阿里云开发者社区
ALIBABA CLOUD DEVELOPER COMMUNITY

飞天大数据平台

SaaS模式
云数据仓库实践手册

MaxCompute帮助企业构建全托管的现代化数仓，简化数据洞察、加速价值实现

- Serverless数据仓库特性及资源管理解读
- BI分析、搜索、近实时分析、高级分析等典型分析场景深入解读
- 面向实际场景的最佳实践，可操作性强

藏经阁

阿里云开发者电子书系列

加入 MaxCompute 开发者社区
扫码关注获取更多资讯

阿里云开发者“藏经阁”
海量免费电子书下载

目录

SaaS 模式云数据仓库概述

作者 | 曲宁 阿里云智能 产品专家

Software as a Service (SaaS)是一种基于 Web 的软件应用交付模式，它改变了用户自己购买软硬件设施、自行部署和运维服务的交付模式，让应用服务直接对客户可用。

数据管理领域的技术演进以及云计算的蓬勃发展催生了基于云提供开箱即用的数据仓库服务的产品模式-Data Warehouse as a Service(DWaaS)，最终用户无需关心基础设施、平台软件管理以及平台运维和优化升级工作。这部分工作由服务提供商完全托管并提供满足 SLA 要求的高质量服务，减少用户的前期投入并加速价值实现，让数据仓库服务回归服务的本质。

阿里云 MaxCompute 正是基于云构建的 SaaS 模式的数据仓库服务，它的核心特点包括：

（1）按需使用的在线服务

- MaxCompute 预先准备了大规模资源池，无需预先资源开通、容量规划，用户可直接使用开展数据管理和分析工作；
- MaxCompute 提供存储计算分离和 Serverless 无服务器的架构设计，面向用户提供 On-Demand 的按需使用服务能力，用户可仅为实际使用付费。

（2）集成现代数据仓库完善功能的多租户服务

- MaxCompute 内建了高性能存储引擎，多种主流的计算分析引擎（SQL、机器学习、Spark 等）和内外部数据管理能力，满足现代化数据仓库分析需求；同时提供完善的 API/SDK/CLI 用户接口，并支持与广泛的生态集成；
- MaxCompute 是个多租户系统，通过完善的多租户隔离和管理能力。提供对不同组织的租户间进行资源、数据、任务的强隔离以保障安全。同时通过权限控制机制，支持组织内或组织间安全、受控地进行资源共享交换；同时为每个租户提供资源监控、任务管理、作业诊断能力，支持用户自助进行必要的管理工作；
- 作为企业级数据仓库服务，MaxCompute 提供完善的安全管理能力，包括：访问控制与授权、多租户/作业级别的安全隔离、操作审计、数据保护（隐私脱敏、数据加密、备份恢复、异地容灾）等能力，满足企业级不同的安全、合规需求。

企业用户在关心数据仓库产品新的交付模式和特性之外，会更加关心如何借助利用这一技术平台满足企业实际的业务需求。本电子书将介绍借助 MaxCompute 这一 SaaS 模式云数据仓库服务的典型使用场景和价值，包括：

云数据仓库+BI：云数据仓库的低成本、高性能，赋能组织内众多用户按需使用，促进数据民主化；

云数据仓库+AI：现代化的数据仓库服务在统一的企业数据资产之上，不仅提供历史分析，更需要是借助数据提供预测性分析，进行业务决策；

云数据仓库+实时分析：传统数据仓库以 T+1 洞察为主，如何为企业提供实时洞察分析能力，让各级业务人员实时决策以提升业务效果成为数据仓库领域的热点话题；

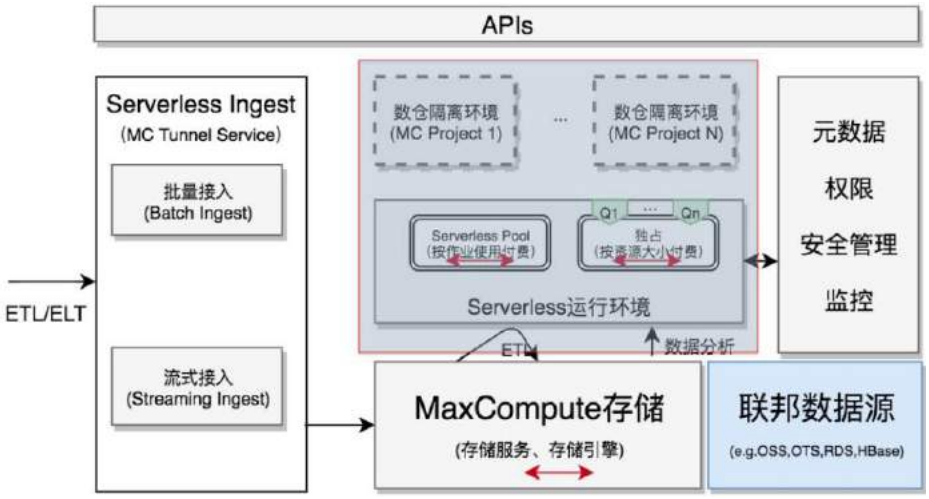
云数据仓库+实时搜索：如何利用数据仓库对多样的企业数据进行加工、整合，为企业内外部客户提供更有效的信息检索能力；

云数据仓库+数据银行：SaaS 模式能够在服务级别与外部服务高价值服务进行预先集成，特别是外部高价值数据的集成能够大大提升。MaxCompute 与友盟数据银行服务集成，低成本扩展企业数据能力；

本期电子书将重点从 SaaS 模式云数据仓库按需使用的核心能力-Serverless 能力、"云数据仓库+价值场景"这 2 个方面进行解读。首先我们先重点介绍 MaxCompute 的 Serverless 能力介绍。

一、Serverless 简介

下图是 MaxCompute 的 Serverless 架构，主要包括数据接入服务、多计算环境、储存服务和管理几个模块。



其中各个模块的主要特点如下：

（1）Serverless 的数据接入服务

- 提供 Tunnel 批量、流式导入，转换为 MaxCompute 列存格式、自动伸缩等功能，且免费；
- 可以免费使用 LOAD/UNLOAD 命令进行 OSS 导入/导出。

（2）Serverless 的多计算环境

- Severless 计算资源池，大规模计算资源池，On-demand 按需提供，按作业付费；
- 独占计算资源：支持包年包月付费、Workload 管理（负载隔离、优先级、分时伸缩等）；
- 运行环境（runtime）支持 ETL/OLAP/ML 等大数据分析使用场景。

（3）Serverless 的存储服务

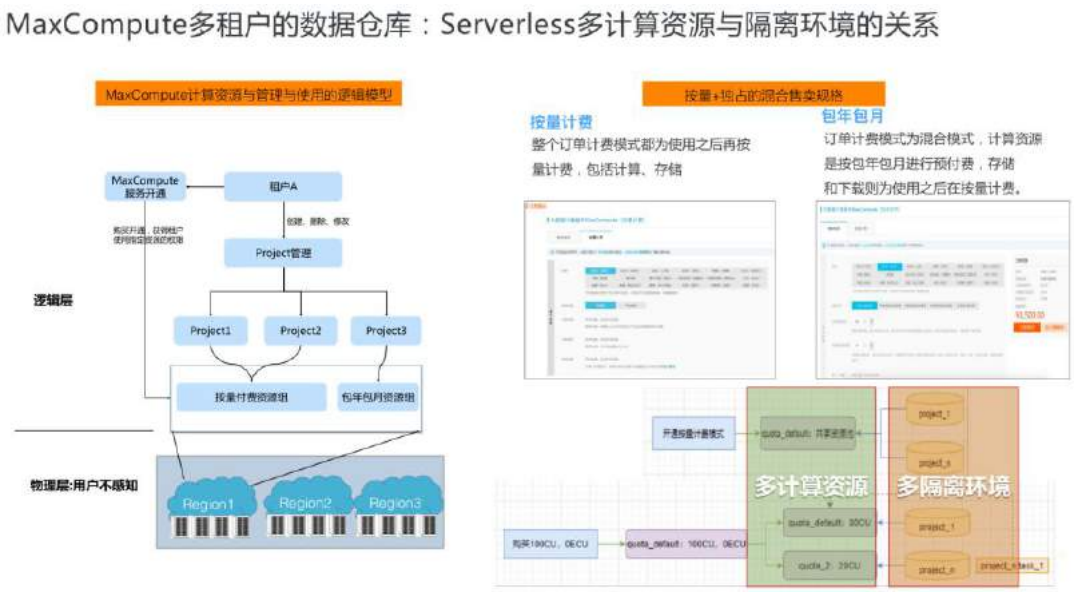
- 与计算无关，独立伸缩，提供 GB-EB 级别的存储服务；
- 按实际存储大小付费，降低成本；
- 无需指定，默认面向分析优化（列压、压缩）；
- 支持区分/分桶/Zorder 等优化手段。

（4）Serverless 的管理

- 开箱即用，内建了完整的管理能力，以 API/sdk/web-console 管理；
- 平台侧无需用户运维，降低成本。

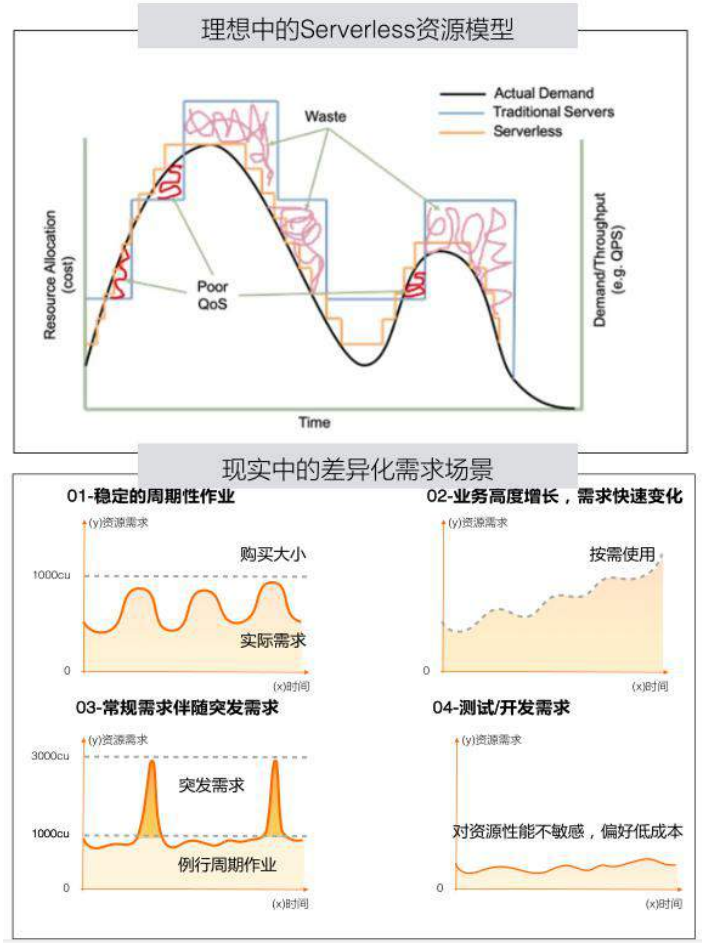
上面是对 Serverless 架构的一个简述，本文的重点是如何利用 MaxCompute Serverless 计算资源来满足数据仓库的需求。

下图所示为 MaxCompute 计算资源管理与使用的逻辑模型。对于 MaxCompute 中的 Project，实际上对应的是一个逻辑的数据仓库隔离单元，我们可以根据不同的管理目标创建不同的 Project，比如我们可以分别创建面向测试的 Project 和面向开发 Project，两个项目之间有独立的数据和权限管理体系，并不互通，达到管理的隔离作用。当然，只有这样的隔离空间还不足够，因为我们的计算任务需要绑定计算资源，我们可以将 Project 与付费方式绑定，根据需求，对不同的 Project 设置不同的计费方式，使得不同的隔离空间使用不同的计算资源。



在上述的体系之下，MaxCompute 有着一些独特的特点，首先就是有一个多租户环境，我们在开通了 MaxCompute 可以根据不同的管理需求创建多个隔离的数据仓库空间，对于企业来说，可以购买多组逻辑上的计算资源，这种多计算资源、多隔离环境，可以更好地满足不同的场景需要。

如下图所示，理想中的 Serverless 资源模型要求我们很好的规划资源的利用方式才能够完美的适配我们的实际需求（图中黑线）。



但是，实际上我们的客户有不同的资源需求，有着众多的差异化需求场景，其场景主要有：

- 稳定的周期性作业场景；
- 业务高度增长、需求快速变化的场景；
- 常规需求伴随着突发需求的场景；
- 测试/开发需求的场景。

从各种场景中我们可以发现，大数据计算对计算资源的需求方式并不是一个完完全全的纯 Serverless 的按需分配的需求，而是不同的阶段有不同的需求，且不同类型的需求有不同的要求，其对计算资源的需求特点主要包括如下：

（1）业务敏捷性需求

- 长期处于成长期，处理能力能满足业务自然增长的需要，特别是业务快速变化的阶段；
- 可以是企业的初期，也可以是创新部门的创业业务。

（2）周期性峰谷差异明显

- 每天、每月周期性的峰谷波动巨大，以峰值容量规划，成本和 SLA 难以平衡；
- 需要常规算力+弹性算力，根据调度/人为指定作业资源策略。

（3）稳定的业务，关注关键任务的按 SLA 产出

- 基线作业，与非关键作业的 SLA 需求不同，基线产出时间需要保障；
- 非关键作业尽可能低成本处理，同时不影响关键作业。

（4）资源治理：算例需求由快速变化转变为稳定可预期

- 对 CU 的容量规划，相互转换以及测算；
- 固定资源的精细化的 Workload 管理。

总得来说，现实中我们的算力需求追求的目标就是在满足现实中的差异化需求的前提下，还能够达到成本最小化的目标。

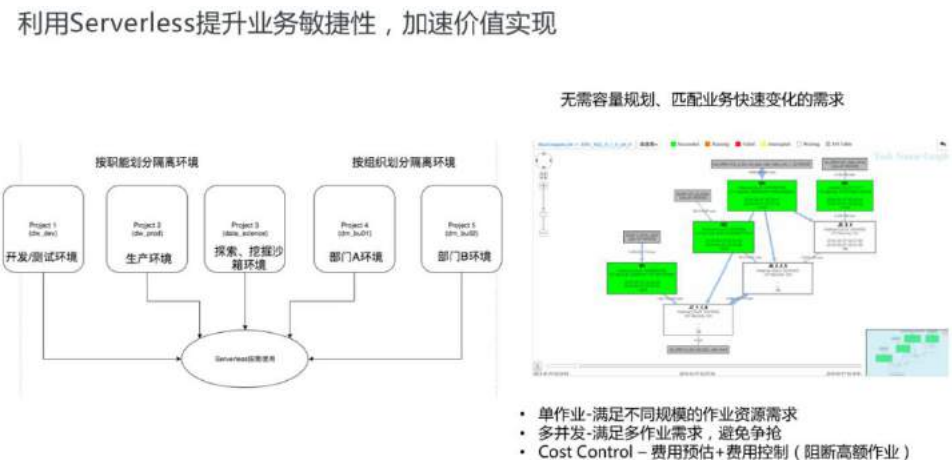
二、Serverless 助力业务敏捷

那么，MaxCompute 的 Serverless 如何满足上述的场景和需求呢？如果是一个业务快速发展、快速变化的企业，我们建议使用 MaxCompute 的 Serverless 按需使用的计算资源。从管理上来讲的话，我们可以建立不同的 Project 去做一些隔离的划分，比如说建立一套开发测试环境，一套生产环境。对于有些分析师来讲，他们往往随机地需要对一些明细数据做大量的探索，或做机器学习分析，往往有一些突发的算力需求，且这个算力需求的规模可能非常大，这个时候往往这些作业要和其他的环境隔离，因为他们是低频的，但是却需要对海量数据做分析。

我们还可以按照组织划分，比如很多企业的组织比较大，可以按照部门来进行划分，使得每个部门有一个隔离的环境，各个部门作为一个独立的组织，他们需要相对独立的数据和计算资源，我们可以使用 Serverless 按需分配的这种模式。有了这种模式之后，企业无需进行容量规划，在初期的时候可以使用按量付费的方式，通过这种超大的资源池来满足各个部门的资源需求，避免资源的争抢。

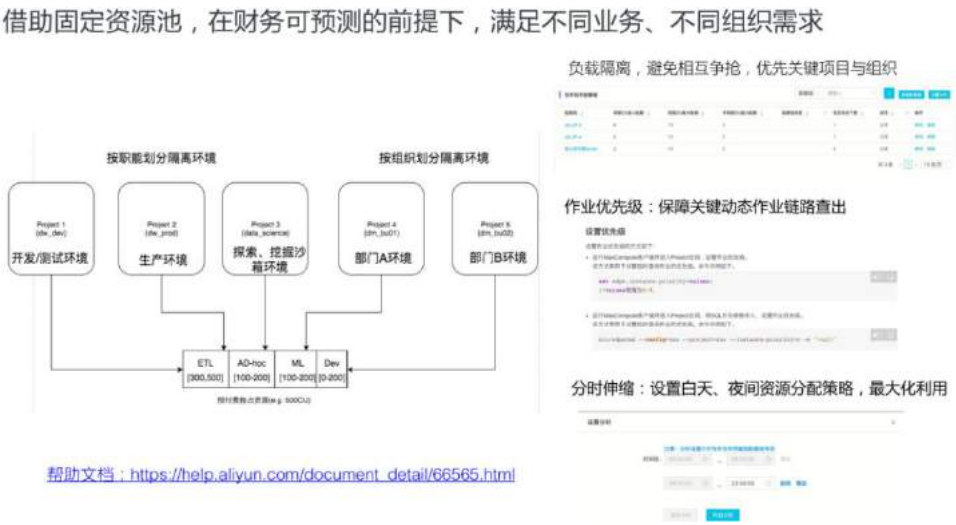
总的来说，利用 Serverless 在各种作业情况下 Serverless 都能够很好的满足需求：在单作业的情况下，无论是规模大小，Serverless 都可以很好的满足不同规模的作业资源需求；在多并发的情况下，Serverless 也能够满足多作业需求，避免出现资源的争抢情况

出现；在某些我们希望能够控制作业费用的情况下，MaxCompute 也可以提供费用预估+费用控制的方式来阻断高额作业。通过上述的方式，MaxCompute+Serverless 可以大大提升业务敏捷性，加速价值实现。



另外，有一些企业结合自身日常的管理环境，更希望有一个相对稳定的资源池，因为其本身有一定的资源规划和资源治理的能力，这种情况下，我们购买一定固定规格大小的资源，然后按照职能或者按照组织划分隔离环境，利用 MaxCompute 提供的配额组管理能力将资源切分成多个资源组，在财务可预测的前提下，满足不同业务、不同组织的需求。这种模式的关键技术特点有：

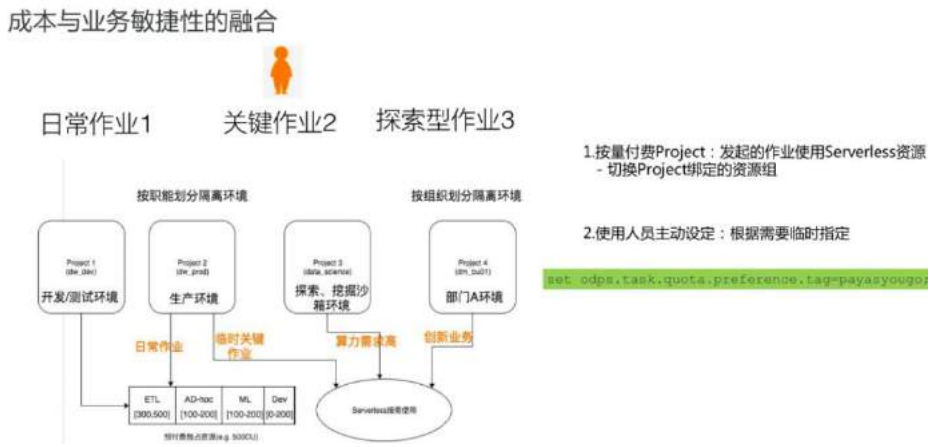
- 负载隔离，避免相互争抢，资源优先分配给关键项目和组织；
- 作业优先级：保障关键动态作业链路查出；
- 分时伸缩：设置白天、夜间资源分配策略，最大化利用资源。



第三种场景是关于成本与业务敏捷性的融合。举个例子，数据平台的管理者可能经常面临多种作业：一类是日常作业，通常把这里作业放在一个固定大小的资源中，成本可控、可预期；另外一类是一些关键作业，我们愿意花费一定的代价把它加速完成去满足业务需求，对于这类作业我们希望能够拿到一些额外的算力；还有一类作业是面向数据科学家的探索型作业，我们希望这类作业对我们的生产作业不要产生干扰，同时数据科学家又能利用强大的算力快速地完成业务假设和设想，我们可以将这类作业放在按需的资源池中；在复杂的企业中可能还会有创新类的业务，他们需要一个新的数据开发环境和应用创新的环境，我们可以新建一个数据仓库的隔离环境，按需分配资源，帮助他们快速的去验证业务假设。

产品侧我们主要提供了两种能力给用户：

- 按量付费 Project：发起的作业使用 Serverless 资源，可切换 Project 绑定的资源组；
- 使用人员主动设定：根据需要临时指定计算资源，作业级别的资源路由。



上面的三种场景都是在日常运营中的真实场景，还有一个场景就是客户在使用了按量付费一段时间之后，业务稳定下来了，希望将项目放在固定的、预付费的资源池上。这时候就会遇到一个问题：如何进行资源的需求评估呢？毕竟之前按量付费的时候是不需要进行资源需求预估的。MaxCompute 提供了容量规划来解决这个问题，其原理是利用 MaxCompute 提供的元数据服务（information schema）根据历史的算力消耗情况来预估项目的整体算力需求，其关键信息有：

- 基于 information schema 按天统计近期项目作业消耗的计算单元时（算力单位：cu 时）；

基于 information schema 按天统计近期项目作业消耗最高的一天，计算每个小时的算力需求（算力单位：cu 时）。

根据上面的信息我们就可以根据一定的规则来预测业务的算力需求，进行容量规划，关于这部分的详细内容大家可以到阿里云社区查找相应的文章进行了解。

三、总结

上文主要分享了如何利用 Serverless 服务来更好的进行资源管理，低成本地满足不同业务的资源需求。总得来说：

（1）按量付费的模式适合业务快速发展及变化阶段，配合 MaxCompute 的 cost control 管理手段，既能满足业务的算力需求，又能有效的控制成本。

（2）对于预付费的资源，我们可以通过 quota 管理，切分多个计算资源，做相应的负载隔离、分时管理，利用 DataWorks+MaxCompute 基线作业优先级保障关键作业 SLA。

（3）对于预付费固定资源和弹性按量付费组合的方式，我们可以根据作业级别选择不同的计算资源：对于突发作业，使用按量付费补充突发算力需求；对于周期性作业中的尖峰需求，也通过按量付费满足，从而达到资源的有效利用，且降低成本。

（4）我们可以利用元数据来进行算力需求评估，进行容量规划，从而在按量付费和预付费方式之间进行转换，还可以利用元数据来进行资源消耗分析，进行资源的优化，降低资源高消耗的作业，做相应的资源治理。

SaaS 模式云数据仓库+BI

作者 | 韦海青 阿里云智能 高级产品经理

简介：本文为大家带来持续定义 SaaS 模式云数据仓库+BI 的介绍。内容包括云数据仓库概述，BI 使用场景与趋势，基于 MaxCompute 云数仓+BI 的特性，以及实践案例。

一、 云数据仓库概述

今天和大家一起探讨一下我们 SaaS 模式下云数据仓库加上商业智能 BI 能有什么新的东西出来。我们先来看一下云数据仓库的一些概述。预测到 2025 年，全球数据增长至 175ZB， 中国数据量增长至 48.6ZB。数据量暴涨这个前提下，我们看一下 BI 市场规模的增长。预测到 2023 年，我们中国 BI 软件市场年复合增长率为 32%。云计算也同样在增速发展，2019 年第四季中国云数据市场的增长率已经达到 66.9%。

云数据仓库可以让企业几分钟内创建并开始使用数据仓库服务，在更低的成本下，专注业务，通过对大规模数据进行多样化的处理、挖掘、分析，快速获得业务洞察。它有四大特点：大规模数据分析，高性能，灵活扩容，低成本。

二、BI 使用场景与趋势

商业智能（BI，Business Intelligence）是一种以提供决策分析性的运营数据为目的而建立的信息系统。随着我们社会发展以及数据量的爆发，在这么大量的数据支持之下，企业希望能快速从这些数据里边挖掘出更科学的一些数据，然后对我们的企业有一个科学化和数据化决策的帮助力。同时，BI 也会助力企业用到一个精细化运营，客户关系维护，还有成本控制等。

我们看一下商业智能建立一个信息系统它主要的一个流程。首先是数据接入，将分散于我们企业内外各种数据集成和进行整合。然后再进入一个数据准备阶段，就是一个 ETL 的阶段。然后再到一个数据分析的阶段，最后将这些成果交给决策层，决策层就可以通过这数据进行一些决策。不管是精细化运营，还是客户维护关系，还是成本控制，都可以从这些数据里边得到一些助力。

随着数据量的暴涨，我们的业务快速的增长，产生了各种分析需求。不仅仅是分析多样，而且还想要实时的，比如说秒级的即时查询。同时在这么大量的数据基础上，数据的安全合规也越来越受到重视。所以需要快速的整合多系统数据和实现信息透明，以及构建一个统一的简单易用的可视化分析平台，提高制表效率。这已经成为 BI 系统的新的趋势。

三、基于 MaxCompute 云数仓+BI 的特性

MaxCompute（原 ODPS）是一项大数据计算服务，它能提供灵活快速、完全托管、高性能、低成本、安全的 PB 级数据仓库解决方案，使您可以经济高效的分析处理海量数据。基于 MaxCompute 云数据仓库的基本架构如下图所示。底层的集群是 MaxCompute 本身搭建好的，用户无需感知。再往上，有多种的计算引擎。引擎之上提供各种的 API，还有深度的集成了一个一站式大数据智能云研发平台 DataWorks。在云数据仓库的这么一个体系下，可以做数据准备，进行各种清洗、加工、分析之后，就可以进入一个数据消费的阶段。

总结一下 MaxCompute 云数仓的特性。

（1）是一个开箱即用的在线服务。免平台运维，总体拥有成本低。
（2）极致弹性能力。弹性扩展，无需容量规划即可应对业务规模的快速变化。
（3）简单易用，多功能计算服务。多种计算模型，多种数据通道，外部数据源联邦计算。
（4）企业级安全能力。多租户安全保障机制，细粒度授权，数据加密、脱敏，备份恢复。
（5）生态融合。支持多样数据源、生态工具和标准。

基于 MaxCompute 云数据仓库，我们和 BI 工具是如何对接的呢。MaxCompute 主要是一个存储和计算服务，加上一个数据开发平台 DataWorks，组成了一个离线的云数据仓库。在这之上，深度的集成了一个阿里云的 Quick BI。它是一个分析报表工具，直接连接 MaxCompute 的数据表即可以自己对这个表进行分析。还有第三方的一些工具，帆软，Tableau。同时我们在生态这一方面，JDBC 同样也是支持。还有一些企业、一些客户对于商业智能这一块有更加多样化的一个需求或者个性的需求，现有对接的这些工具有可能不支持，那么它也可以通过 SDK 的方式来连接，从而实现基于 MaxCompute 云数据仓库对接的一个商业智能的信息平台。

我们看一下 MaxCompute 离线数仓是怎么实现一个高性能低延迟的分析查询。它可以直接读取离线数仓，支持多样化的查询分析，包括一些简单的查询、复杂的查询、点查询、联邦查询等等。它底层也可以有丰富的数据源，通过 MaxCompute + MC-Hologres 组成一个交互式分析。这么一个大数据生态下，它都可以无缝的对接。比如说 Quick BI，Tableau，帆软。所以它可以做到很快的上手，通过这么一个组合我们可以很快速的实现一个企业的信息平台。

四、实践案例

新零售的一个行业案例，需求背景： 基于 Hadoop 开源生态打造，软硬件维护成本高昂，稳定性问题不断，严重影响业务经营分析；线上业务爆发，需求积压严重，期望有整体解决方案，能够快速灵活支持业务发展所需的技术扩展。通过这么一个大数据解决方案，直接用了阿里云的 Quick BI 这个产品，实现了快速数智化转型，拥抱新零售，降低 TCO 的同时，更好的依托云上生态，实现数据资产业务化闭环。最终新零售这个案例，基于我们的 MaxCompute + DataWorks，提高了他的数据业务的开发效率。

我们再看一个新金融的案例。需求背景：金融业务数据，对安全管控有极强要求，需要一个完整的安全管理体系，同时还要满足个性化安全需求；业务快速发展，需要能快速搭建、成本低、秒级扩展的数据中台体系。我们给客户创造的价值：基于 MaxCompute 开箱即用的应用满足其在安全审计过程中的数据安全需求，缩短了需求响应时间并满足其在数据安全上的个性化需求。

SaaS 模式云数据仓库+AI

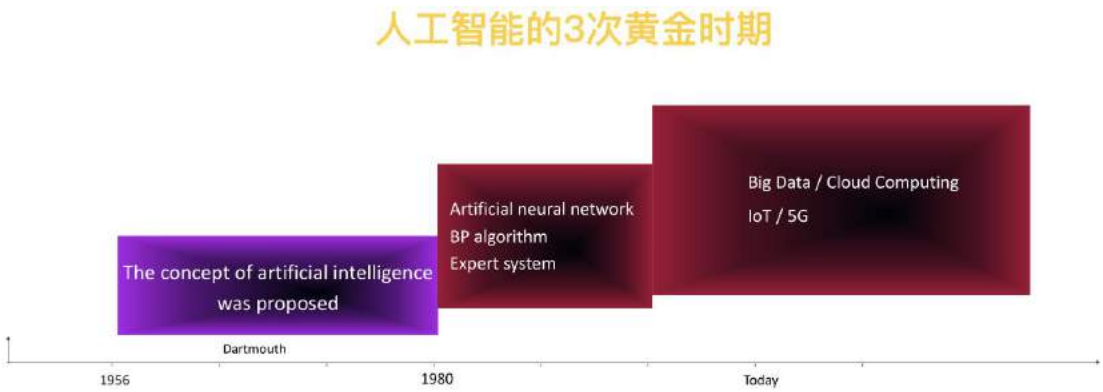
作者 | 孟硕 阿里云智能 产品专家

简介：本文由阿里云计算平台事业部 MaxCompute 产品经理孟硕为大家带来《持续定义 SaaS 模式云数据仓库+AI》的相关分享。

一、Why：概述与价值

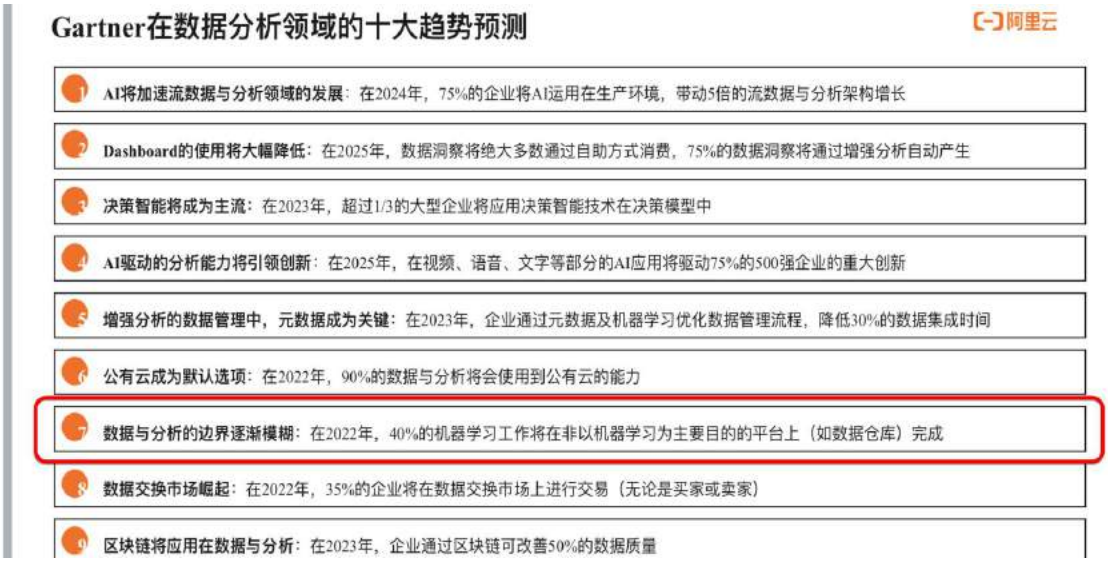
（一）人工智能的发展历史

人工智能是很早就出现的一个概念，起源于上个世纪 50 年代，之后由于种种原因人工智能经历了几十年的漫长的消沉的过程，直到最近几年人工智能才火热起来。人工智能的发展其实有三次黄金时期：第一次是人工智能概念提出的时候，学者们以为 AI 技术能改变世界，但是实际上并没有；第二次是上个世纪 80 年代左右，此时已经提出了神经网络等模拟人脑思考的算法，但是也并没有得到很快的发展；第三次可以认为是从 2010 年左右开始的，与前两次不一样的是这次我们有大数据为生产资料，以强大的算力、云计算为基础设施，包括 IOT 和 5g 技术的发展，有应用场景驱动，比如说搜索就是一个应用人工智能算法的众多场景之一，所以这次是人工智能发展真正的黄金时期



（二）为什么需要 MaxCompute+AI

Gartner 在数据分析领域的是大趋势预测如下：



从中可以看出，Gartner 认为在未来数据与分析的边界逐渐模糊，并且预测在 2022 年，40%的机器学习工作将在非以机器学习为主要目的的平台上（如数据仓库）完成。因此，可以说 MaxCompute+AI 是大势所趋。

因为数据仓库承载的是整个企业的数据资产，尤其是 MaxCompute，它是一个从 TB 到 EB 级，能够弹性扩展大量存储能力的数据平台，所以数据仓库内置机器学习的优势非常明显：

（1）无需移动数据（数据量大），降低基础设施成本、人工成本、减少数据安全风险；
（2）数据访问速度快（让算法找数据）；
（3）可扩展性强；
（4）纯 SQL ML / Python 更易用。

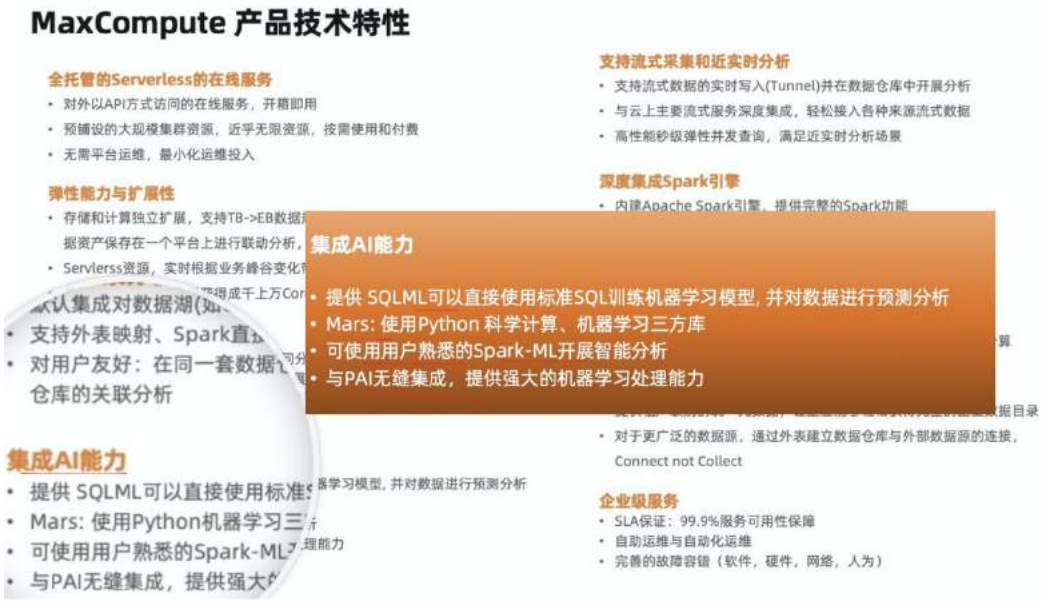
而且数据仓库内置机器学习是各角色均收益的一种集成：对于商务人士来说，新想法可以快速得到快速试验，ROI 得到提升；对于数据科学家和数据分析师来说，大部分工作通过 SQL/Python 实现，易用高效，且模型开发和生产环境可以无缝对接；对于数据库管理员（DBA）来说，数据管理更加简单，安全性更高。

（三）MaxCompute 现有的 AI 能力

MaxCompute 的产品特性在之前的讲座中已经具体讲过了，这里不再赘述，其中 MaxCompute 集成 AI 的能力主要有：

- 提供 SQLML，可以直接使用标准 SQL 训练机器学习模型，并对数据进行预测分析；
- Mars：使用 Python 科学计算、机器学习三方库；
- 可以用用户熟悉的 Spark-ML 开展智能分析；
- 与 PAI 无缝集成，提供强大的机器学习处理能力。

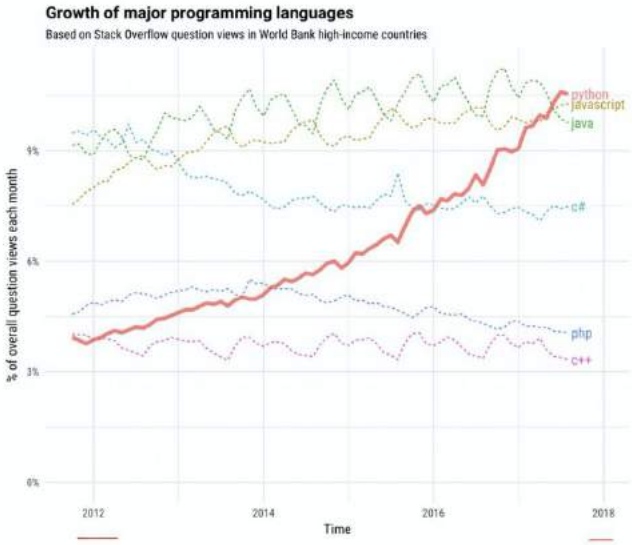
上述的集成 AI 能力中，SQLML 和 Mars 是 MaxCompute 的两个原生 AI 扩展能力，本文我们重点介绍这两个能力。



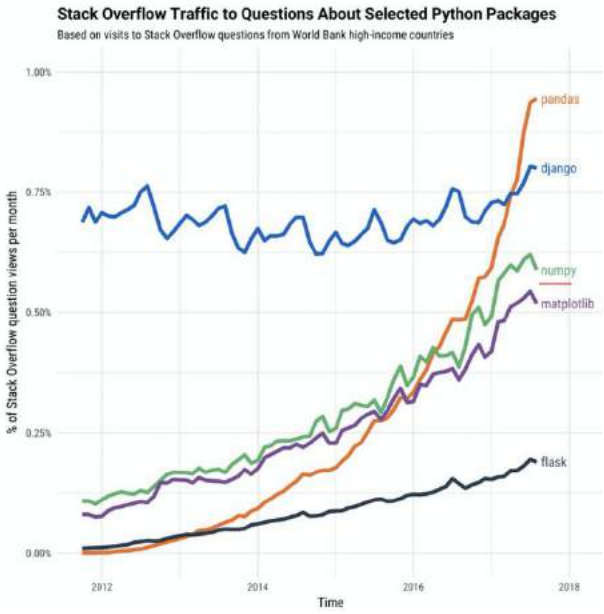
为什么选择 SQL 和 Python 这两种语言呢？主要是因为 SQL 和 Python 是当前数据处理和机器学习领域中最火的两种语言。下面两张图是 SQL 查询语言的发展及现状以及 Python 的发展。

Python has grown to become the dominant language both in data analytics and data science

Image credit to Stack Overflow blogposts



Python has grown to become the dominant language both in data analytics, and general programming



对于数据处理语言来讲，关系型数据库，也就是以 SQL 为基础的关系型数据库，包括类似的数据库目前仍然占据了数据处理引擎的前几名，有着稳健的生态；而 Python 已经逐渐称为数据分析领域和数据科学领域的主流语言，其有着强大的机器学习生态。因此选择这两种语言作为 MaxCompute 的 AI 集成，既是大势所趋，又能减轻使用者的学习成本和迁移成本。

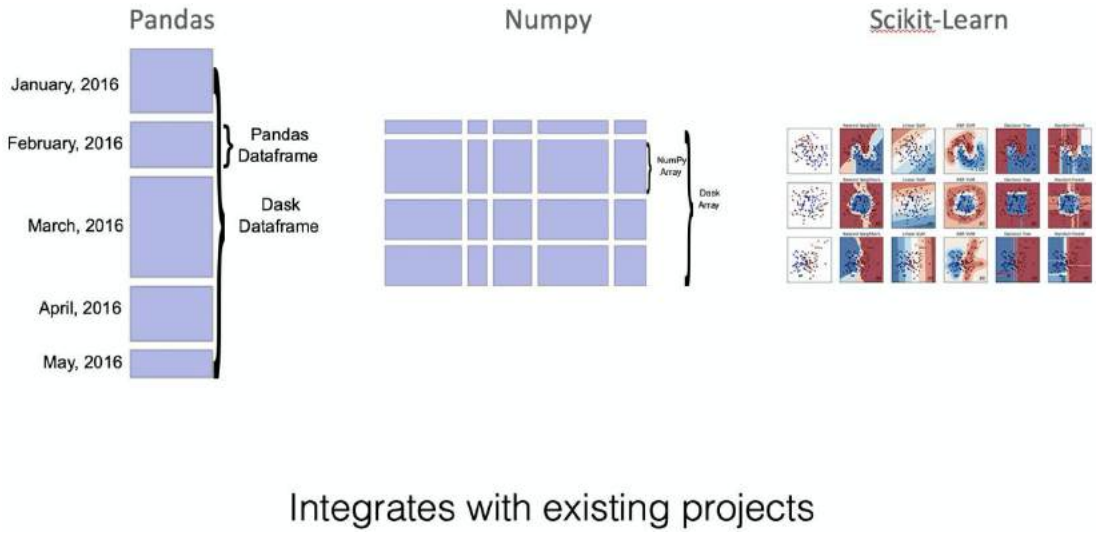
二、What：能力与应用

我们将该项目的名字叫 Mars，其最早是意味着 Matrix 和 array，当然现在已经不再局限于这两者，数据维度可以达到非常高的程度；第二是意味着我们向着比登月更高的目标出发，不断的挑战自己。

那么我们为什么要做 Mars 呢？其主要原因有：

- 为大规模科学计算设计的：传统的大数据引擎编程接口对科学计算不太友好，框架设计也不是为科学计算模型考虑的；
- 传统科学计算大多基于单机，而大规模科学计算需要用到超算，并非普通人所能寄予的能力；
- 传统 SQL 模型科学计算的处理能力不足，做一些简单的科学计算，比如矩阵转置等等，效率也是非常低；
- 目前 R 和 Python 基本上基于单机，其分布式扩展能力比较弱。

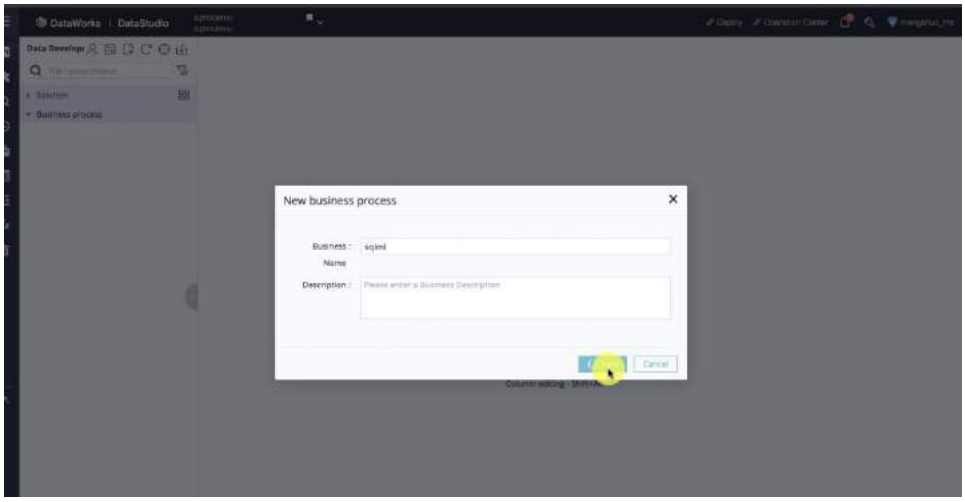
目前，Mars 是唯一的商业化的大规模科学计算引擎，关于 Mars 的更多信息大家可以到阿里云官网查找。Mars 的基本思路如下图所示，主要是将 Python 中的主流科学计算和机器学习的库做相应的分布式化处理。



三、How：最佳实践

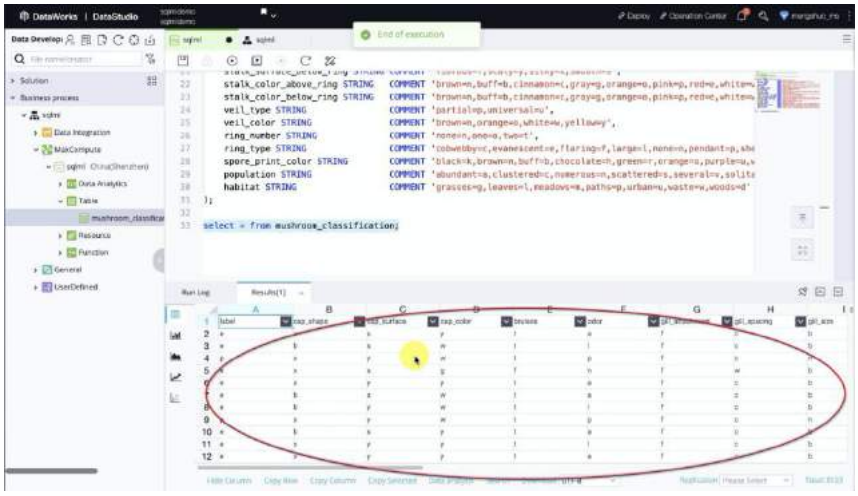
下面是一个简单的 SQLML 的 Demo 介绍。

首先，我们在 DataWorks 中新建一个工作流，会发现工作流中有很多组件，我们先建一个临时查询，如下图所示：

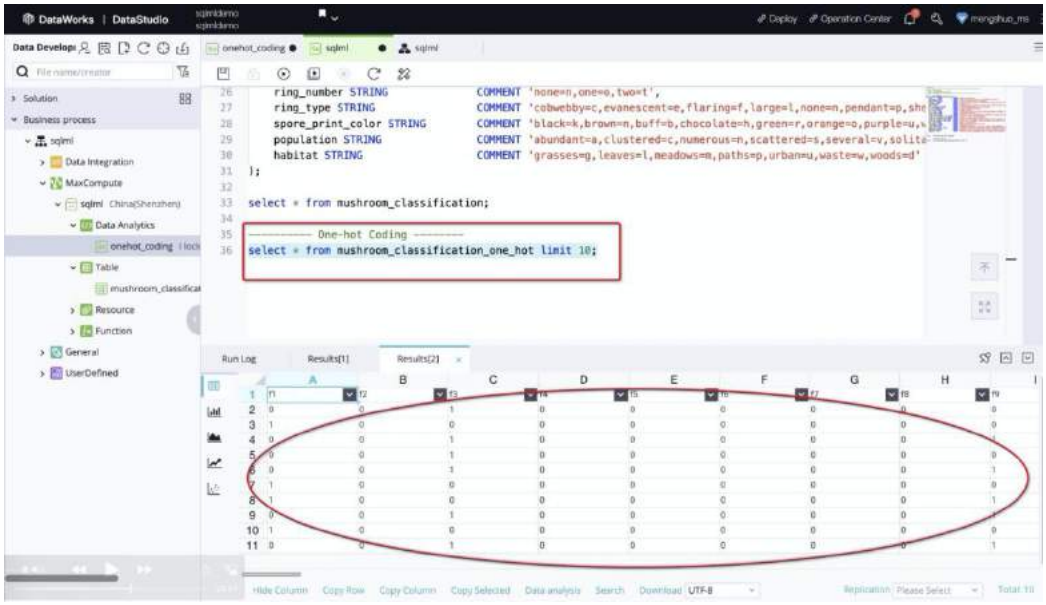


然后新建一张表，其中保存的是关于蘑菇的一些属性，根据这些属性数据，我们可以对其进行分类。

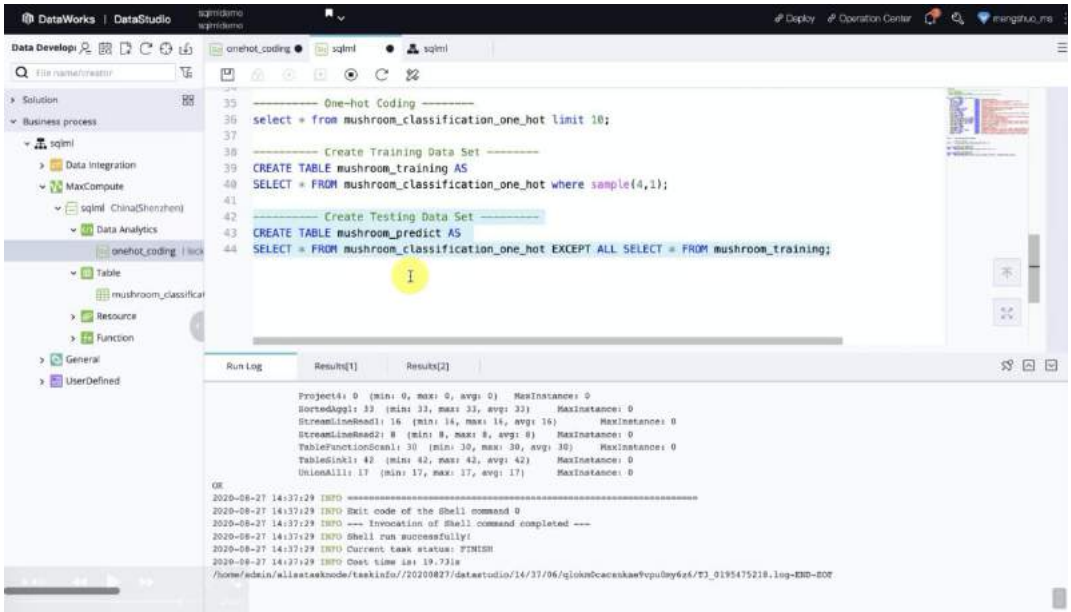
表建立好之后，我们可以将数据导入，因为该数据集比较小，所以我们从本地上传 csv 文件，将列与表中的字段对应即可：



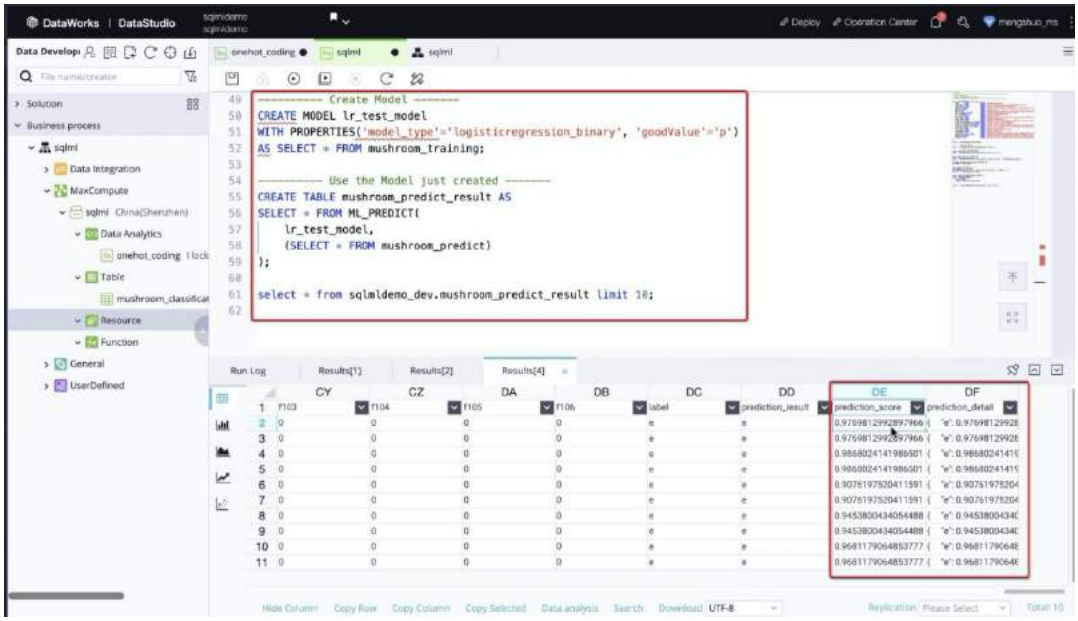
之后，我们需要对特征进行 onehot 编码，其结果如下图所示：



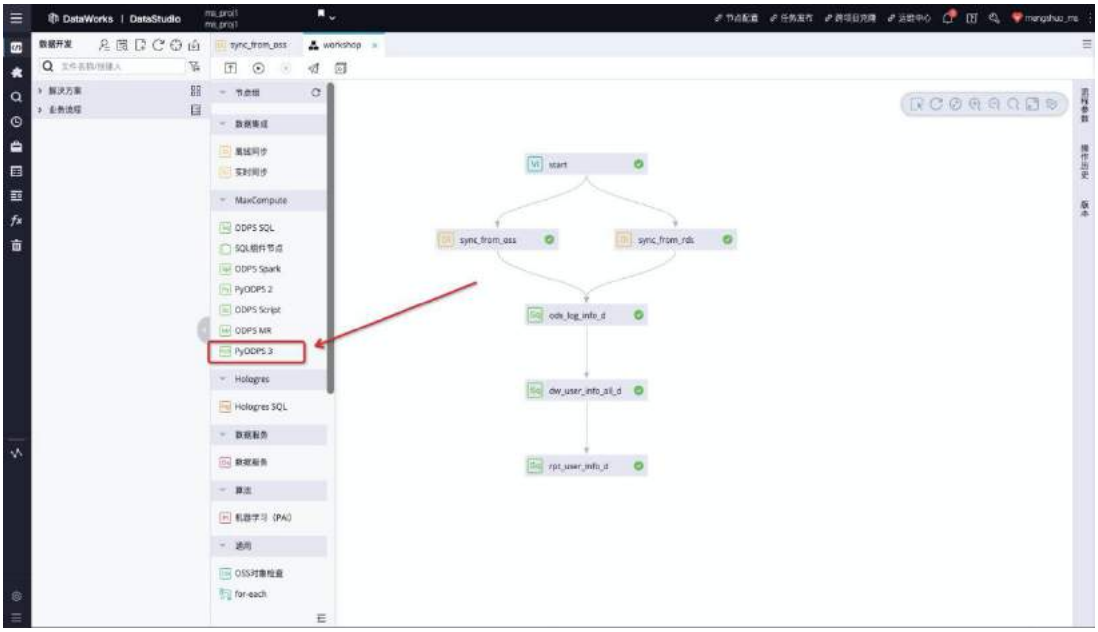
然后，我们将数据分成训练集和测试集，并且分别将训练集和测试集导入一张单独的表中，之后就可以创建模型了，这里我们用的是逻辑回归，一个常用的二分类算法：



运行模型，很便捷地就可以得到训练结果：



通过上面的 Demo，我们很容易的就完成了一次机器学习的训练过程，其过程类似与使用 SQL 中的 UDF，简便、高效。上面 Demo 介绍的是 SQLML，如果想使用 Mars 也非常简单，我们只需要拖拽 PyODPS3 组件即可，如下图所示。



目前，Mars 已经可以试用，SQLML 马上就会和大家见面，欢迎大家进行试用。

SaaS 模式云数据仓库+实时分析

作者 | 孔亮 阿里云智能 产品专家

简介：从实时分析的价值、场景和数据流程，以及用户对平台能力要求展开，讲述云数据仓库 MaxCompute 的产品能力优势 ，面对实时分析场景的能力演进要求。进而以实时分析典型场景的全数据流程处理、建模和分析的最佳实践，讲解 MaxCompute+MC-Hologres 的解决方案，展现强强组合应的能力优势。

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。

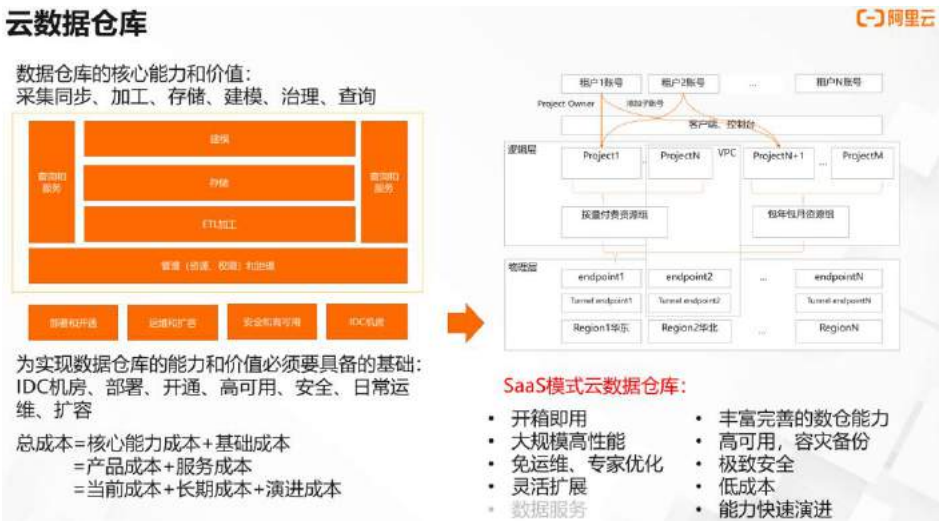
数据仓库的核心能力和价值包括：采集同步、加工、存储、建模、治理、查询。但是为了实现数据仓库的能力和价值必须要具备的基础包括：IDC 机房、部署、开通、高可用、

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。

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。

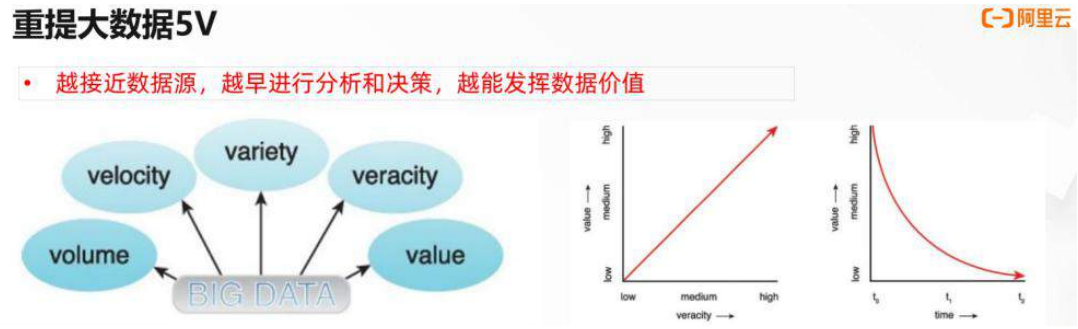
（2）速率（Velocity）即数据生成、流动速率快。数据流动速率指指对数据采集、存储以及分析具有价值信息的速度。因此也意味着数据的采集和分析等过程必须迅速及时。

（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。

（1）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。

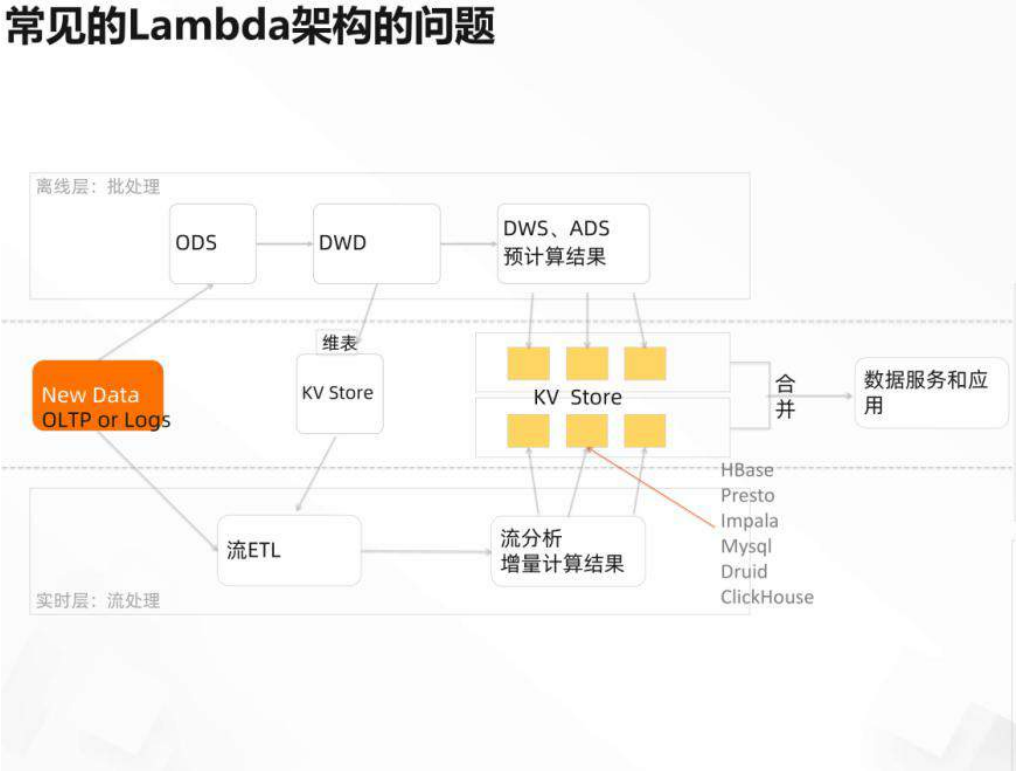
（2）环环相扣、多套系统、运维复杂、成本高昂：

- 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。

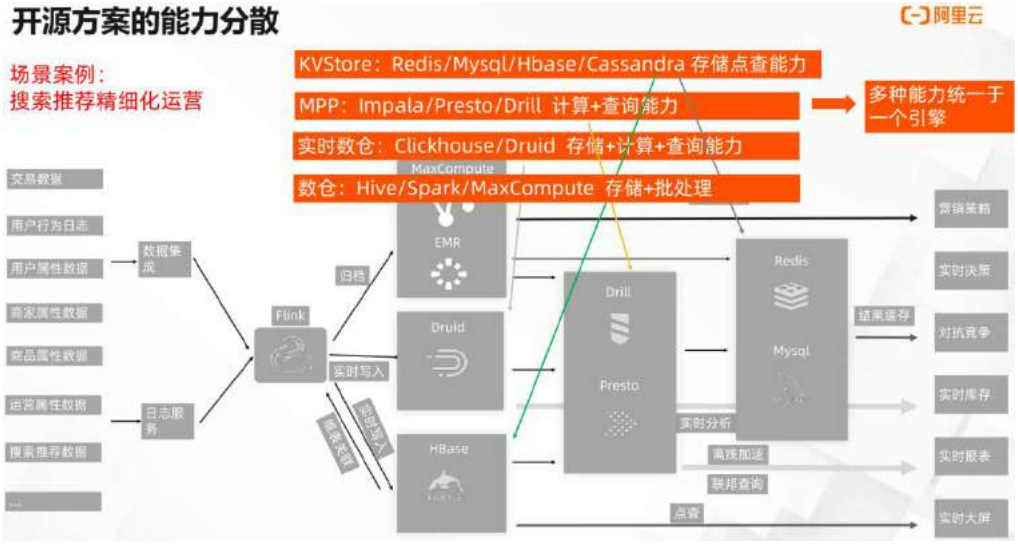
（3）开发周期长、业务不敏捷：

- 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。

常见的Lambda架构的问题



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。MC-Hologres 即是这个产品和解决方案。



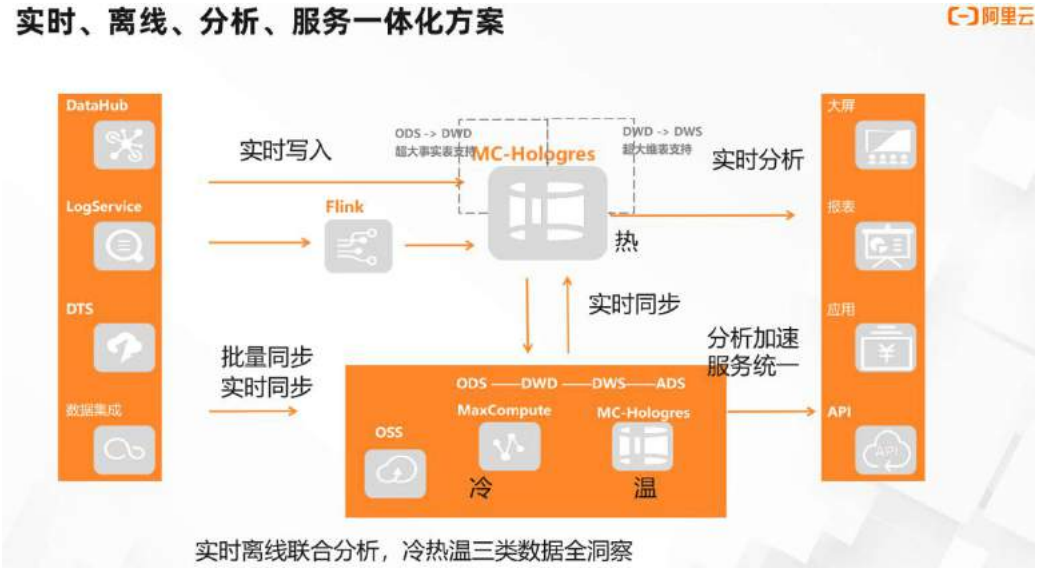
MC-Hologres 将实时分析的架构变得简单和高效。以实时分析为中心设计，MC-Hologres 能够实现实时写入和实时分析、查询。MC-Hologres 提出云原生 HSAP 架构中，一份数据同时用于实时分析、在线服务和实时离线数据统一存储，与 SaaS 模式云数据仓库 MaxCompute 完美结合。

另一种场景，MC-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，但是施难度相对较高，尤其是对于数据回放部分。

如下图所示，MC-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 MC-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。

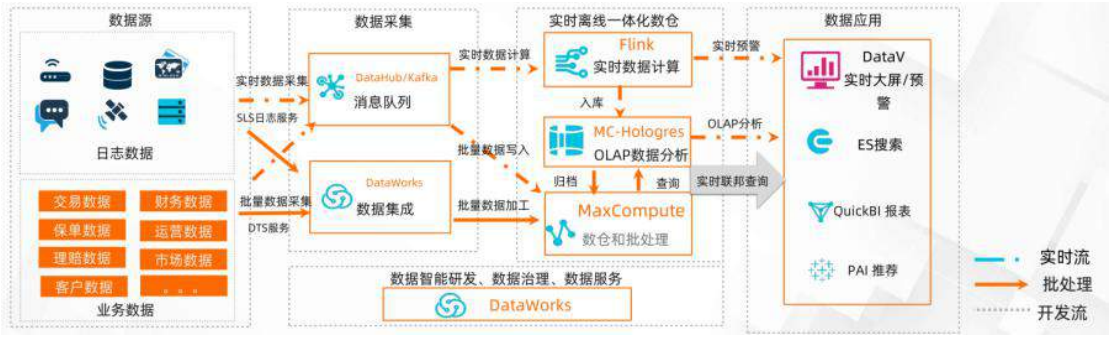
方案优势：阿里巴巴最佳实践的大数据平台，

（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 MaxCompute 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。

SaaS 模式云数据仓库+实时搜索

作者 | 孟硕 阿里云智能 产品专家

简介：本文由阿里云计算平台事业部 MaxCompute 产品经理孟硕为大家带来《持续定义 SaaS 模式云数据仓库+实时搜索》的相关分享。以下是视频内容精华整理，主要包括以下三个部分：1.Why：概述与价值；2.What：应用场景；3.How：最佳实践。

一、Why：概述与价值

（一）MaxCompute

我们把 MaxCompute 定义为 SaaS 模式的企业级云数据仓库。在之前，我们可能会认为 MaxCompute 是一个离线数据处理引擎，也就是一个传统的数仓，但 MaxCompute 所能做的事情要比传统数仓多的多。因此，我们更倾向于把 MaxCompute 看成一个数据处理的平台，在它上面我们可以做离线数据的处理，包括数据库的应用，传统数据仓库的应用，以及近实时的数据采集和近实时的数据查询，现在将其与 MC-Holegres 组件结合，我们还能做到实时数仓的应用场景。

MaxCompute 是阿里云的一个托管服务，它依托于阿里云强大的基础设施，为用户提供优质、便捷的服务，其架构如下图所示。

MaxCompute 有着广泛的应用场景，传统数仓所能做的，MaxCompute 都能做，主要包括：

- 广告场景：用户标签计算、分析等；
- 业务运营场景：交互式业务指标计算、查询等；
- 各行业搭建数据仓库，比如流批一体、湖仓一体等；
- 云上弹性扩展大数据计算和存储。

得益于可靠的架构和强大的技术实力，MaxCompute 有着非常优秀的产品技术特性，主要包括：

（1）全托管的 Serverless 的在线服务

- 对外以 API 方式访问的在线服务，开箱即用；
- 预铺设的大规模集群资源，近乎无限资源，按需使用和付费；
- 无需平台运维，最小化运维投入。

（2）弹性能力与扩展性

- 存储和计算独立扩展，支持 TB 到 EB 级别数据规模的扩展能力，可以让企业将全部数据资产保存在一个平台上进行联动分析，消除数据孤岛；
- Serverless 资源按需分配，实时根据业务峰谷变化带来的需求变化分配资源，自动扩展；
- 单作业可根据需要秒级获得成千上万 Core。

（3）数据湖探索分析

- 默认集成对数据湖（如 OSS 服务）的访问分析，处理非结构化或开放格式数据；
- 支持外表映射、Spark 直接访问方式开展数据湖分析；
- 对用户友好：在同一套数据仓库服务和用户接口下，实现数据湖分析和数据仓库的关联分析。

（4）集成 AI 能力

- 与阿里云机器学习平台 PAI 无缝集成，提供强大的机器学习处理能力；
- 可使用用户熟悉的 Spark-ML 开展智能分析；
- 提供 SQLML 可以直接使用标准 SQL 训练机器学习模型，并对数据进行预测分析；
- Mars：使用 Python 机器学习第三方库。

（5）支持流式采集和近实时分析

- 支持流式数据的实时写入（Tunnel），并在数据仓库中开展分析；
- 与云上主要流式服务深度集成，轻松接入各种来源流式；
- 高性能秒级弹性并发查询，满足近实时分析场景。

（6）深度集成 Spark 引擎

- 内建 Apache Spark 引擎，提供完整的 Spark 功能；
- 与 MaxCompute 计算资源、数据和权限体系深度集成。

（7）统一而丰富的运算能力

- 离线计算(MR，DAG，SQL，ML，Graph)；
- 实时计算(流式，内存计算，选代计算)；
- 涵盖通用关系型大数据，机器学习，非结构化数据处理，图计算。

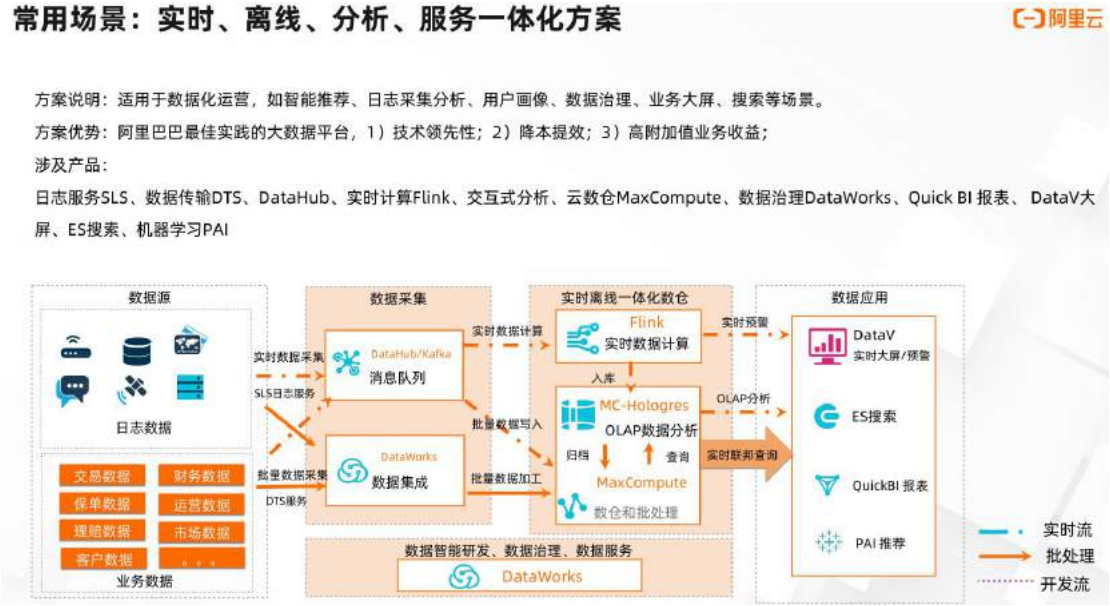
（8）提供统一的企业数据视图

- 提供租户级别的统一元数据,让企业能够轻松获得完整的企业数据目录；
- 对于更广泛的数据源，通过外表建立数据仓库与外部数据源的连接，Connect not Collect。

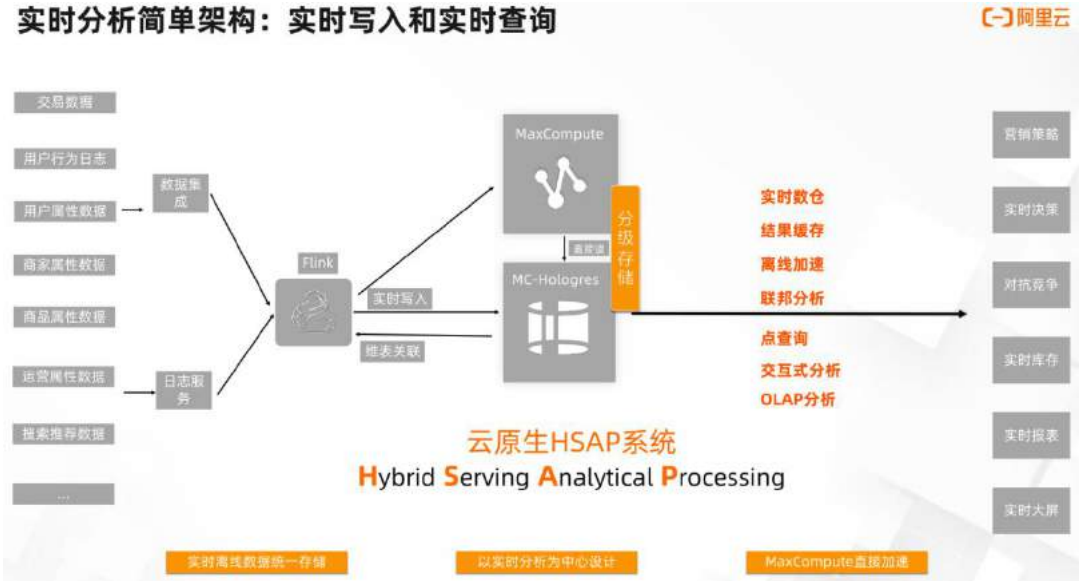
（9）企业级服务

- SLA 保证：99.9%服务可用性保障；
- 自助运维与自动化运维；
- 完善的故障容错（软件，硬件，网络，人为）机制。

一般来讲我们的大数据项目是需要很多个组件才能完成的，包括离线组件和实时组件。下图一个常用的场景，它是集实时、离线、分析、服务于一体的一套方案，适用于数据化运营，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。这套方案是阿里巴巴最佳实践的大数据平台，具有技术领先性，降本提效，高附加值业务收益等优势。当然，整个方案涉及到的产品也非常多，包括日志服务 SLS、数据传输 DTS、DataHub、实时计算 Flink 等等，具体如下图所示。



下图是 MaxCompute 和 MC-Hologres 两个组件融合之后的实时分析简单架构，即云原生 HASP 系统，通过该架构我们可以实现实时写入和实时查询。与其他的 OLAP 应用不同的是这种架构下 MC-Hologres 和 MaxCompute 是一体的，可以共享存储，也就是说 MC-Hologres 可以直接读取 MaxCompute 的数据，大大降低了存储成本。通过这两个组件，我们还可以解决离线加速、联邦分析、交互式分析等问题。

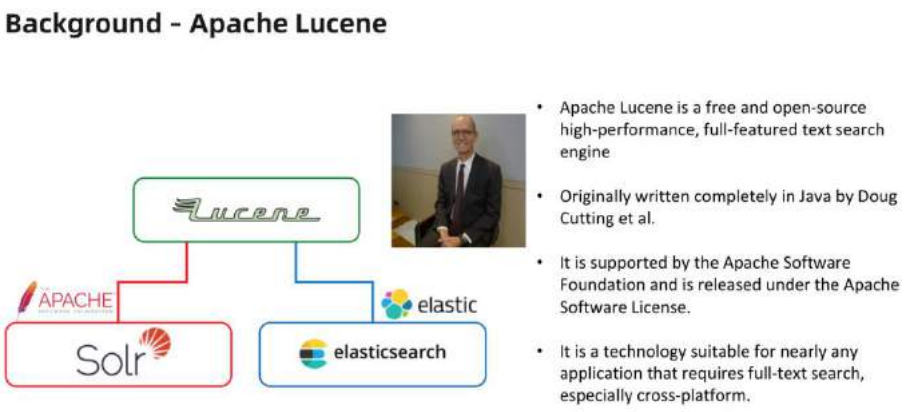


（二）Elasticsearch

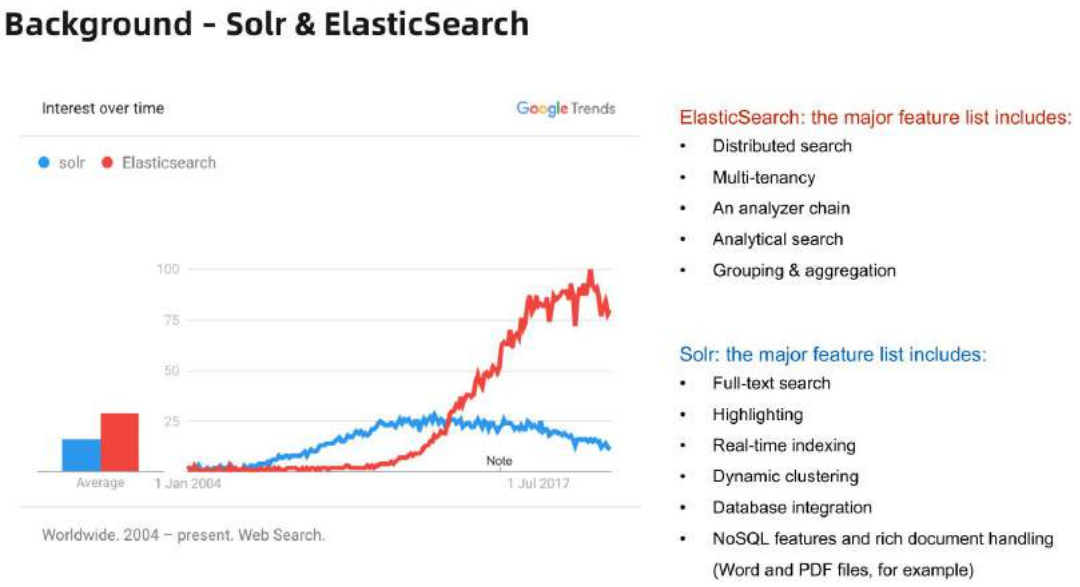
Elasticsearch 是一个开源的分布式、RESTful 风格的搜索和数据分析引擎，它的底层是开源库 Apache Lucene。Elasticsearch 解决了 Lucene 使用时的繁复性，功能强大，使用简单，能够提供实时搜索服务。Elasticsearch 应用场景广泛，比如打车的场景中（例如滴滴打车），使用查询附近的车辆等功能时候，后台就是 Elasticsearch 在为搜索做支撑，又比如在 Github 中，Elasticsearch 可以帮助我们利用关键字等在站内进行检索。当然，不只是网站应用，包括手机 APP，只要用到站内搜索服务，都能够用到 Elasticsearch 或者其他的搜索服务应用。

我们为什么需要搜索引擎呢？实时搜索为什么现在这么火呢？之前我们在做数据分析的时候，可以通过写程序的方式，但是写程序对于一些数据分析师来说是一个高门槛的任务，需要一定的学习成本，包括使用 SQL 也有一定的学习成本。但是有了搜索引擎之后，我们只需要按照一定的条件进行筛选就可以得到我们想要的信息，大大降低了学习成本。

当前，主流的搜索引擎有两个：Solr 和 Elasticsearch，两者都基于 Lucene 发展而来。Lucene 是当今最先进，最高效的全功能开源搜索引擎框架，但是 Lucene 只是一个框架，且比较复杂，要充分利用它的功能，需要在其基础上进行扩展开发，因此有了 Solr 和 Elasticsearch。



下图是 Google Trend 中两个搜索应用的趋势分析，可以看出在实时搜索领域，近几年 Elasticsearch 的热度已经超过了 Solr，因为在实时搜索领域 Elasticsearch 的效果要好于好于 Solr，但是不得不提的是 Solr 在现有数据的基础上进行查询搜索的速度会更快一些。



目前，Elastic 已经宣布与阿里云建立长期合作和战略伙伴关系。未来阿里云 Elasticsearch 将会兼容开源 Elasticsearch 的功能，以及 Security、Machine Learning、Graph、APM 等商业功能，致力于数据分析、数据搜索等场景服务，与 Elastic 合作，共同为客户提供企业级权限管控、安全监控告警、自动报表生成等场景服务。

（三）为什么需要 MaxCompute+实时搜索

- MaxCompute 能解决日增量数据超大场景，提供强大算力

  MaxCompute EB 级别 Saas 云数仓，支持流式采集

- 需要更多的计算模型做处理

  MaxCompute (SQL/Spark/Python/Java/ML) ->高维度聚合查询、like、分词 (Lucene Query)

- 辅助 ElasticSearch 做数据存储

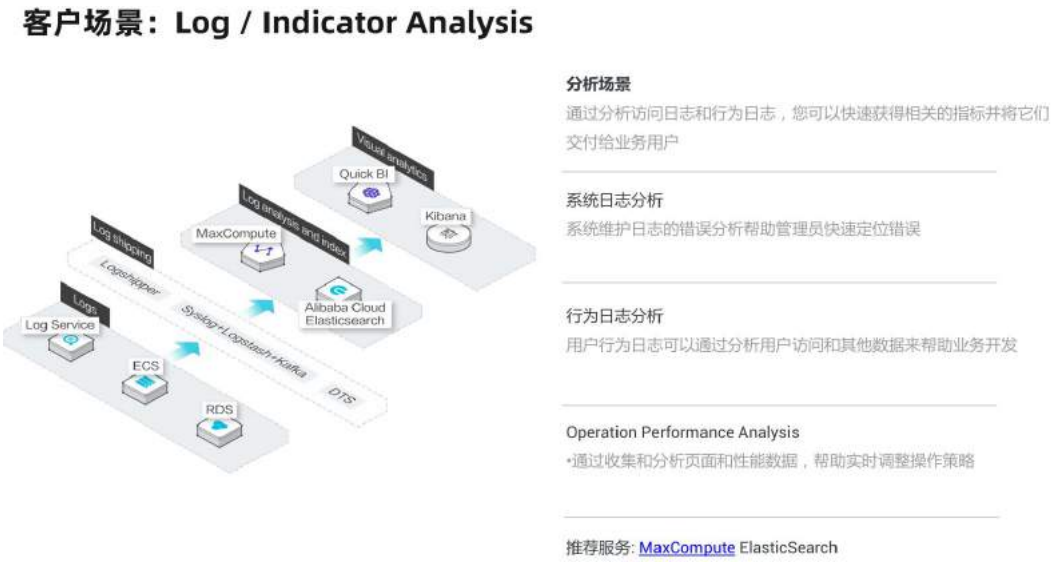
  MaxCompute + OSS / OTS

二、What：应用场景

当前，实时搜索的主要应用场景有三个：

（1）日志和指标分析（Log/Indicator Analysis）；
（2）安全（Security）；
（3）站内检索（WebHosting）。

其场景的应用逻辑如下面三张图所示。



三、How：最佳实践

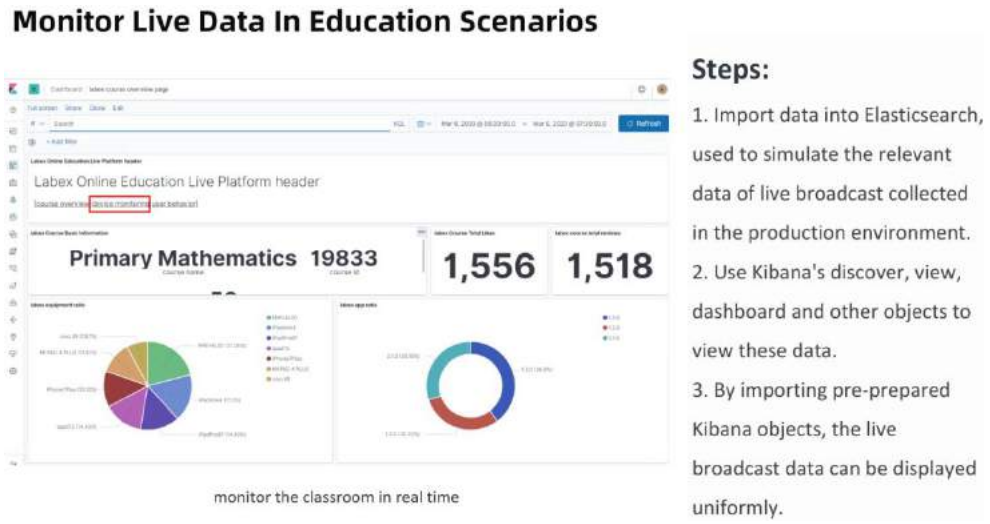
最佳实践主要包括数据集成和数据监控两部分，其中数据集成指的是 MaxCompute 和 Elasticsearch 两个组件之间怎么做数据交互。

（一）数据集成

下图是一个在线教育的案例，该案例大的背景是要监控企业内部包括用户的 C 端产生的日志，还有内部的服务端产生的日志，它由 MaxCompute 做数据的预分析，然后交由 Elasticsearch 做数据监控，其痛点主要有如下三点：

（1）监控数据实时性要求高；
（2）流量波动大，自建集群投入成本高；
（3）数据权限粒度要求高。

对于以上痛点，一般通用的解决方案如上图所示，包括数据采集和集中、数据 ETL、数据分析和展现三部分，最后会得到如下图所示的 DashBoard。



MaxCompute 和 Elasticsearch 之间的数据交互是非常重要的一部分，将 MaxCompute 的数据导入 Elasticsearch 主要分为如下五步：

（1）准备工作

创建 DataWorks 工作空间并开通 MaxCompute 服务，准备 MaxCompute 数据源、创建阿里云 Elasticsearch 实例。

（2）步骤一：购买并创建独享资源组

购买并创建一个数据集成独享资源组，并为该资源组绑定专有网络和工作空间，独享资源组可以保障数据快速、稳定地传输。

（3）步骤二：添加数据源

将 MaxCompute 和 Elasticsearch 数据源接入 DataWorks 的数据集成服务中。

（4）步骤三：配置并运行数据同步任务

配置一个数据同步的脚本，将数据集成系统同步成功的数据存储到 Elasticsearch 中，然后将独享资源组作为一个可以执行任务的资源，注册到 DataWorks 的数据集成服务中，这个资源组将获取数据源的数据，并执行将数据写入 Elasticsearch 中的任务（该任务将有数据集成系统统一下发）。

（5）步骤四：验证数据同步结果

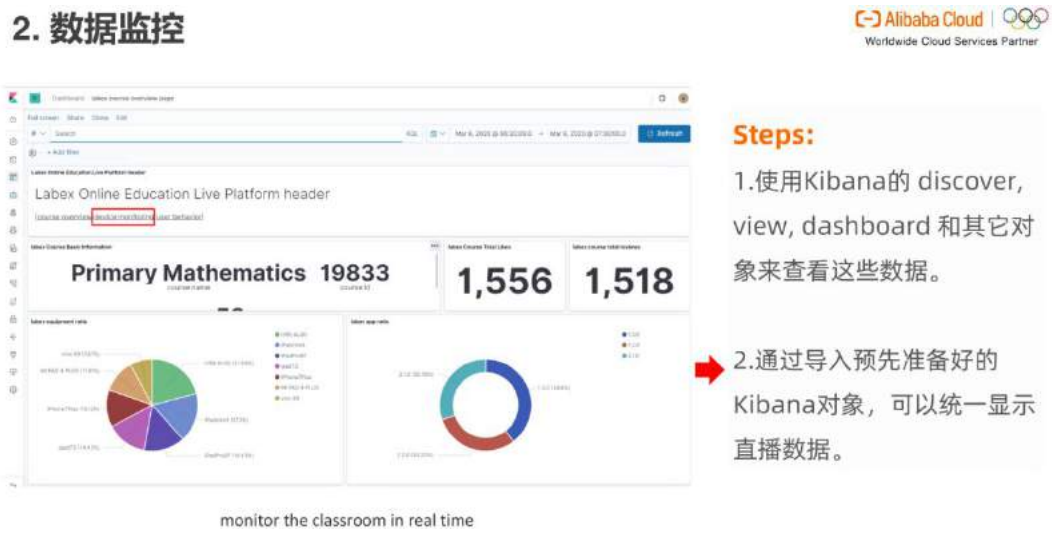
在 Kibana 控制台中，查看同步成功的数据，并按条件查询数据。

（二）数据监控

经过上面的步骤，数据已经导入了 Elasticsearch，我们做数据监控主要有如下两步：

（1）使用 Kibana 的 discover，view，dashboard 和其他对象来查看这些数据；

（2）通过导入预先准备好的 Kibana 对象，可以统一显示直播数据。



SaaS 模式云数据仓库+数据银行

作者 | 隆志强 阿里云智能 高级产品专家

简介：本文将介绍 SaaS 模式云数据仓库 MaxCompute，如何助力数据银行 SaaS 模式云战略和一体化数据开放场景介绍。

一、云数据仓库

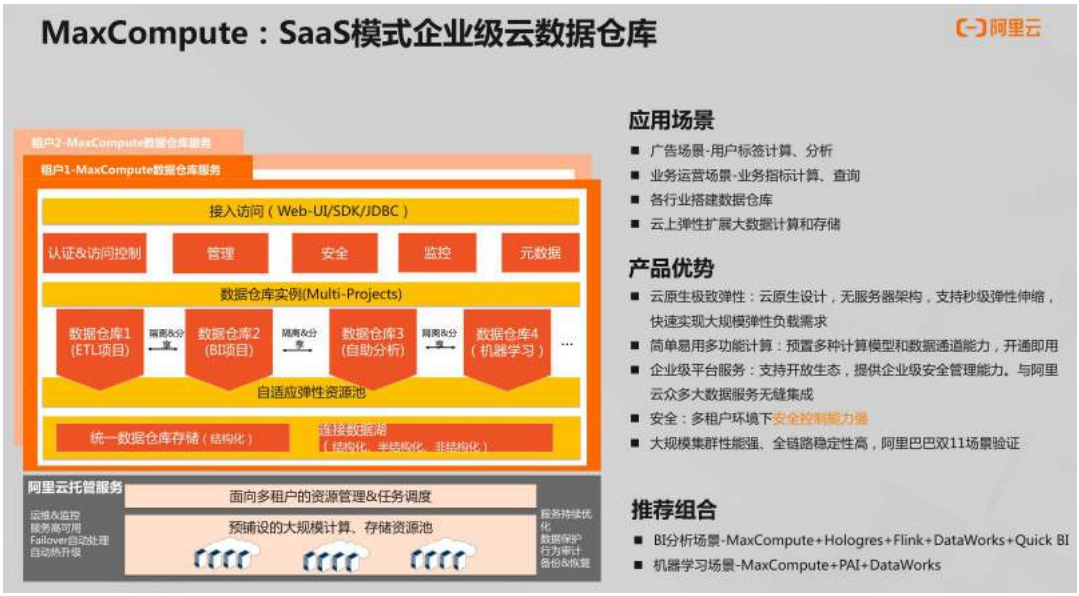
MaxCompute：SaaS 模式企业级云数据仓库的应用场景包括：

广告场景-用户标签计算、分析；
业务运营场景-业务指标计算、查询；
各行业搭建数据仓库；
云上弹性扩展大数据计算和存储。

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。

推荐组合包括 BI 分析场景和机器学习场景，分别为 MaxCompute+MC-Hologres+Flink+DataWorks+Quick BI，以及 MaxCompute+PAI+DataWorks。



MaxCompute 算力资源产品解决方案如下图所示。

（1）包年包月

满足常规需求，稳定财务支出；

支持作业优先级，保障关键任务稳定产出；

支持存储与计算资源包购买。

（2）按需使用

无服务器架构，超大规模的存储和计算扩展能力；

自动匹配业务需求，完美适配业务的高速变化；

不使用不付费。

（3）多计算资源打通

融合打通包年包月与按需使用的弹性资源，只需联合开通，即可实现更优的成本与性能平衡的资源解决方案。

（4）抢占空闲资源

非预留计算资源，抢占并使用服务空闲计算资源，价格较包年包月标准计算资源下降 74%。

安全事件频发，云上大数据服务如何保障企业数据和服务安全。MaxCompute 构建全面、多层次的安全管理能力，持续保护云上数据及服务安全。包括 MaxCompute 安全生态，平台系统安全，基础设施安全三大部分。

飞天大数据平台解决方案适用于电商、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。

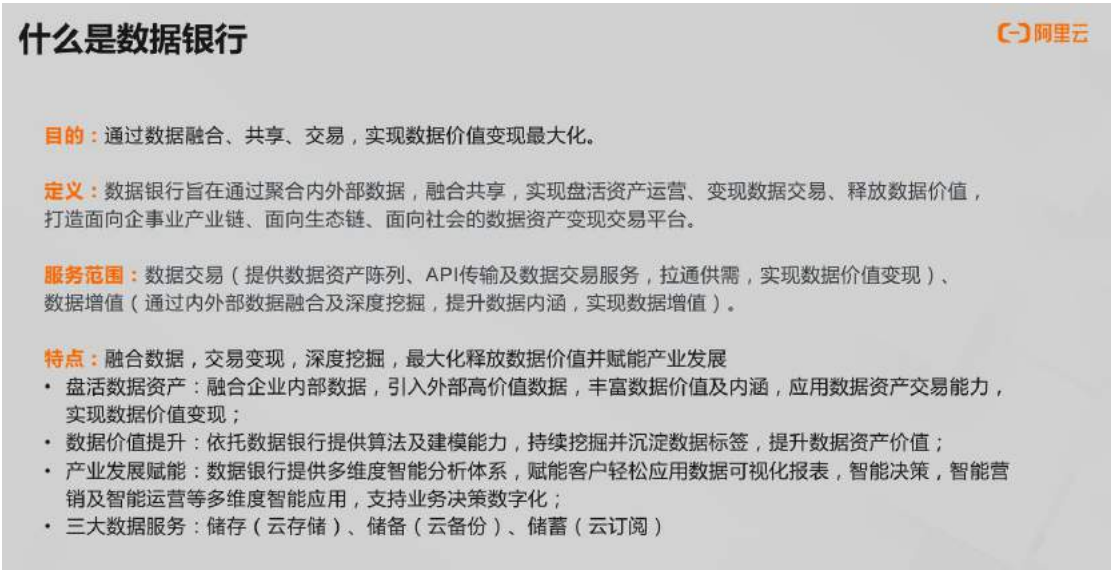
二、数据银行

数据银行旨在通过聚合内外部数据，融合共享，实现盘活资产运营、变现数据交易、释放数据价值，打造面向企事业产业链、面向生态链、面向社会的数据资产变现交易平台。

目的是通过数据融合、共享、交易，实现数据价值变现最大化。

服务范围包括数据交易（提供数据资产陈列、API 传输及数据交易服务，拉通供需，实现数据价值变现）和数据增值（通过内外部数据融合及深度挖掘，提升数据内涵，实现数据增值）。

特点是融合数据，交易变现，深度挖掘，最大化释放数据价值并赋能产业发展。其中，包括盘活数据资产，数据价值提升，产业发展赋能，以及三大数据服务，具体内容如下图所示。



行业应用架构-友盟，具体架构如下图所示。

三、MaxCompute+数据银行

主题数据包及数据来源包括三个部分。

（1）统计分析；

（2）开发者工具；

（3）营销增长。

我们如何共享、转让、公开披露您以及您最终用户的个人信息。

（1）共享。承担保密义务，不会为满足第三方的营销或非法目的而向其出售或出租您的信息，会与这些合作伙伴合作以多种形式将经 U-DIP 数据中台处理、加工后的脱敏数据用于包括优化广告投放和提升营销效果等商业化使用。

（2）转让。不会向任何第三方转让您以及您最终用户的个人信息。

（3）公开披露。

- 获得您或您最终用户明确同意；
- 基于法律的披露：在法律、法律程序、诉讼或政府主管部门强制性要求的情况下，我们可能会公开披露您或您最终用户的个人信息；
- 在紧急情况下，经合理判断是为了保护我们、我们的客户、最终用户或其他人的重要合法权益。

友盟数据银行已实现产品功能和价值“一键通”模式。一体化消费体验包括三个部分。

（1）主题数据包。每日高性能采集加工海量数据，自动生产 APP/WEB/小程序/广告/PUSH 主题数据包。

（2）一键数据订阅开放。与 MaxCompute(DataWorks) 云数据仓库无缝对接，一键订阅数据。

（3）主题分析模板与自助分析。预置分析模板和拖拽式自助分析能力，业务人员无需麻烦开发跑数即可完成分析。

友盟数据银行通过和 MaxCompute 共创带来的客户体验提升，如下图所示。从账号登陆，到应用配置，现在比过去更加智能，更加便捷。

开放多端、多主题的明细数据与指标数据，为开发者构建私域数据体系。

指标数据开放，将友盟+9 年行业经验沉淀回馈于开发者：

（1）实时指标大屏展示；

（2）多维指标分析监控。

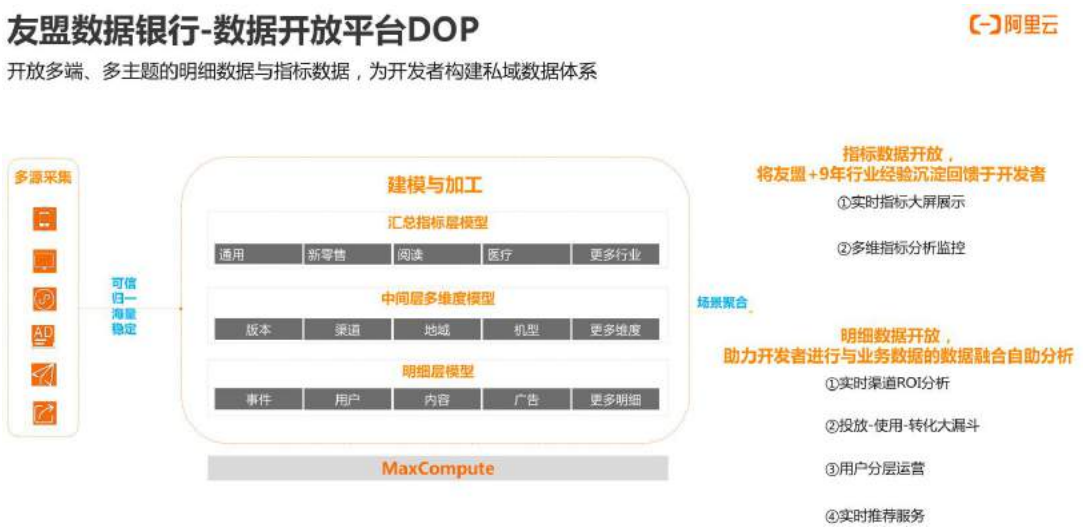
明细数据开放，助力开发者进行与业务数据的数据融合自助分析：

（1）实时渠道 ROI 分析；

（2）投放-使用-转化大漏斗；

（3）用户分层运营；

（4）实时推荐服务。



友盟数据银行支持云上数仓无缝链接，为开发者提供一键式数据模型体系开放的体验。开发者云上数仓，高性价比交互式查询服务，兼容接入异构数据源进行查询和分析。为您提供快速、完全托管的 PB 级数据仓库解决方案，经济并高效的批量分析海量数据。

四、案例介绍

第一个案例：本地生活行业客户，业务数据化+数据可视化。

客户：本地生活类，智慧社区服务平台。

痛点：数据化运营程度低，数据分散，业务人员的数据需求实现周期长。

实施方案：

（1）规范化的多端数据采集。基于业务需求梳理进行埋点方案设计，APP、H5、小程序等多端 SDK 采集。

（2）实时数据和离线数据的订阅返还。经过友盟统一 ETL 服务的采集数据分别投递至客户 SLS（实时）、DLA （离线）。

（3）数据报表设计与开发。离线数据自动联通 QBI，除 4 个预置看版外，根据具体业务需求搭建业务分析监测。

方案结果：

（1）业务数据化。多端采集行为数据纳入数仓体系建设。

（2）数据可视化。日常数据监测看版，让业务人员快速看到产品迭代、运营动作的效果。

第二个案例：游戏行业客户，多源数据融合。

客户：独立游戏工作室。

痛点：APP 行为数据与后台业务数据割裂。

实施方案：

（1）数据采集。使用游戏行业埋点方案进行 APP 端数据采集，获取多种用户识别 ID。

（2）数据迁移。存入其他云厂商的用户付费、广告收入等数据迁移入阿里云。

（3）数据融合。采集行为数据一键投递至阿里云数据库，通过用户唯一识别将数据融合。

方案结果：数据融合分析。结合用户留存行为和收入数据，测算用户生命周期价值，判断渠道回本周期、渠道投放优选。

加入 MaxCompute 开发者社区
扫码关注获取更多资讯

阿里云开发者“藏经阁”
海量免费电子书下载