DTCC

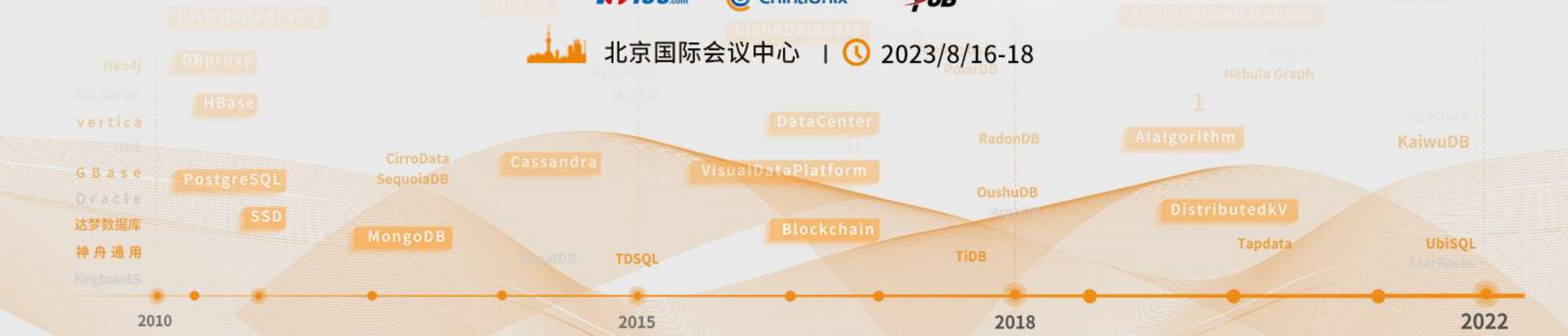
第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix ITPUB

北京国际会议中心 | 2023/8/16-18



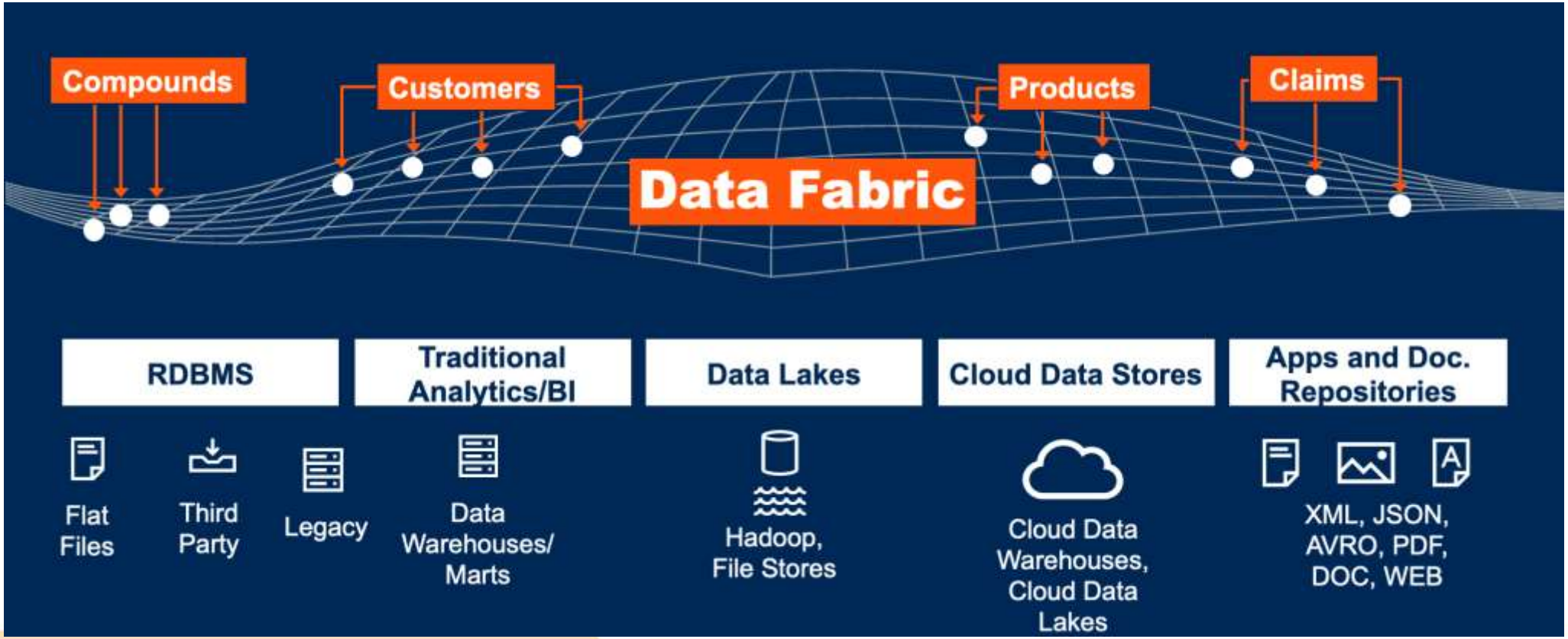
基于Data Fabric 的实时湖仓平台技术实践

北京滴普科技有限公司
FastData产品线 DLink 产品总经理 冯森

目录

Data Fabric 的定义

Gartner 将 Data Fabric 定义为一种新兴的数据管理设计理念，可实现跨异构数据源的增强数据集成和共享，通过对现有的、可发现和可推断的元数据资产进行持续分析，来支持数据系统跨平台（包括混合云和多云）的设计、部署和使用，从而实现灵活的数据交付。



Data Fabric 的发展

作为全球数据分析领域顶级研究机构， Gartner自2019年起，已连续三年将Data Fabric列入十大数据分析技术趋势。而在最新发布的2022年重要战略技术趋势中，Data Fabric更是荣登数据分析领域十大技术趋势之首，其重要性可见一斑。

Gartner 还预测至2024年，Data Fabric 可以帮助企业减少70%的数据管理工作，25% 的数据管理厂商将提供 Data Fabric 完整的框架

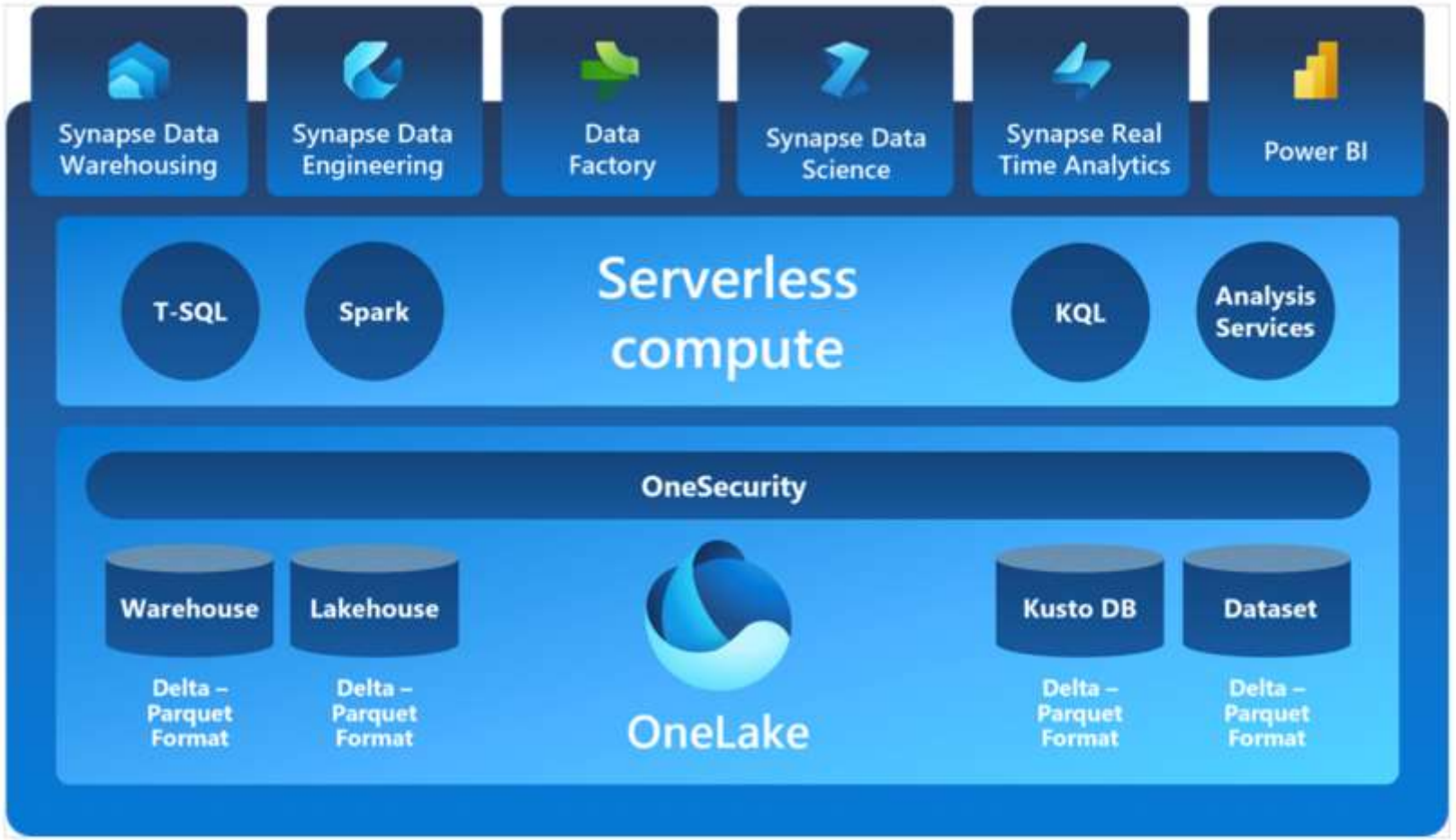
Data Fabric 的价值

- 【降本】：Data Fabric 能够帮助用户减少在数据开发，数据分析和数据管理的70%工作量，帮助企业快速分析数据价值
- 【增效】：Data Fabric 能够解决多数据源、多模态数据、跨部门、跨平台数据整合问题，打破数据孤岛，让业务人员和数据团队可以通过统一的管理平台更加高效的协作

Data Fabric + Lakehouse

- 微软于5月23日首次推出了Microsoft Fabric。该产品与OneLake（其Lakehouse产品的新名称）一起，旨在为企业的所有数据管理、分析和机器学习需求提供一站式服务。
- HPE于5月16日发布了Ezmeral， Data Fabric的升级版
- 在5月9日举行的THINK大会上，IBM发布了watsonx.data lakehouse，与IBM云数据中心(IBM Cloud Pak for Data)紧密相连，后者更多扮演Data Fabric的角色，内置治理、集成、隐私和安全功能

Microsoft Fabric: All your data. All your teams. All in one place.



目录

FastData 3.0 一站式实时智能湖仓平台

FastData是新一代数据智能基础设施，核心是基于Data Fabric架构的一站式湖仓一体数据平台，相比2.0最核心的是实现了融合的可组装的数据智能分析，并基于云原生架构扩展了国内MDS生态。

FastData 工作场景

数据工程师
数据科学家

工程化能力
工具效率提升
自动化/智能

Jobs
APIs
Data Pipeline
Notebooks
Models
Dashboards

数据分析师
业务人员

行业场景数据资产
标准/指标/模型等
数据智能应用基础

FastData 产品族

开发套件 Dev Suite

Zero-ETL集成
DataOps敏捷数据开发
开发治理一体化、全域数据资产

分析套件 Analyze Suite

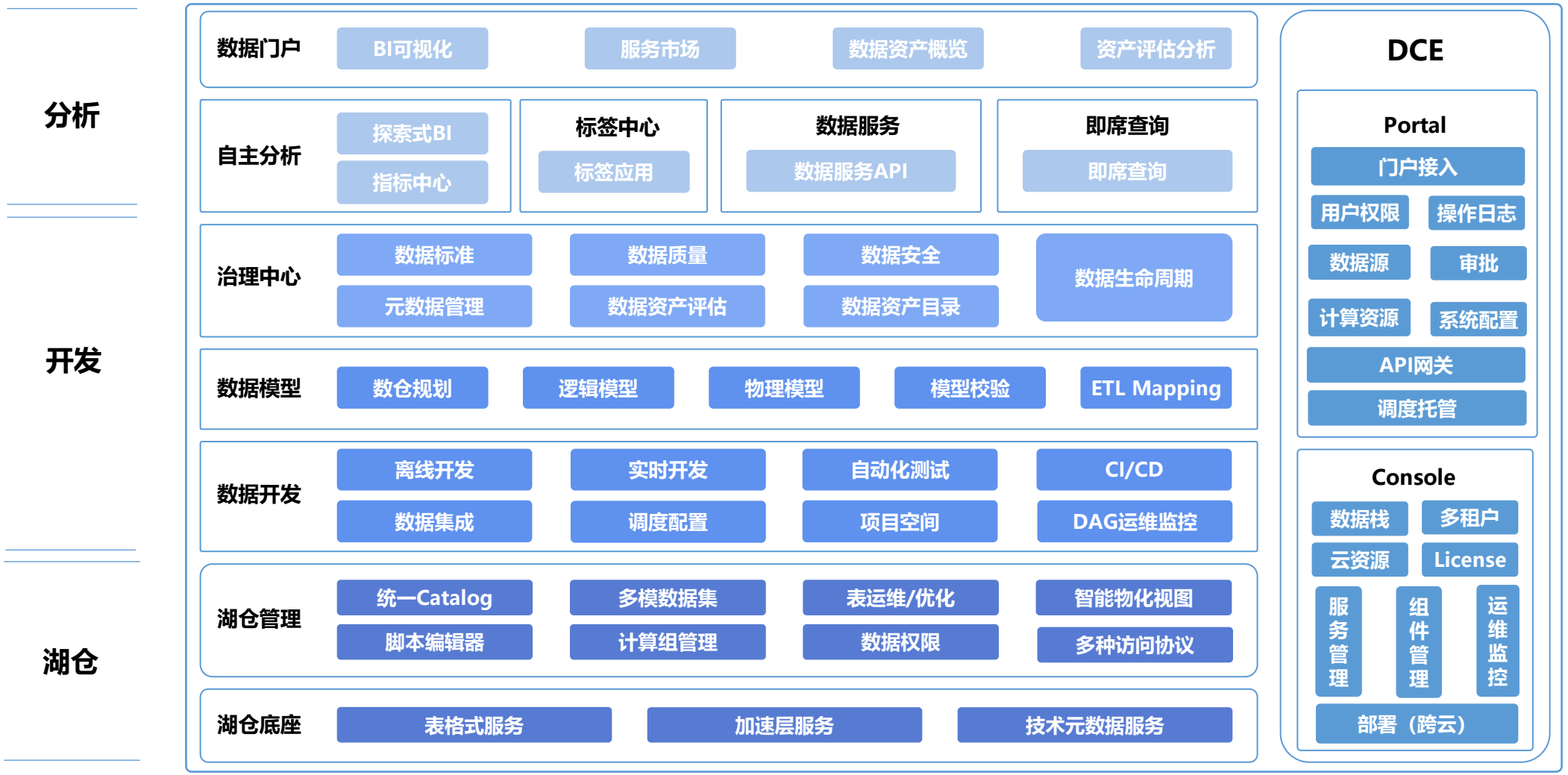
智能语义模型
探索式BI

湖仓引擎

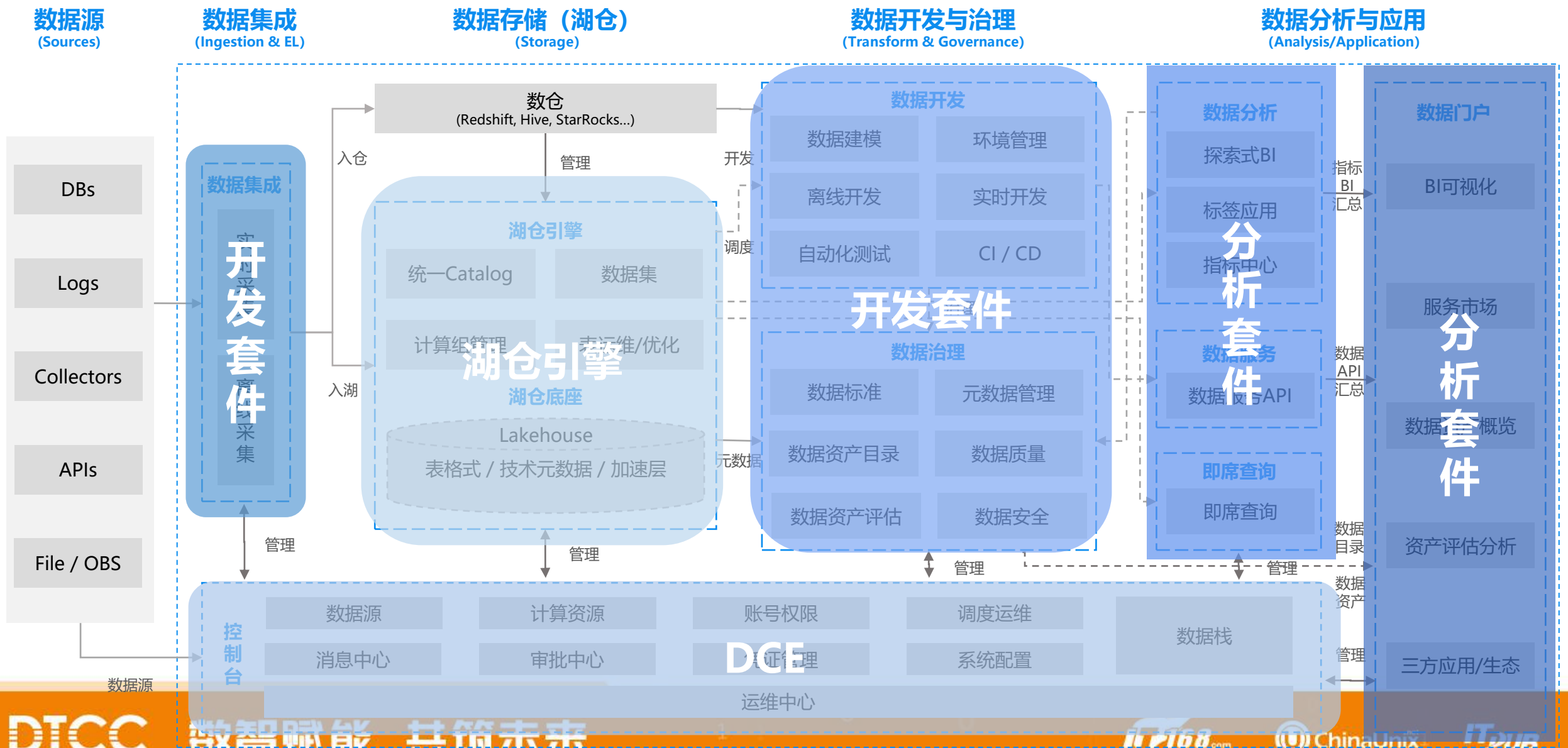
批流一体
湖仓一体

amazon web services
Azure
HUAWEI 华为云
京东云
阿里云
天翼云 e Cloud
联通云
移动云
私有云

FastData 产品功能架构



基于MDS 的FastData 全流程架构



FastData 的产品核心优势

① 低成本

- 产品基于云原生、存算分离的湖仓架构，降低企业数据存储与计算的成本。
- 支持多云平台部署

② 易使用

- 提供敏捷数据开发、主动数据治理、低代码指标分析等产品套件，让用户极易使用。

③ 可组装

- 提供可组合、可配置的现代数据栈（MDS），为数据应用与场景提供端到端组装以及全链路管理能力

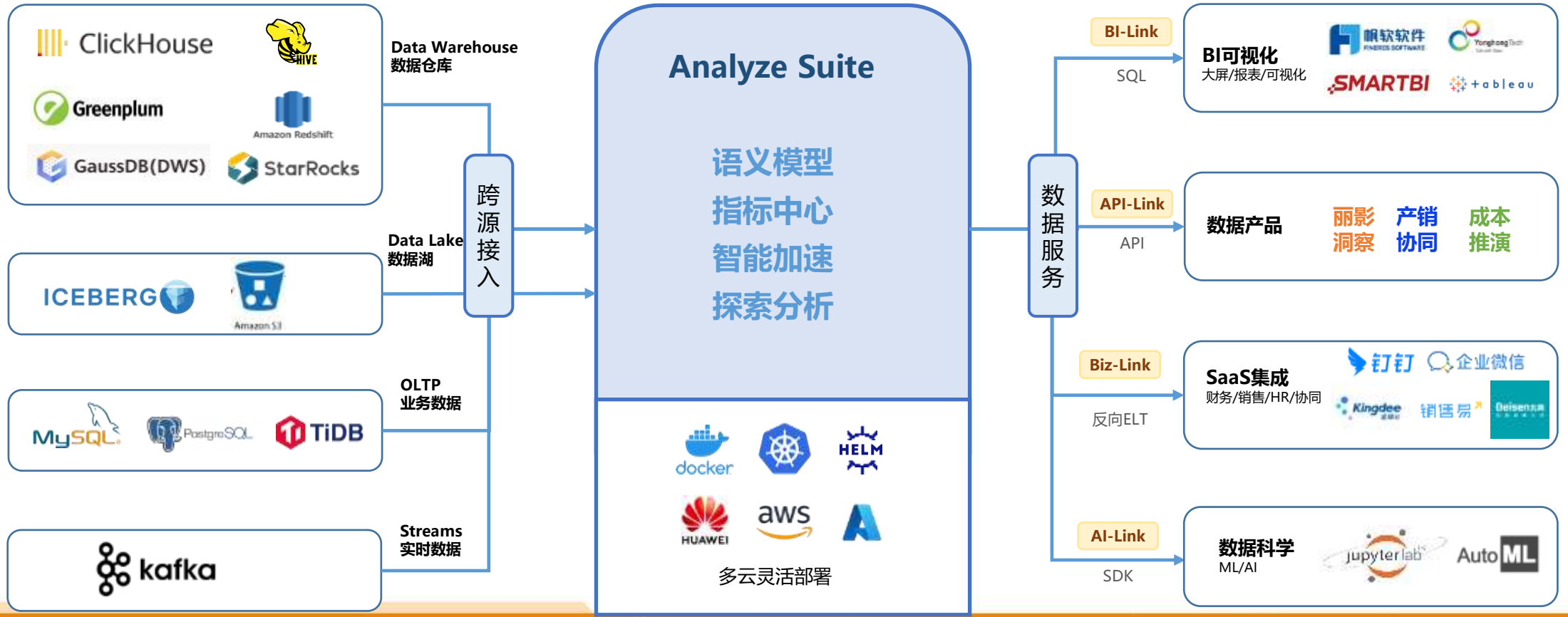
④ 易扩展

- Hadoop生态的大数据平台向湖仓一体的新一代大数据平台的演进
- 支持国产化信创要求的适配

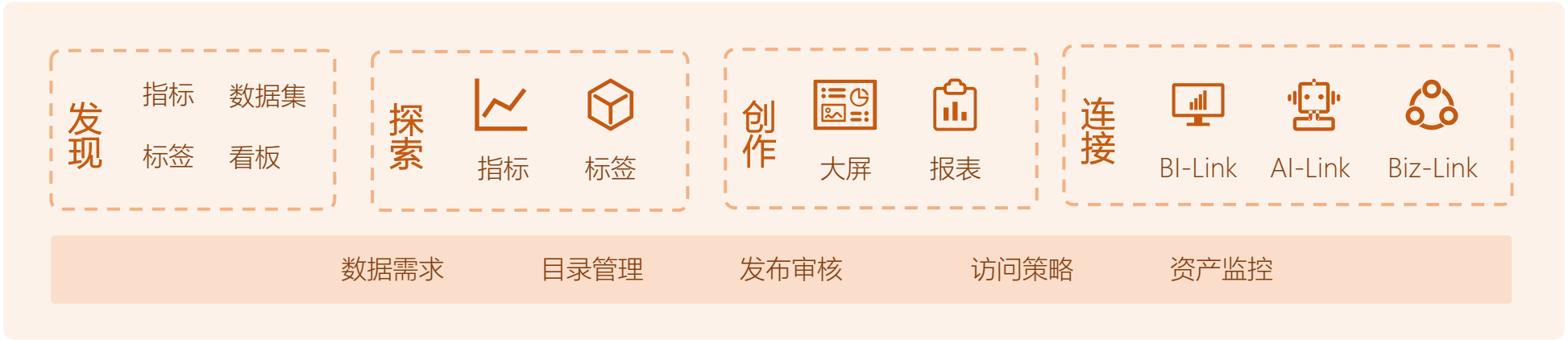
FastData 分析套件

在已有湖仓数据资源之上，旨在面向业务场景人人都能可控访问、使用业务数据、轻松获得实时洞察。

产品内核基于统一ML语义模型用业务语言定义数据，以Metrics Store指标中心低代码构建数据资产，同时通过 Link+组件为下游应用提供统一的数据消费服务，形成数据价值的多场景闭环



FastData 分析套件功能架构



发布　查询

指标中心

指标　指标加速　指标视图　探索　监控

模型

数据仓库/数据湖

FastData 分析套件价值

无门槛数据洞察

- 零基础业务人员低门槛实现数据探索及洞察

指标探索

- 仅需通过简单拖拉指标即可创建高度交互的仪表盘，展现关键业务指标信息，洞察隐藏在数据背后的见解

统一指标服务

- 通过MeticQL提供API、JDBC、多语言SDK多种协议快速接入数据，无缝集成主流BI

CubeLess智能加速

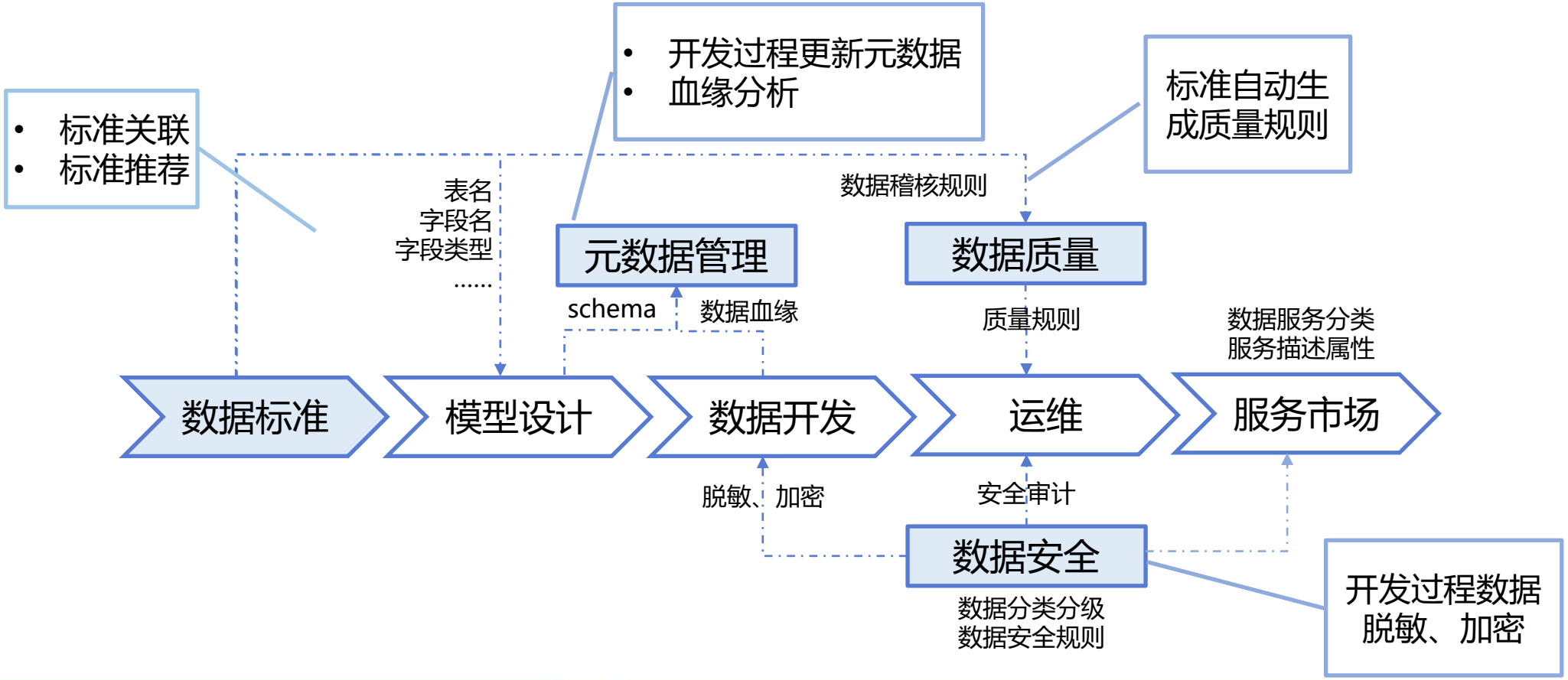
- 通过策略自动匹配到缓存层、主题加速层或者即席查询层，实现数据消费时敏捷与速度兼顾

BI 加速查询

- 通过BI-LINK，外部BI能快速连接指标体系，创作指标可视化，过程中自动加速查询

FastData 开发治理套件

- 传统场景是企业先找厂家做数据开发建设，做完后发现数据质量有问题，再找数据治理厂家做数据治理项目；
- 产品规划将数据开发与治理一体化，在开发环节同时把治理动作执行到位，从根源上保证数据质量。



FastData 湖仓引擎

生态

- 数据源
  - SaaS应用
    - OLTP
    - ERP
    - CRM
  - 文件
  - 传感器
  - 设备
  - 社交
- Kafka
- DCT
- DataFacts

实时湖仓引擎

- 批引擎服务
  - 作业管理
  - 作业集群监控
  - 集群服务
- 任务管理
  - 非结构化入湖任务
  - Hive转换任务
  - 特征提取任务
- 实时引擎服务
  - 作业管理
  - 自定义函数
  - 作业集群监控
  - connector帮助
  - 计算编辑器
  - SQL脚本
- 分析引擎服务
  - 数据源探查
  - 分析编辑器
  - 数据应用
  - 数据安全

实时湖仓底座

- 租户管理
  - 项目空间
  - 操作日志
  - 系统设置
  - 数据源/目录管理
- 湖仓管理
  - 数据集
  - 表运维
  - 数据权限
  - 统一元数据
- 湖仓服务
  - 计算组管理
  - 实例管理
  - 多种访问协议
  - 智能物化视图

基础设施

- 计算集群
- 存储集群
- 运维管控平台/DMR

基础服务

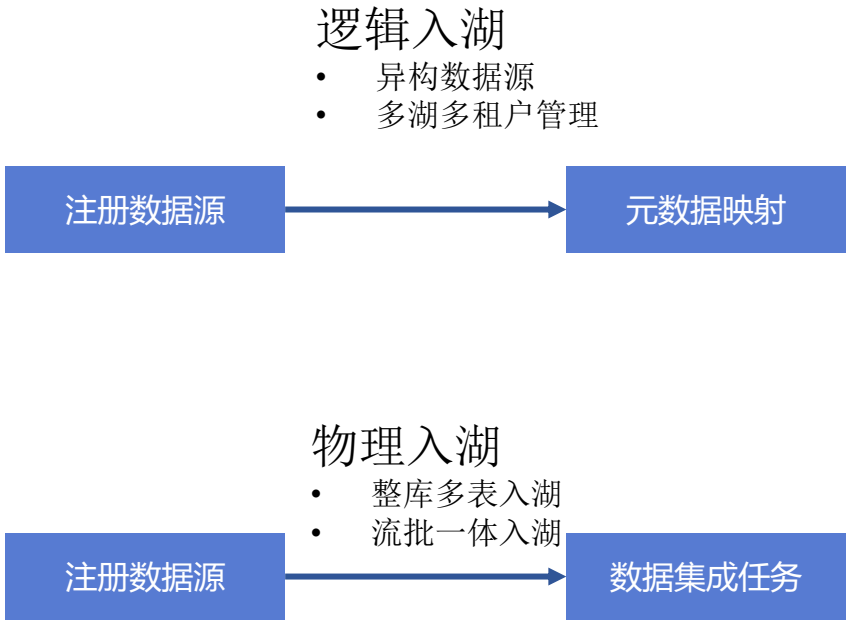
- 自动化部署
- 调度
- 监控与告警
- 日志服务
- 集群管理
- 安全控制
- 统一运维
- 云服务

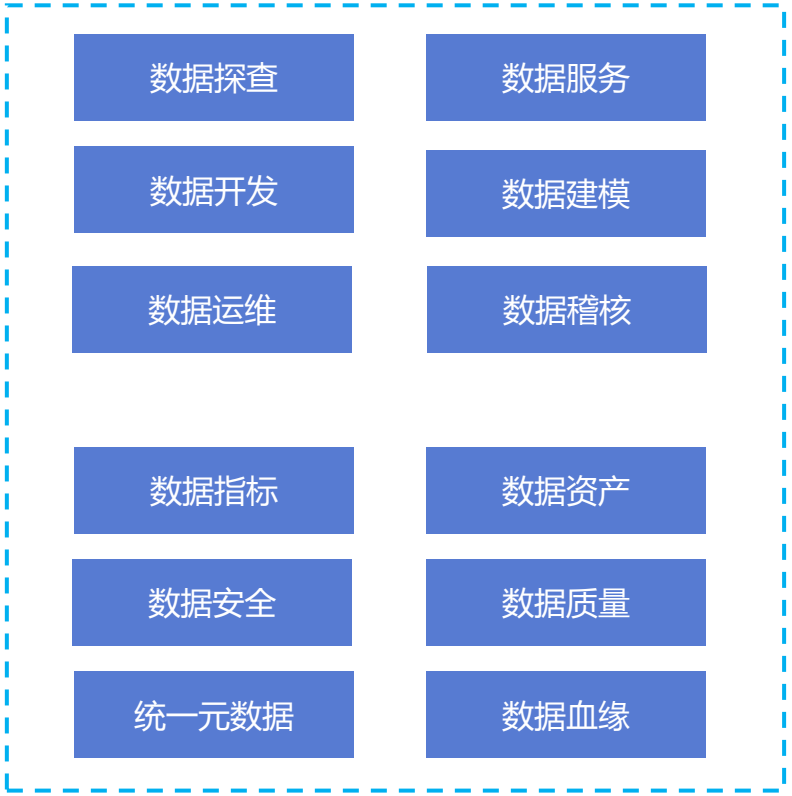
FastData 湖仓引擎价值

- 实时湖仓引擎能够支撑国产化基础数据平台演进

- 分布式数据湖架构适应大型企业多级+多租户的数据场景
- 支持PB级别海量数据存储+秒级计算处理
- 采用存算分离架构，支持EB级多模数据的存储与处理，支持流批数据处理、数据分析、数据科学等
- 多工作负载提供，并能够以云化方式提供数据服务，与传统Hadoop大数据平台对比优势明显

FastData 数据逻辑入湖与物理入湖

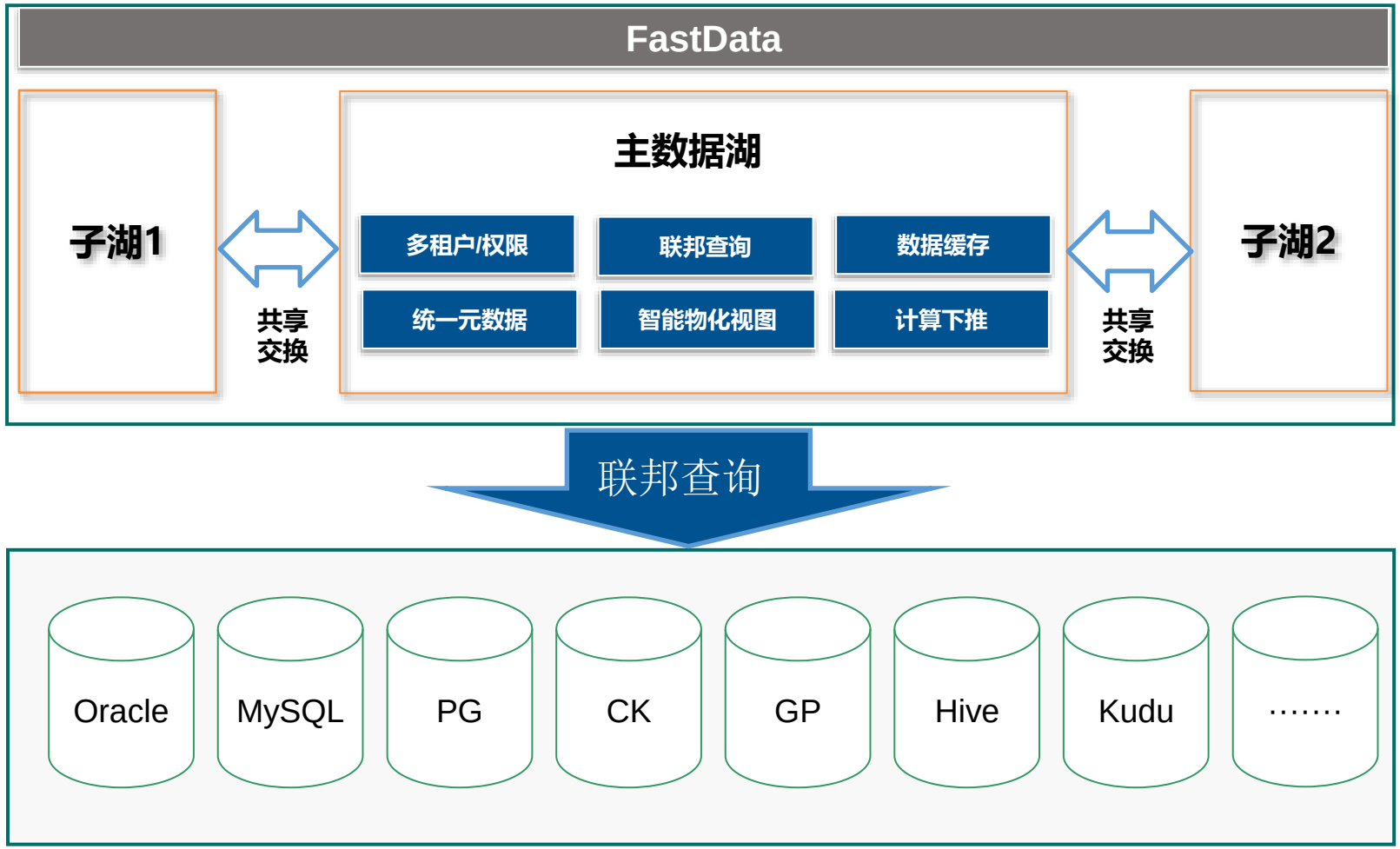




FastData 多湖/多数据源联邦查询分析

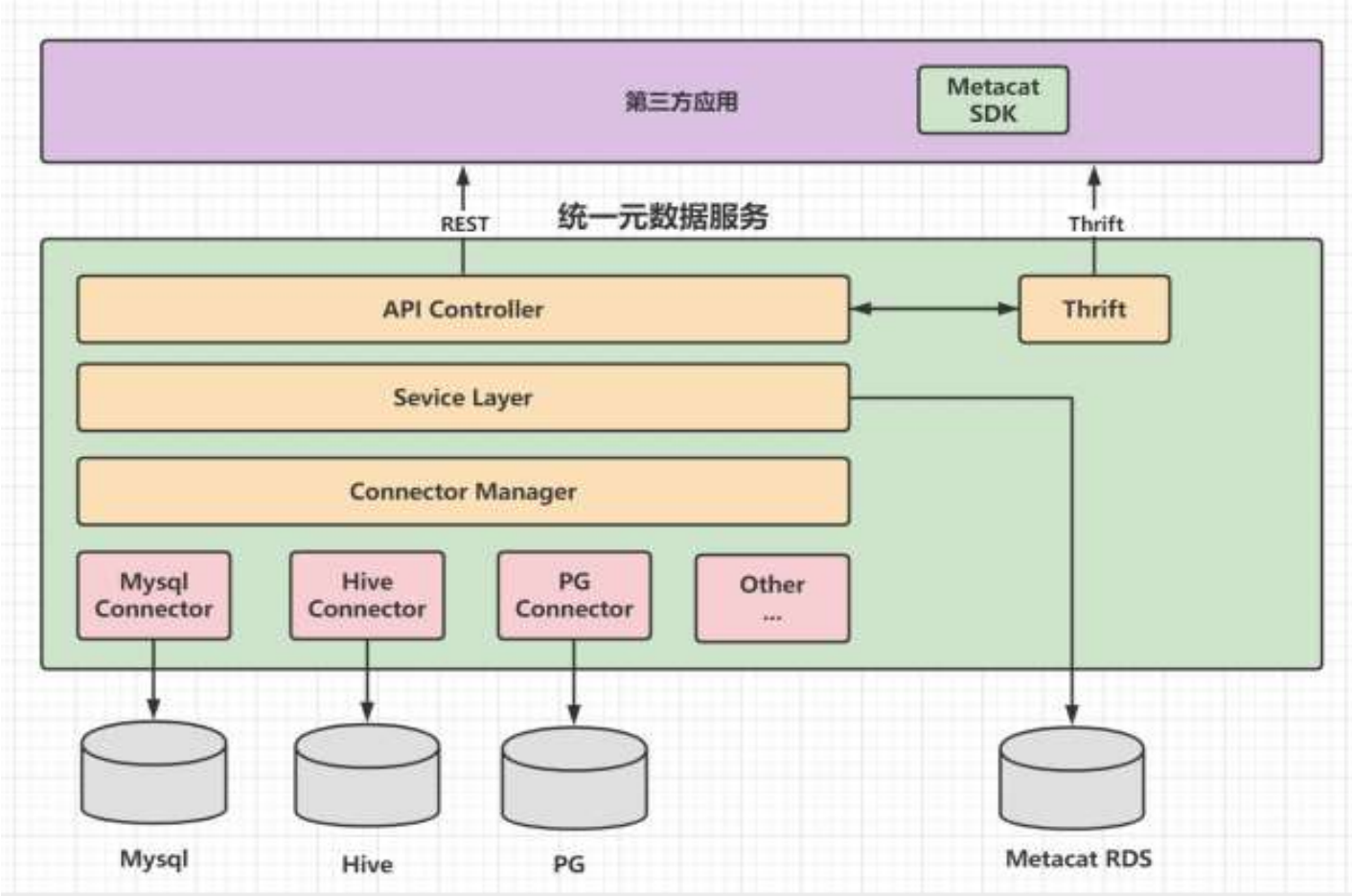
联邦分析需求

- ETL需要消耗大量计算资源。
- 数据分析时效性低
- 传输过程链路长，容易造成数据不一致
- 部分公司数据即资产，不允许数据外传
- 数据访问权限控制和多租户需求



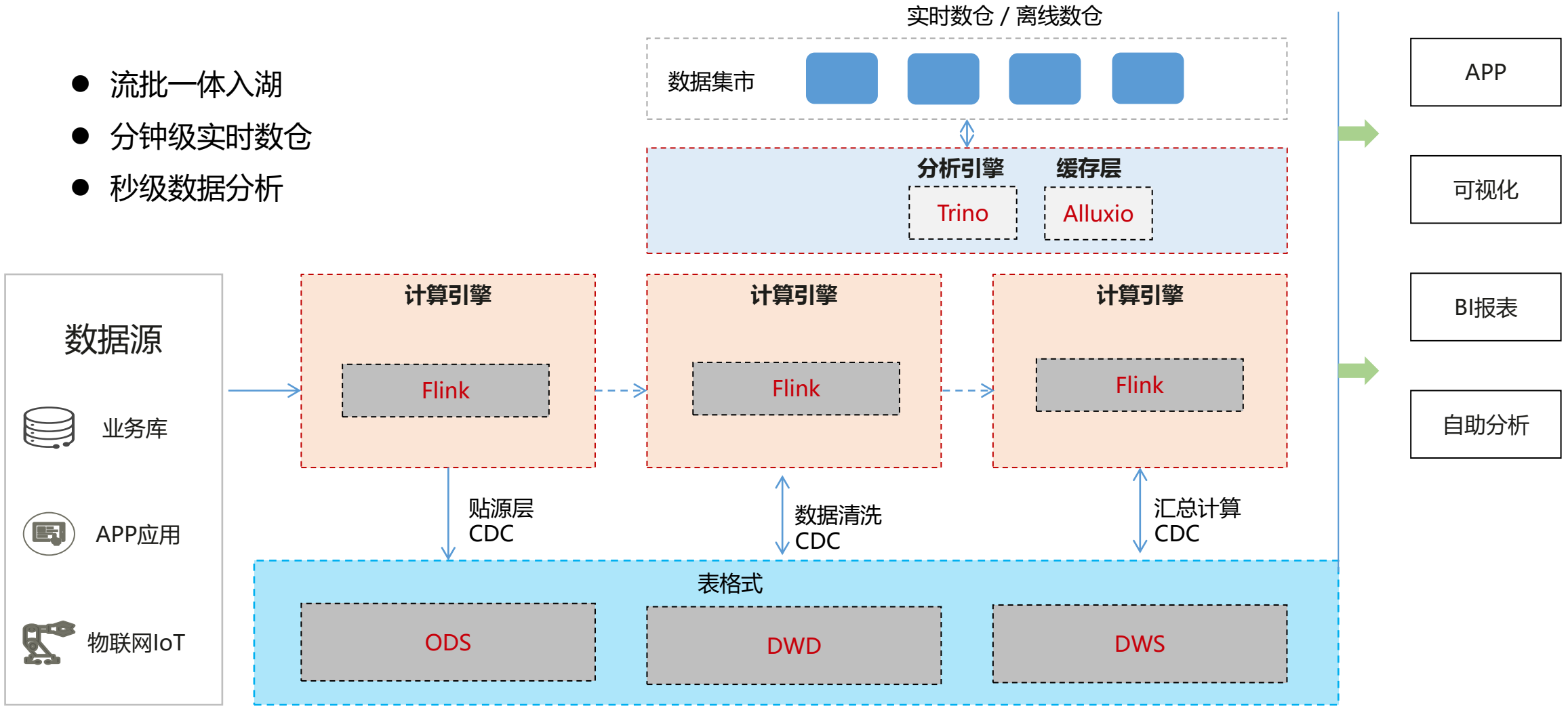
FastData 统一Catalog 服务

- 协议兼容：通过HMS多实例，支持多版本HMS协议（2.x/3.x）；
- 存储兼容：通过HMS多实例，支持多种存储介质（S3和HDFS）；
- 统一数据目录管理：提供针对不同数据源如Hive、Iceberg、MySQL 等的统一数据目录管理。
- 权限控制：所有DDL 操作鉴权在统一元数据层完成，不依赖引擎。
- 多租户：支持通过租户（和/或项目空间）对catalog进行隔离；
- 多引擎兼容：不同引擎能够使用同一份元数据，比如Spark、Flink、Hive、Trino不需要独立维护自己的元数据。



FastData 湖内构建分钟级近实时数仓

- 流批一体入湖
- 分钟级实时数仓
- 秒级数据分析



FastData 表自动化运维

让用户忘记繁重的表运维任务，让湖仓引擎开箱即用。

支持独立部署和应用

支持独立部署运维服务

无缝接入DLink

自动生成运维任务

支持手动和自动的运维方式。

根据监控指标自动生成运维任务

资源隔离

DLink运维任务的资源与业务资源隔离

运维任务之间资源隔离

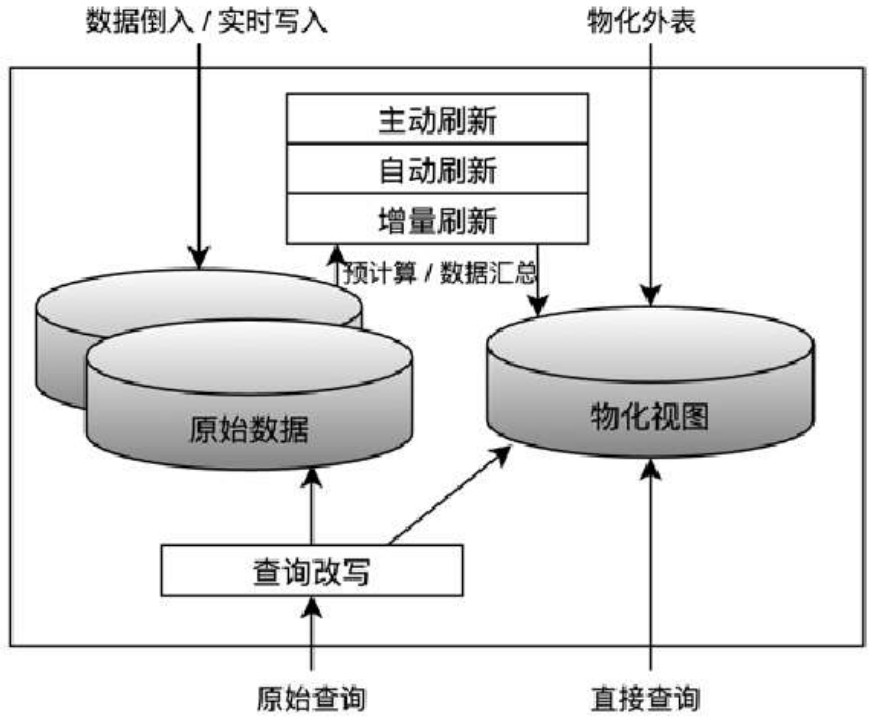
多种任务策略

用户可以自定义运维任务的策略

FastData 跨多源物化视图查询加速

应用场景：

数据分析师经常检索类似结果集，如果每次都实时计算，比较浪费系统资源。如果可以通过缓存回答，比重新计算结果或从速度较慢的数据存储中读取要快得多，消耗更少的系统资源。



左图：查询加速的流程

1.原始查询转换成关系代数表达式树。

2.匹配与查询语义等价的候选物化关系代数。

3.从候选物化列表找cbo代价最优的物化进行查询改写。

Notes:

查询改写场景支持基于规则或者SPJG结构信息（Select,Project,Join,Group）改写。

目录

FastData 已经服务约200余家企业

金融科技

兴业证券 | 朴道征信 PUDAO CREDIT | 江苏联合征信

先进制造

长安新能源 CHANGAN EV | 纳爱斯集团 | vivo | oppo
JEZETEK九洲 | 植电智能制造 | BOE | ESWIN
TCL | GREE格力 | dreame | CATL 宁德时代

生物医药

科伦药业 | 高济医疗 | HealthLink 远盟康健 | 益丰大药房 Yifeng Pharmacy
国科恒泰 | PUJIANYIYAO | 陕西医药控股集团有限责任公司

国央企/政务

中国石油 | 昆仑数智 | 国家电网 STATE GRID | China unicom中国联通
中国烟草 | 中国南方电网 | 深圳地铁 SHENZHEN METRO | CDCI 成都城投
新建元 SUNGENT | LONGHUA DISTRICT | 中核储新技术研究(上海)有限公司 | CETC 中国电子科技集团有限公司

商业综合/文旅

招花信 | ITG 国贸地产 | 华发股份 Huafa Properties | 新华联集团
HENDERSON LAND | 惠多港 HUIDUO MALL | ZYF 中亿丰控股

能源双碳

CNTRC 中汽中心 华诚认证 | 潭华能源 Tongxin Energy
亨通集团 | 国贸智谷

商品流通

百丽时尚 | Top sports | 秀域
DHgate.com | Vanguard 华润万家 | GAMBOL
AIMER 爱慕 | PAGODA
元气森林 | 泸州老窖 | 特驱

某零售行业客户数据平台建设

数据平台建设-湖仓一体时代

多模数据入湖

结构化： 1.8PB
图片/文档等：15000张
日增量：450GB
实现 货品域、交易域、办公域的源业务系统数据入湖。

数据治理

梳理183个数据域，230个系统；
实现统一元数据管理，支撑数据全链路追踪；

湖上建仓

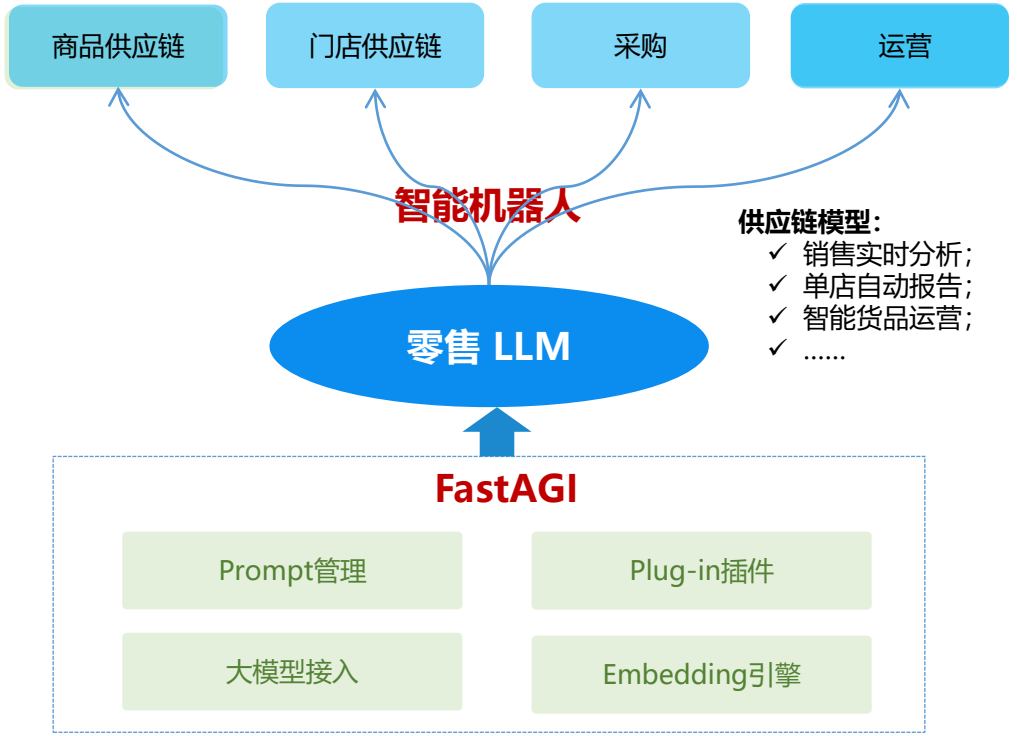
基于数据湖的业务数据字典 迁移，以及湖上建仓实现货 品域、交易域、办公域的任 务迁移。

数据资产

数据标准：177维度/474指标
数据开发：11000任务/天
数据质量：4200+监控点
数据安全：351张表700+敏感字段

会议、直播、行为数据等半/非结构化数据缺少统一存储与挖掘

数据智能决策-模型机器人时代



某新能源汽车客户数字化转型

项目背景及目标

消费者习惯、车企营销模式，以及消费者与车企的交互方式正在发生深刻而不可逆转的改变，汽车营销数字化转型势在必行。汽车行业中各家企业均在积极探索以数字化技术为依托，优化终端用户选车、购车、用车全生命周期的用户体验，以更先进，更高效的方式服务客户，提升用户满意度与用户体验。

痛点1：与用户不直连、少互动

- 与消费者触达不够，营销触达的及时性和有效性大打折扣，消费者选车、购车、用车的旅程中，参与度不高、主动性有限。

痛点2：用户价值缺乏运营

- 以”产品+渠道为核心“的营销模式，没有以“用户为中心”，客户价值无法充分被挖掘，难以转化为车企长期资产。

痛点3：营销不精准，被动服务

- 营销粗放不精准，营销渠道选择、资源分配及运营手段与目标客群不匹配。被动式服务，未给消费者提供端到端服务体验。

FastData：

- 建立以Data Fabric为核心的企业应用新型基础设施，解决数据孤岛、业务割裂及协同效率低等痛点，实现消费者全生命周期一体化运营。拉通公私域流量，并进行贯穿全体系的数字化平台建设，实现以数据赋能销售运营。

数据资产构建与业务价值

100% 线上销售	**89.7%** 覆盖客户旅程触点	**↑50%** 线索转化率
↑60% 服务效率	**80%** 服务标准化	

✓ 战略价值

①实现从传统”产品+渠道为核心“营销到”用户为中心“经营，落地商业模式

②实现价值链完整覆盖的数字化运营：形成产品和服务创新的双轮驱动

③实现某新能源汽车数字化战略，赋能业务持续数字化运营和创新

✓ 业务价值

①量体裁衣打造了贴合某新能源汽车实际业务的数字化平台，提升运营效率

②引入汽车新势力领先的业务运营管理理念和经验，提升业务价值

③数字化赋能运营、产品和服务创新，支撑企业经营持续增长

✓ IT价值

①为某新能源汽车数字化新营销和服务实现了 0 到 1 的突破！

②通过领先的中台架构，为其信息化打下坚实基础

③基于高扩展、高性能、强稳定的系统沉淀数据资产，赋能业务价值创新

目录

FastData 未来规划

- 高性能，低成本，易使用的大数据平台
- 湖仓真正一体，提升湖内数据服务性能
- 统一Gateway 服务
- 支持多云环境
- 支持大模型

……

THANKS

关注社区

个人微信