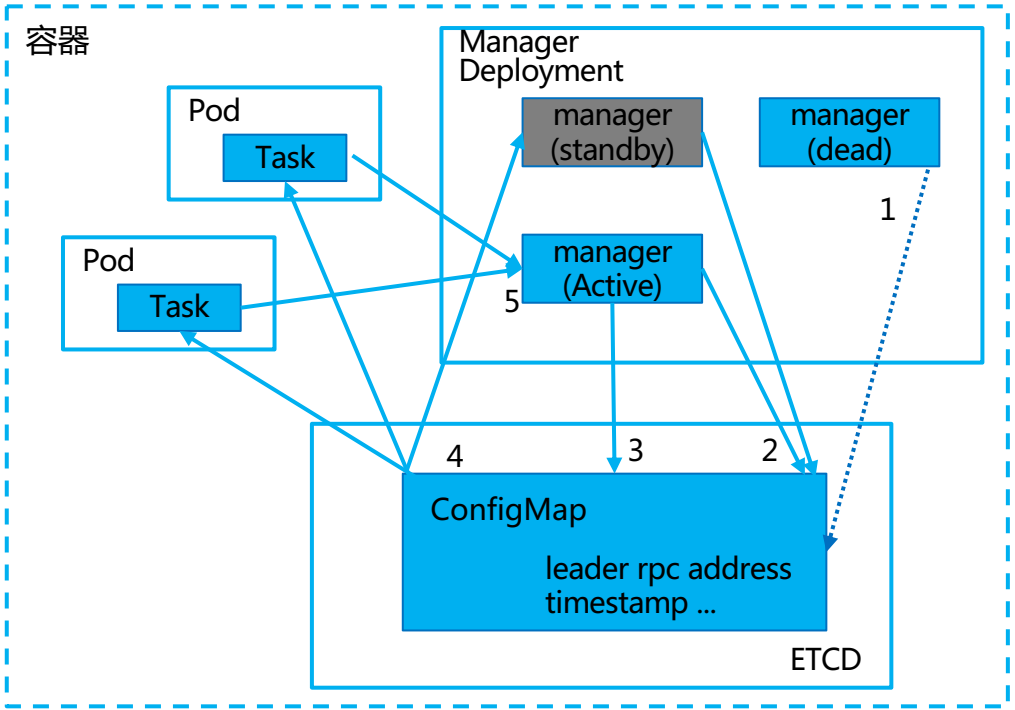
1.每个pod只属于一个任务

2. 一个任务可以使用多个pod

3. manager主备高可用

4. manager和task在不同的pod

数据集成平台高可用设计



Leader选举方案

1. 第一个创建出ConfigMap的成为Leader
2. Leader定期续租
3. Follower检查租约时间，过期则重新选举

Manager故障恢复流程

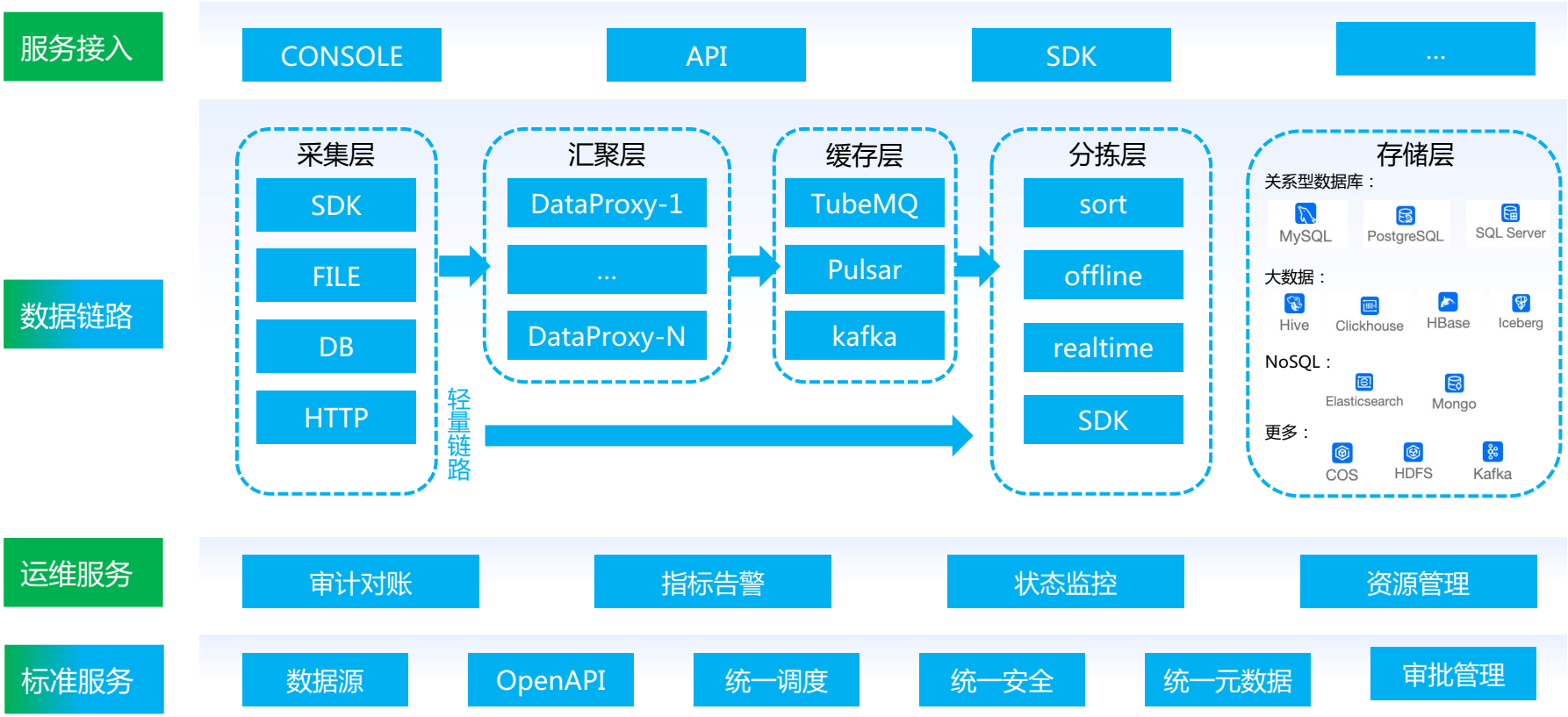
1. Manager发生故障
2. 选举Leader
3. 发布Leader信息，并开始服务
4. Task感知Leader变化，重新汇报信息
5. Leader正式服务

腾讯云 | DataFun.

03 云上数据集成产品落地

DataInlong产品简介

数据集成DataInLong：整体架构图



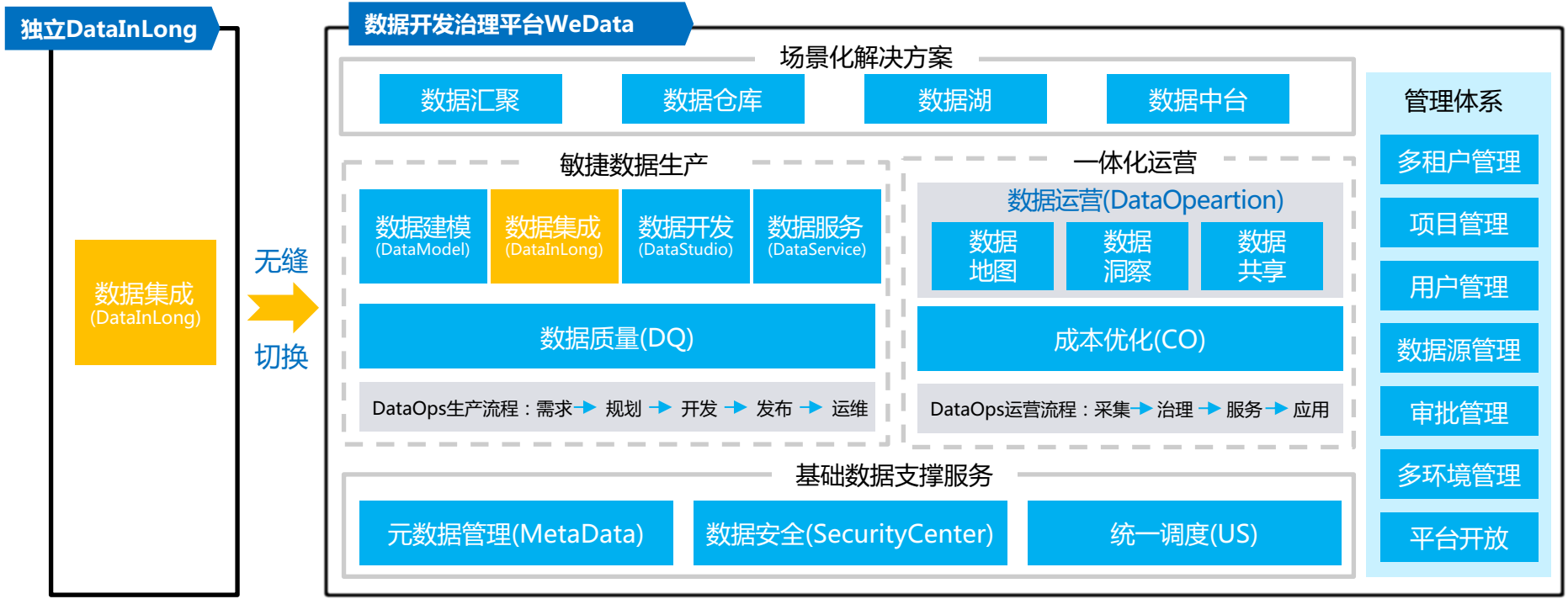
数据集成DataInLong：全场景海量数据集成服务

帮助企业建设全业务场景的海量数据传输通道

产品要素	海量集成框架	一站式敏捷体验	一体化云生态融合
	【自研技术增强+内部业务孵化】	【开箱即用】	【链路融合】
	【云原生算力融合+安全增强】	【无代码可视化配置】	【产品融合】
	【插件化灵活扩展】	【资源全托管】	【组件融合】
业务价值	稳定、安全、高性能	无门槛、免运维、低成本	平滑适配、全链路解决方案

数据集成DataInLong：全链路数据开发与治理平台搭建

DataInLong 提供了开放的技术能力，可与统一调度、统一元数据、统一安全等技术/产品服务快速深度融合。同时，支持平滑无缝升级至腾讯云数据开发与治理平台WeData，帮助企业更好应对快速变化、日益增长的业务数据需求。



数据集成DataInLong：全场景生态融合

异构数据总线构建

全链路数据开发与治理平台 / 数据中台搭建

离线/实时数据入仓入湖分析

实时报表展示

...

异构数据同步

实时/离线数据传输

数据拉取/主动上报

产品融合

稳定、高效、安全的海量数据传输通道，覆盖数据集成全业务场景，无缝融合数据生态

腾讯云 | DataFun.

04 云上数据集成的实践

客户经典案例分析

DataInLong：云上海量数据集成实践

稳定、安全

- **多集群部署**

百万亿级数据量

- 实时高性能消息队列同步
- **全链路数据指标监控**

微信支付

为支持商户/个人两大微信支付业务场景， InLong内部部署交付两套系统，每套系统支持集群三副本容灾，支持三地多活保障支付业务平稳、稳定、安全运行。

腾讯广告

为了解决广告部门数据来源广泛、采集点众多的难题， InLong为腾讯广告提供包括MQ在内的多种类型消息通道和接入方式，支持近百万亿级数据接入和处理，最终实现广告业务统一监控、告警和核心指标运营实时呈现。

DataInLong：云上海量数据集成实践

全链路数据平台

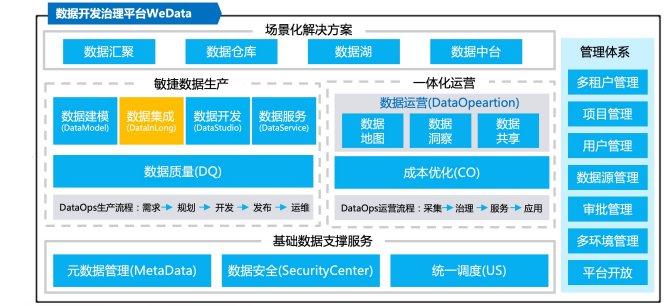
- 离线数仓与数据同步
- 离线数据开发与调度
- 元数据、数据资产管理与治理

企业云原生数据湖构建

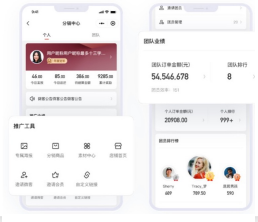
- 多种同步方式，多种数据源快速搭建云数据湖
- 实时数据秒级同步iceberg，完成冷热数据计算

某电商平台：DataInLong in WeData

方案架构



应用场景

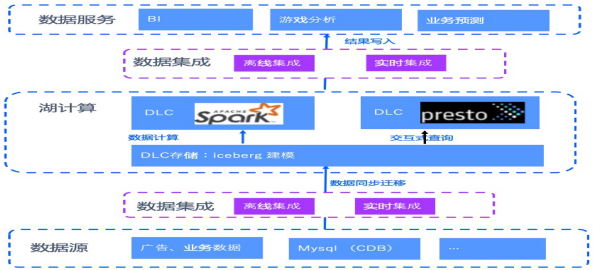


个性化推荐 用户洞察

商品/订单/库存数仓数据开发

某传统企业：DataInLong + DLC

方案架构



应用场景

信息采集 日志分析

实时数据入湖 业务预测

DataInLong：云上海量数据集成实践

某商业银行客户数据能力中心建设

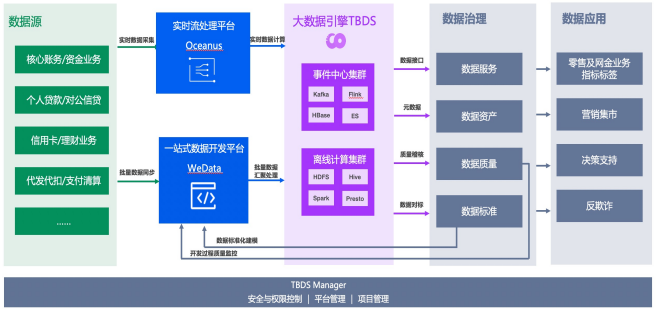
客户痛点

- 数据的开发、调度分工分散情况严重
- 数据质量低、数据落标困难
- 大数据组件运维复杂、缺乏集中统一的有效运维中心

业务数据

核心账务、对公信贷、个人贷款、在线贷款、代发代扣、外汇资金、人民币资金、理财、基金、网银、信用卡、柜面、支付清算等

方案架构



- 2021金融业新技术应用创新突出贡献奖
- 2021年度农村中小金融机构科技创新优秀案例

应用场景

零售和网金业务的指标标签 营销集市 反欺诈 智能决策 数据探索

统一开发 ： 18000+个数据任务

- 统一对接MYSQL/ORACLE/DB2/文件等多种数据源
- 统一开发HIVE/SPARK/Shell/Python等多种任务
- 基于事件和时间的统一任务调度及运维

统一落标 ： 1000个标准项

- 数据标准平台建表5000个标准项
- 通过数据开发建模平台事前落标1000个

统一管控 ： 19000+张数据表

- 数据资产平台展示完整字段级数据血缘
- 统一数据权限的申请、授权、审批等管控
- 精确到人到表的行列权限与动态脱敏控制

感谢观看

腾讯云 | DataFun.

官网：https://inlong.apache.org

代码：https://github.com/apache/inlong

Apache InLong 公众号

Apache InLong 交流群 5

该二维码 7 天内 (9月2日前) 有效，重新进入将更新