持续定义 Saas 模式云数据仓库+实时分析

原创

阿里云技术 于 2020-11-03 10:40:59 发布 107 收藏
文章标签： 数据库 saas

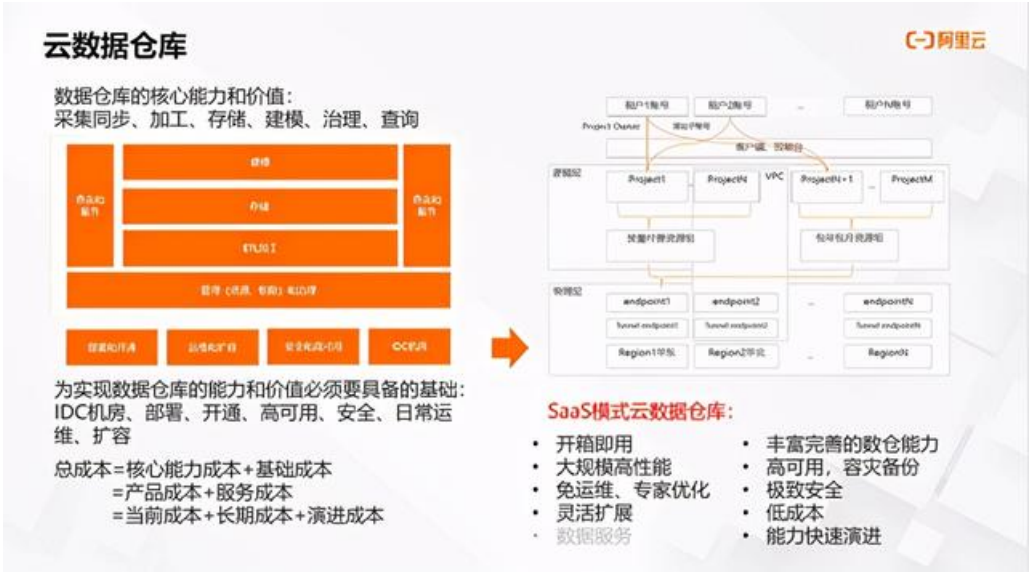
一、云数据仓库概述

数据仓库的定义是面向主题、集成性、稳定性和时变性，用于支持管理决策。 数据仓库的意义在于对企业的所有数据进行归集，为企业各个部门提供统一的， 规范的数据出口。
数据仓库（模型）本质是人收集和存储数据，认识数据，组织和管理数据，使用数据决策的最佳实践形成的方法论。 模型本身与在哪、用什么技术无关。 但逻辑模型和物理模型在最终方案中又是紧密结合的。用户需要的是数仓的业务能力和技术能力。

数据仓库的核心能力和价值包括：采集同步、加工、存储、建模、治理、查询。但是为了实现数据仓库的能力和价值必须要具备的基础包括： IDC机房、部署、开通、高可用、安全、日常运 维、扩容。这些构成了数仓总拥有成本。从各个角度看，总成本=核心能力成本+基础成本 =产品成本+服务成本 =当前成本+长期成本+演进成本。

MaxCompute是SaaS模式企业级云数据仓库。SaaS模式云数据仓库具有如下特点：

- 开箱即用
- 大规模高性能
- 免运维、专家优化
- 灵活扩展
- 数据服务
- 丰富完善的数仓能力
- 高可用，容灾备份
- 极致安全
- 低成本
- 能力快速演进。能够为企业免去拥有数据仓库的基础建设成本、维护成本、长期演进成本等非核心能力之外的投入。



SaaS模式云数据仓库可能的应用场景举例如下：

- 实时数据入仓和分析决策
- 业务运营场景-交互式业务指标计算、查询
- 各行业搭建数据仓库-流批一体、湖仓一体 □ 云上弹性扩展大数据计算和存储。

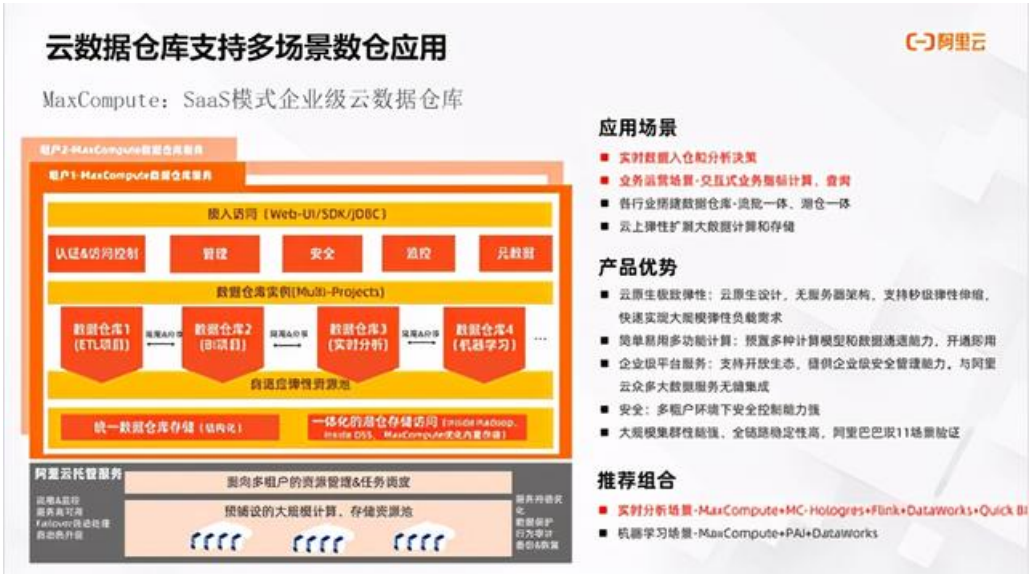
SaaS模式云数据仓库的产品优势包括：

- 云原生极致弹性：云原生设计，无服务器架构，支持秒级弹性伸缩， 快速实现大规模弹性负载需求
- 简单易用多功能计算：预置多种计算模型和数据通道能力，开通即用
- 企业级平台服务：支持开放生态，提供企业级安全管理能力。与阿里 云众多大数据服务无缝集成
- 安全：多租户环境下安全控制能力强
- 大规模集群性能强、全链路稳定性高，阿里巴巴双11场景验证。

SaaS模式云数据仓库推荐场景和产品组合例如：

- 实时分析场景-MaxCompute+MC-Hologres+Flink+DataWorks+Quick BI
- 机器学习场景-MaxCompute+PAI+DataWorks。等。

今天重点讲解实时分析场景。

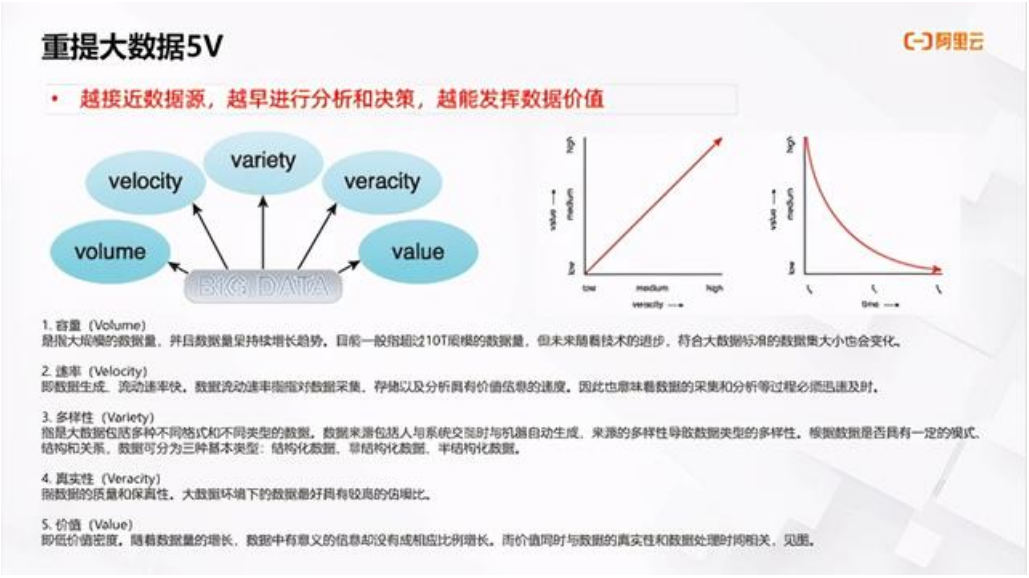


云数据仓库包含的面向用户的功能和数据流程，如下图所示。开通MaxCompute云数仓即可拥有如下全部功能和能力。

二、实时分析场景与价值

再提一遍大数据的5V能力

1 容量（Volume） 是指大规模的数据量，并且数据量呈持续增长趋势。目前一般指超过10T规模的数据量，但未来随着技术的进步，符合大数据标准的数据集大小也会变化。
2 速率（Velocity） 即数据生成、流动速率快。数据流动速率指指对数据采集、存储以及分析具有价值信息的速度。因此也意味着数据的采集和分析等过程必须迅速及时。
3 多样性（Variety） 指是大数据包括多种不同格式和不同类型的数据。数据来源包括人与系统交互时与机器自动生成，来源的多样性导致数据类型的多样性。根据数据是否具有一定的模式、 结构和关系，数据可分为三种基本类型：结构化数据、非结构化数据、半结构化数据。
4 真实性（Veracity） 指数据的质量和保真性。大数据环境下的数据最好具有较高的信噪比。
5 价值（Value） 即低价值密度。随着数据量的增长，数据中有意义的信息却没有成相应比例增长。而价值同时与数据的真实性和数据处理时间相关，见图。
其中最关键的一点是：越接近数据源，越早进行分析和决策，越能发挥数据价值。



实时分析的场景可以用以下两个类比演化出来：

类比1：大酒店同时具备其他综合业务，发展出餐 饮（实时）业务，用以更好的发挥协同作用。
演化1：以数仓分析为主场景，根据业务实时性需 求进行实时分析，构建实时通道和实时交互式分析， 形成Lambda架构。

类比2：饭店从餐饮（实时）业务发展而来， 需要更好的外围支持作用，并向综合性发展。
演化2：以实时分析为主场景，形成流式架 构，又需要能从数仓快速提取数据，和数据 源回放，形成kappa架构，后续还要考虑实 时数据和模型如何入仓。

详细分析这两种演化场景如下：

以数仓分析为主场景，根据业务实时性需求进 行实时分析，构建实时通道和实时交互式分析， 形成Lambda架构 例如IOT设备监控分析，下发策略，设备接收 后上报新数据立即进行分析，对比之前的结果， 反复分析调优。

以实时分析为主场景，形成流式架构，又需要能从 数仓快速提取数据，和数据源回放，形成kappa 架构，后续还要考虑实时数据和模型如何入仓 例如欺诈监控，必须第一时间获取分析结论，并关 联标签精准识别，最后实时数据落入数仓与其他数 据融合形成知识。

进一步的，实时分析的主要能力要求如下：

1 应用生态：

- 开发者生态
- 丰富的API、SDK
- BI工具无缝对接
- 流式处理工具和分布 式消息队列无缝对接。

2 极速查询响应：

- 毫秒级响应速度，轻 松满足客户海量数据 复杂多维分析需求
- 千万QPS点查
- 上千QPS简单查询。

3 实时存储：

- 亿级写入TPS
- 写入即可查询。

4 数仓查询加速：

- 直接分析
- 无数据搬迁
- 无冗余存储
- 统一权限。

5 联合计算：

- 统一建模方法
- 统一元数据
- 统一的管控治理体系
- 分层划域架构下的演 进和整合。

三、MaxCompute云数仓+实时分析

常见的Lambda架构有三大问题：

首先，一致性难题：

- 两套代码，两套逻辑
- 流和批语义完全不同
- 离线层和实时层数据存储和变换方式完全不同。

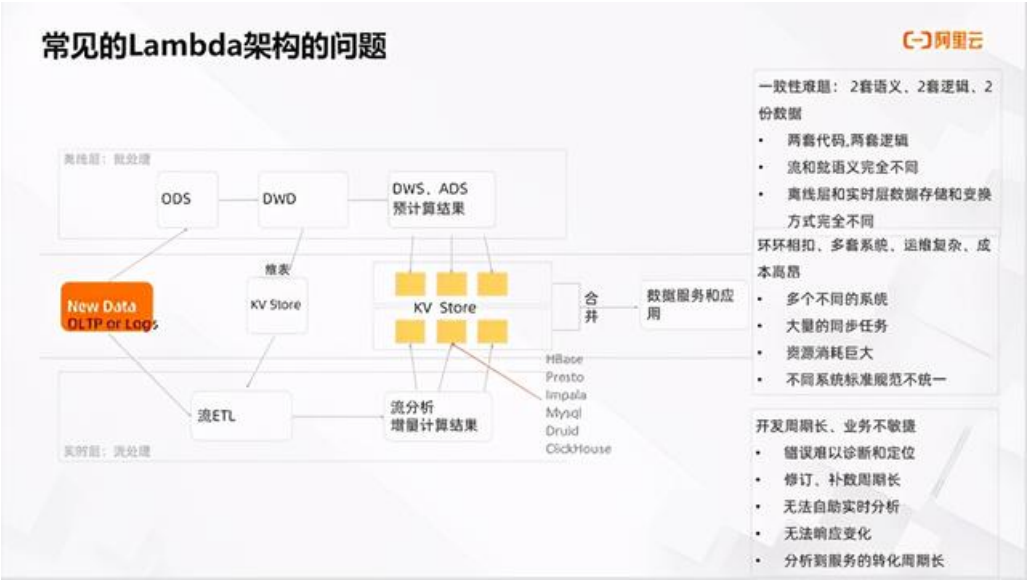
第二，环环相扣、多套系统、运维复杂、成本高昂：

- 多个不同的系统
- 大量的同步任务
- 资源消耗巨大

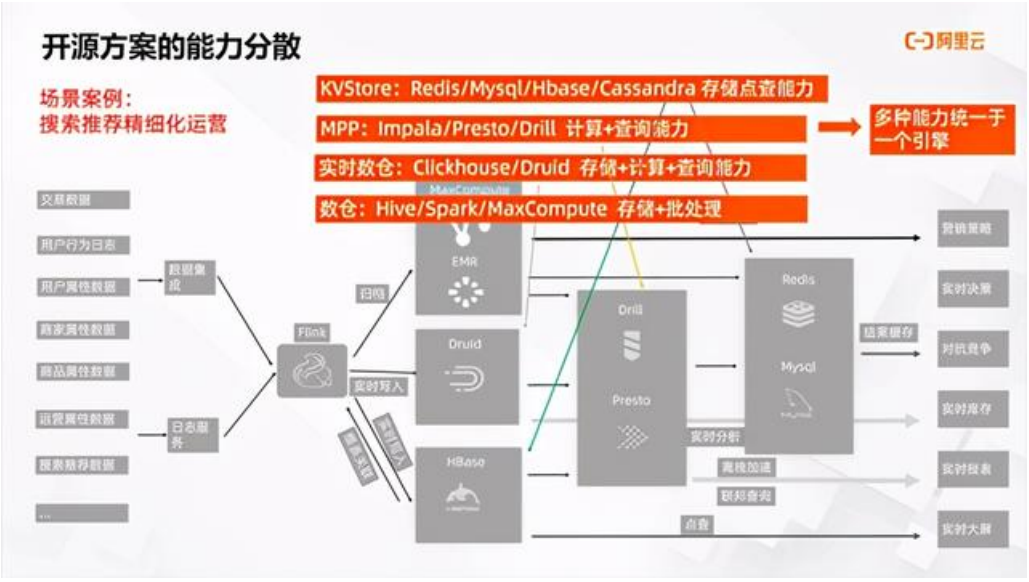
- 不同系统标准规范不统一。

第三，开发周期长、业务不敏捷：

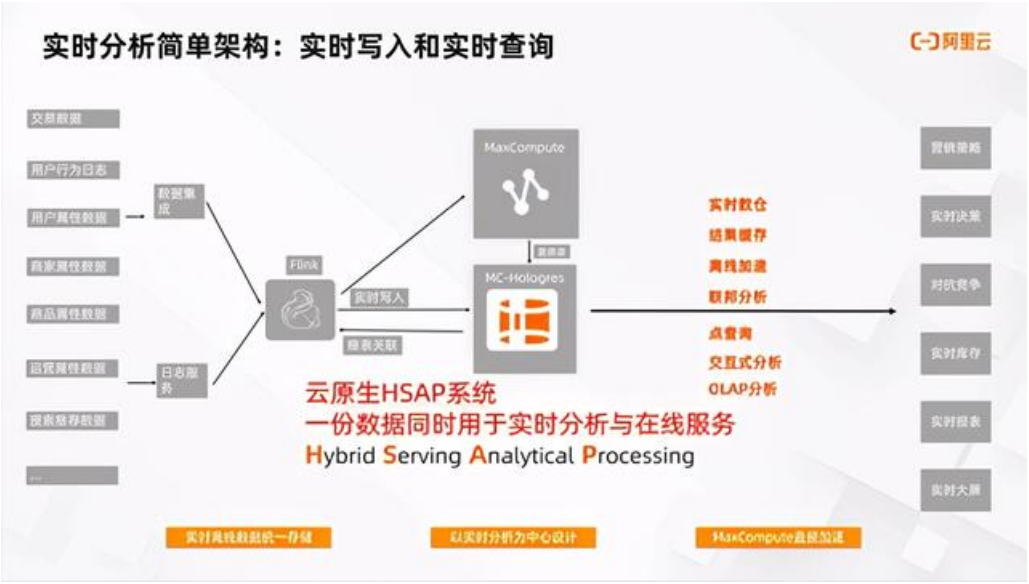
- 错误难以诊断和定位
- 修订、补数周期长
- 无法自助实时分析
- 无法响应变化
- 分析到服务的转化周期长。



以搜索推荐精细化运营的场景案例进行分析，开源方案的能力分散。如下图所示，KVStore，MPP，实时数仓，数仓具有多种能力，最好能有一种技术方案将多种能力统一于一个引擎。将存储、实时数仓、交互式分析、点查、OLAP分析等能力集于一身。MaxCompute Hologres即是这个产品和解决方案。



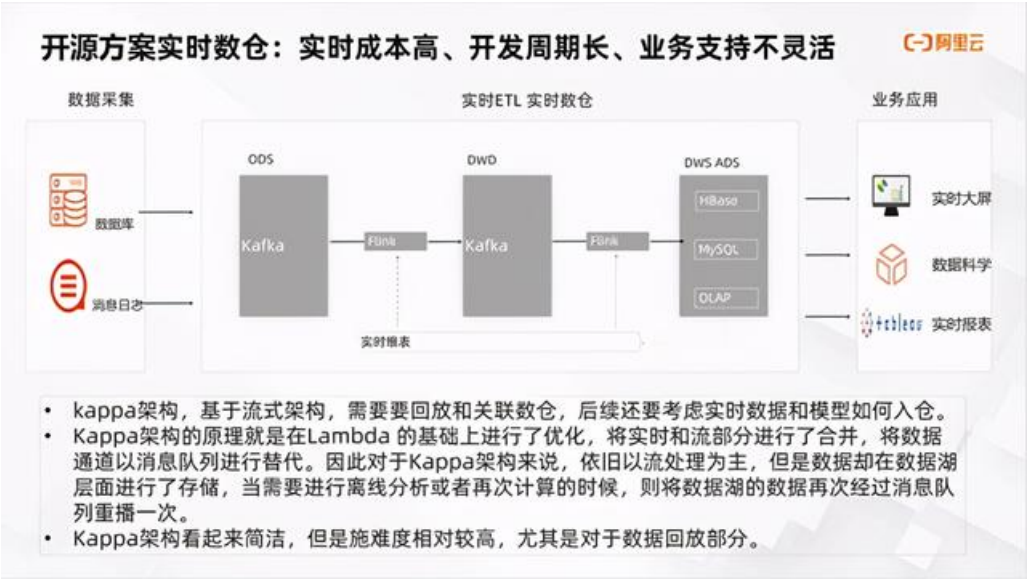
MaxCompute Hologres将实时分析的架构变得简单和高效。以实时分析为中心设计， Hologres能够实现实时写入和实时分析、查询。MaxCompute Hologres提出云原生HSAP架构中，一份数据同时用于实时分析、在线服务和实时离线数据统一存储，与SaaS模式云数据仓库MaxCompute完美结合。



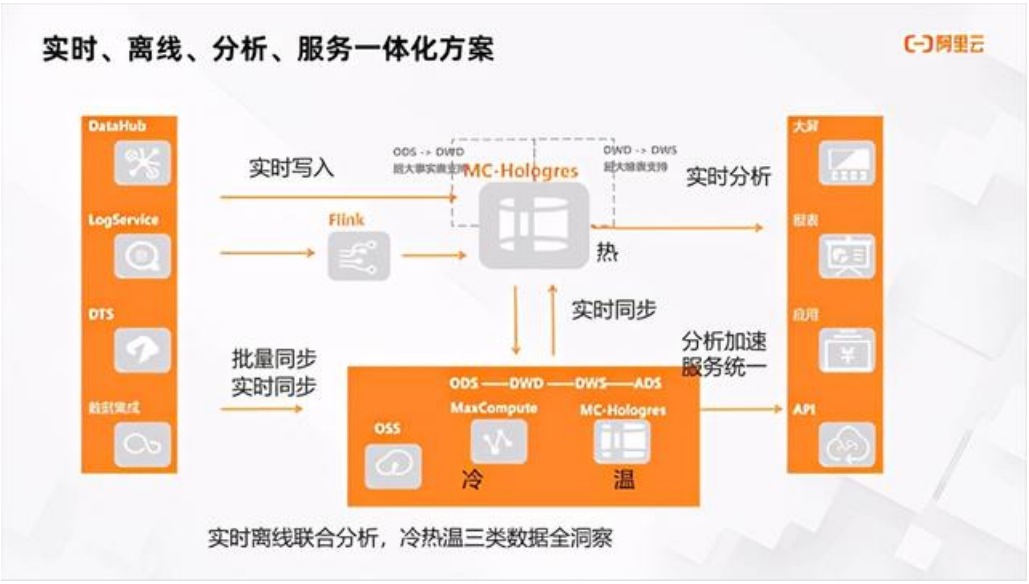
另一种场景，MaxCompute Hologres可以作为云数据仓库MaxCompute分析加速能力模块和ADS层建模能力模块。无数据搬迁、数据分析效率高。ADS层建模+服务统一、OLAP增强，如下图所示。

再看kappa架构，Kappa架构是基于流式架构的升级，需要回放和关联数仓，后续还要考虑实时数据和模型如何入仓。开源方案实时数仓有以下问题：实时成本高、开发周期长、业务支持不灵活。

Kappa架构的原理就是在Lambda 的基础上进行了优化，将实时分析和流部分进行了合并，将数据 存储和通道以消息队列进行替代。因此对于Kappa架构来说，依旧以流处理为主，但是数据却在数据湖 层面进行了存储和简单建模，当需要进行离线分析或者再次计算的时候，则将数据湖的数据再次经过消息队 列重播一次。Kappa架构看起来简洁，但实施难度相对较高，尤其是对于数据回放部分。



如下图所示，MaxCompute Hologres可以将实时、离线、分析、服务一体化，做到了实时离线联合分析，冷热温三类数据全洞察。



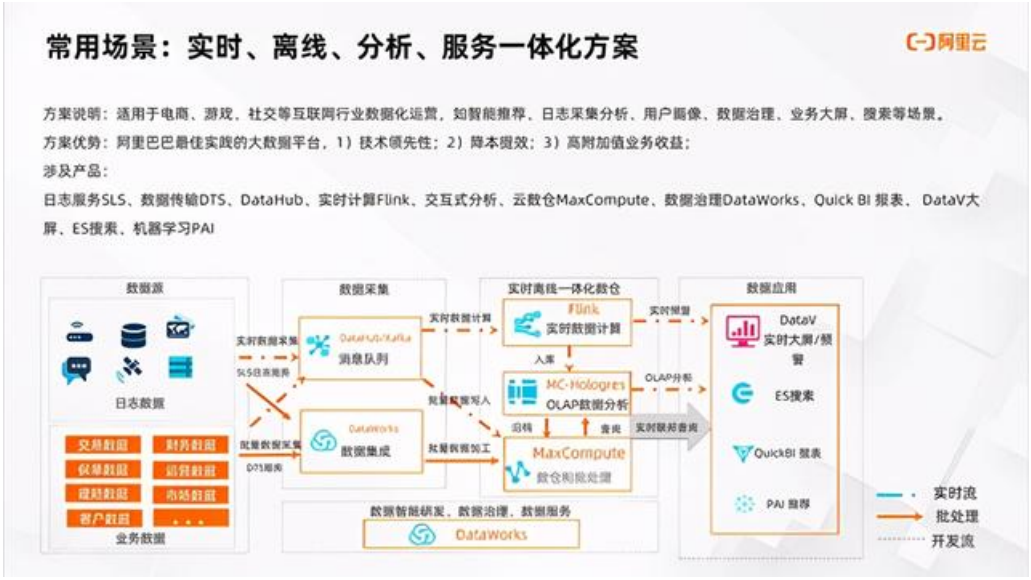
四、实时分析案例

针对实时分析的常用场景，Saas模式云数据仓库MaxCompute在拥有了Hologres后提出了：实时、离线、分析、服务一体化方案。即前文描述的Lambda架构简化、交互查询增强、kappa架构增强，实时离线联合分析，冷热温三类数据全洞察的方案能力。

此方案适用于电商、游戏、社交等互联网行业数据化运营，如智能推荐、日志采集分析、用户画像、数据治理、业务大屏、搜索等场景。

方案优势：阿里巴巴最佳实践的大数据平台，1）技术领先性；2）降本提效；3）高附加值业务收益；

涉及产品： 日志服务SLS、数据传输DTS、DataHub、实时计算Flink、交互式分析、云数仓MaxCompute、数据治理DataWorks、Quick BI 报表、 DataV大 屏、ES搜索、机器学习PAI。



小影是一款原创视频、全能剪辑的短视频社区APP，面向大众提供短视频创作工具，包括视频剪辑、教 程玩法、视频拍摄，谷歌应用商城收入榜前五，全球累计用户突破8.9亿。

用户标签数据开发：客户通过 MaxCompute 针对每 天APP产生的客户基础属性数据、 行为日志数据、内容数据等进行 计算，每天离线更新用户标签的 数据，支持营销业务的使用。

用户画像实时洞察：客户基于MC离线计算好的用户标 签，通过MC-Hologres进行多标 签、多维度的实时分析，了解用户 属性标签与内容标签之间的关联性， 洞察交叉销售机会，并通过人群圈 选，进行APP消息PUSH。

实时视频推荐：客户通过Flink + MaxCompute +MC- Hologres +PAI搭建个性化 实时推荐系统，基于用户特征和实 时行为特征，实时推荐个性化的短 视频内容。

原文链接