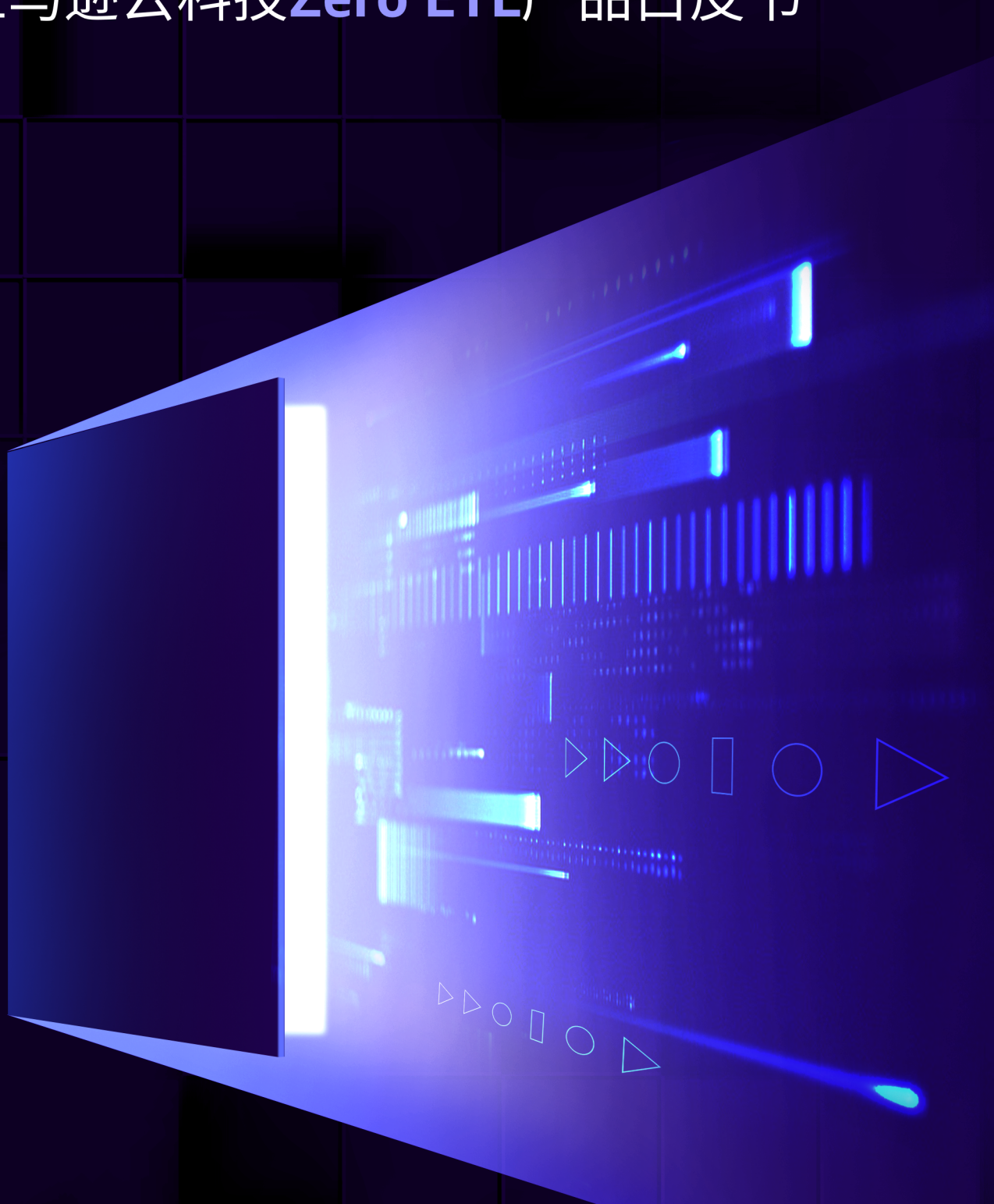
亚马逊云科技

数据工程师的未来照进现实

亚马逊云科技Zero ETL产品白皮书



亚马逊云科技

目录

CONTENTS

Zero ETL
——数据工程师的未来照进现实

ETL 是将业务系统的数据经过提取(Extract)、转换清洗(Transform)和加载(Load)到数据仓库、大数据平台的过程，目的是将企业中的分散、零乱、标准不统一的数据整合到一起，为企业的决策提供分析依据，ETL 是各类数据创新项目(比如 BI 辅助决策，反欺诈与内部合规项目等)重要的一个环节。通常情况下，在 BI 项目中 ETL 会花掉整个项目至少 1/3 的时间。自传统数据仓库理论形成,动辄占据数据工程师 70% 工作量的 ETL 构建与维护已经成为常识。每当一个数据项目开展，背后就会牵引出复杂的 ETL 工作，且这样的工作随着项目的开展就像一个工作量黑洞一样吞噬着项目组的资源。

图：ETL-一个不受欢迎的、不可持续的黑洞

而到了大数据时代，本以为可以找到新的解决方案，但是这种痛苦不减反增，这些痛苦体现在：

- 市面上的 ETL 工具多，代表选择多，那么第一步，ETL 工具选型就是个难题。
- ETL 任务的多少，往往和项目的复杂度呈“指数级正相关”，越复杂的项目，调度任务越多，动辄数千个 ETL 任务的项目已屡见不鲜，任务调度与排查的复杂与维护是一个巨大的难题。

图：“天下苦 ETL 久矣”

02

什么是 Zero ETL?

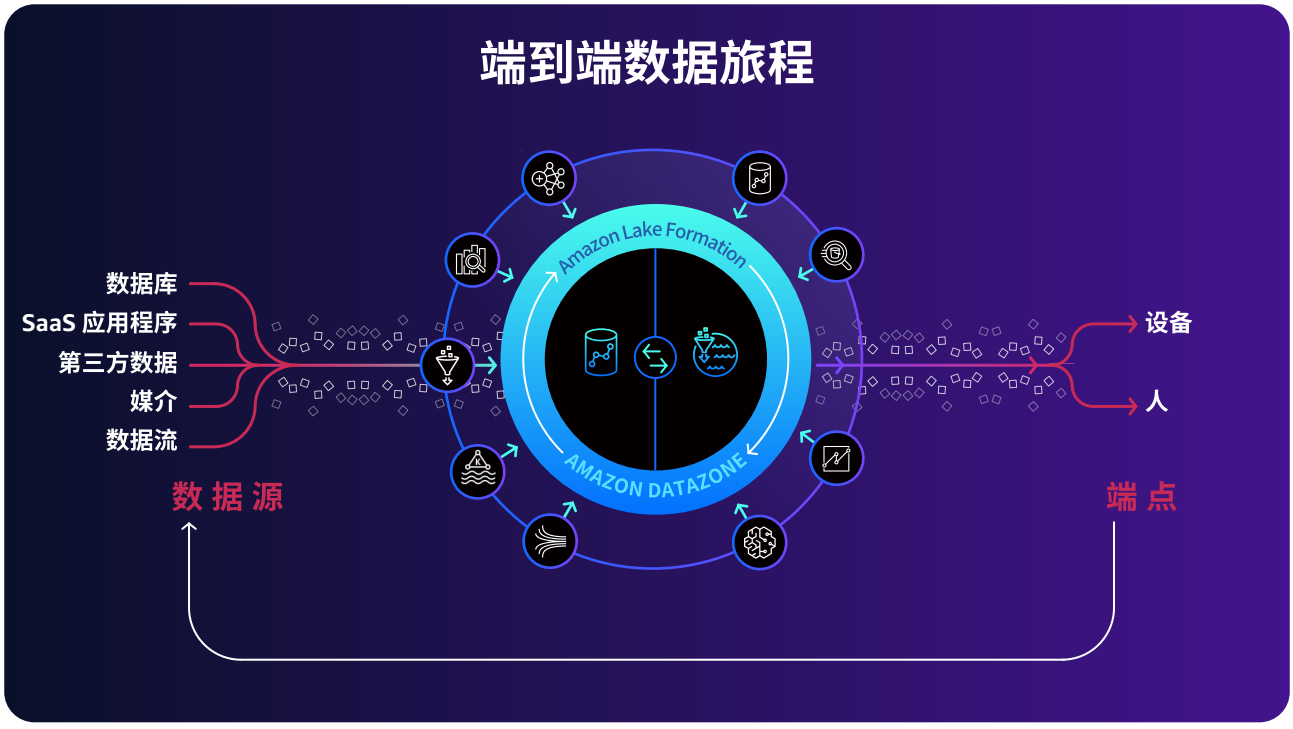
在 2022 re:Invent 全球大会**亚马逊云科技数据库、分析和机器学习副总裁 Swami Sivasubramanian** 表示：

“当前，客户管理的数据既庞大又复杂，这意味着他们不能只用单一技术或几个工具来分析和探索这些数据。我们的许多客户都通过亚马逊云科技的众多数据库和分析服务从数据中提取价值。确保他们能够使用正确的工具完成工作，对于他们的企业成功非常重要。

今天发布的新功能帮助我们的客户在亚马逊云科技上迈向一个 ‘Zero ETL 的未来’，借助 Zero ETL 减少在不同服务间手动迁移或转换数据的工作。无论企业和数据的规模有多大，复杂度有多高，通过为客户消除 ETL 和其它数据迁移任务，助力客户专注于分析数据，面向业务获取新的洞察。”

企业需要全面了解其业务的真实情况，才能让数据帮助企业在整个价值流程之中创造价值。数据一体化融合需要让企业打破数据孤岛，并以一种一体化的方式实现数据的共享与安全访问，以解锁不同企业用户和不同目的的数据价值。企业可以通过“智能湖仓”架构实现这一目标，将湖、仓、库连接成为一个整体，通过专门构建的数据分析服务实现“用正确的工具完成正确的任务”，进而实现优势整合与成本效益最大化。任何阶段的企业都可以从这种敏捷的架构中快速获益，轻松打破数据及技能孤岛，并以迭代及增量的方式获得数据分析的敏捷性，缩短企业提取数据价值的创新周期。

亚马逊云科技发布了多项全新的集成功能，帮助客户在亚马逊云科技上迈向一个“Zero ETL 的未来”。亚马逊云科技一直在投入开发基于 Zero ETL 理念的功能，例如 Amazon Aurora ML 和 Amazon Redshift ML，让客户可以在机器学习用例受益于 Amazon SageMaker 的功能，而无需在不同服务间迁移数据。还有流式服务（如 Amazon Kinesis 和 Amazon MSK）向数据存储服务（如 Amazon S3）无缝注入数据，从而助力客户及时分析数据。同事，我们也发布了更多新的 Zero ETL 功能助力客户更快、更轻松、更经济的实践“数据一体化融合”。



图：实践数据一体化融合，打造端到端数据创新旅程

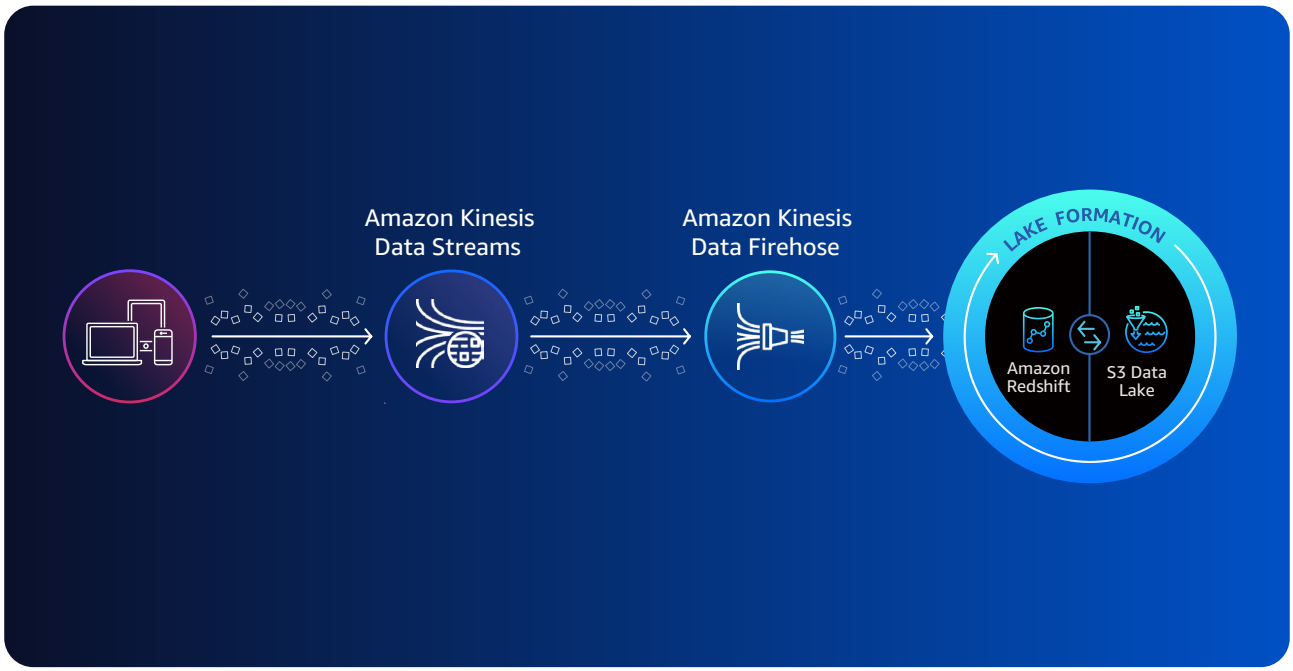
03

亚马逊云科技当前提供的 Zero ETL 选项

1.以往已提供的 Zero ETL 能力

亚马逊云科技一直致力于 Zero ETL 的研发投入，以往我们已经发布了一系列 Zero ETL 的功能特性，比如自动化流式数据入湖，联邦查询，Amazon Redshift Spectrum 等。

1.1 自动化的流式数据摄入



应用程序使用 Amazon Kinesis Data Streams 和 Amazon Managed Streaming for Apache Kafka (MSK, 完全托管、高度可用的 Apache Kafka 服务)安全地流式传输数据来存储和处理流数据，例如点击流日志。Amazon Kinesis Data Firehose 也是一种 Zero ETL 方式，可将数据直接加载到 Amazon S3 数据湖、Amazon Redshift 和 Amazon Open-Search 服务中。

1.2 联邦查询能力

以前，我们需要先将数据从 PostgreSQL 数据库提取至 Amazon S3 数据湖，而后使用 COPY 将其加载至 Amazon Redshift，或者使用 Amazon Redshift Spectrum 对 Amazon S3 进行直接查询，其背后往往并伴随着复杂的 ETL 调度。

1

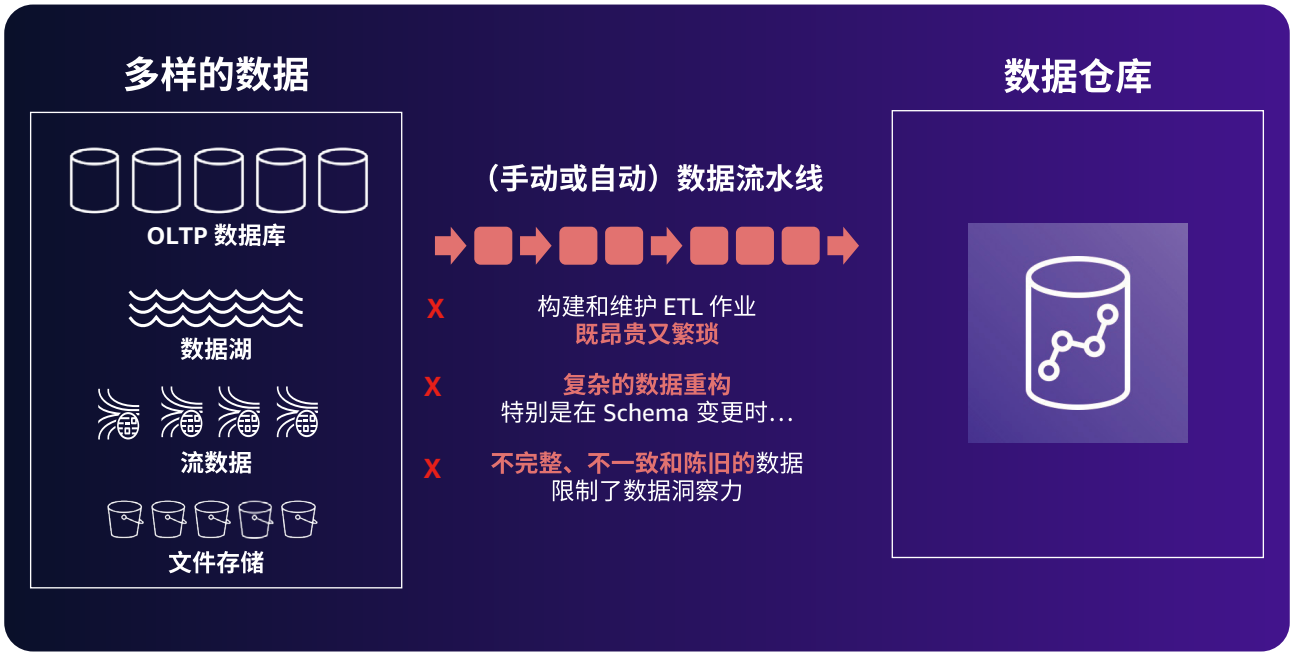
Amazon Redshift Federated Query 旨在帮助用户使用 Amazon Redshift 提供的分析功能直接查询存储在 Amazon Aurora PostgreSQL 与 Amazon RDS for PostgreSQL 数据库内的数据。Federated Query 可实现实时数据集成的 Zero ETL 处理流程。

2

Amazon Athena Federated Query，您可以对存储在关系数据源、非关系数据源、对象数据源和自定义数据源中的数据运行 SQL 查询。您可以通过点击了解更多关于 Athena 联邦查询目前以支持的外部数据源。

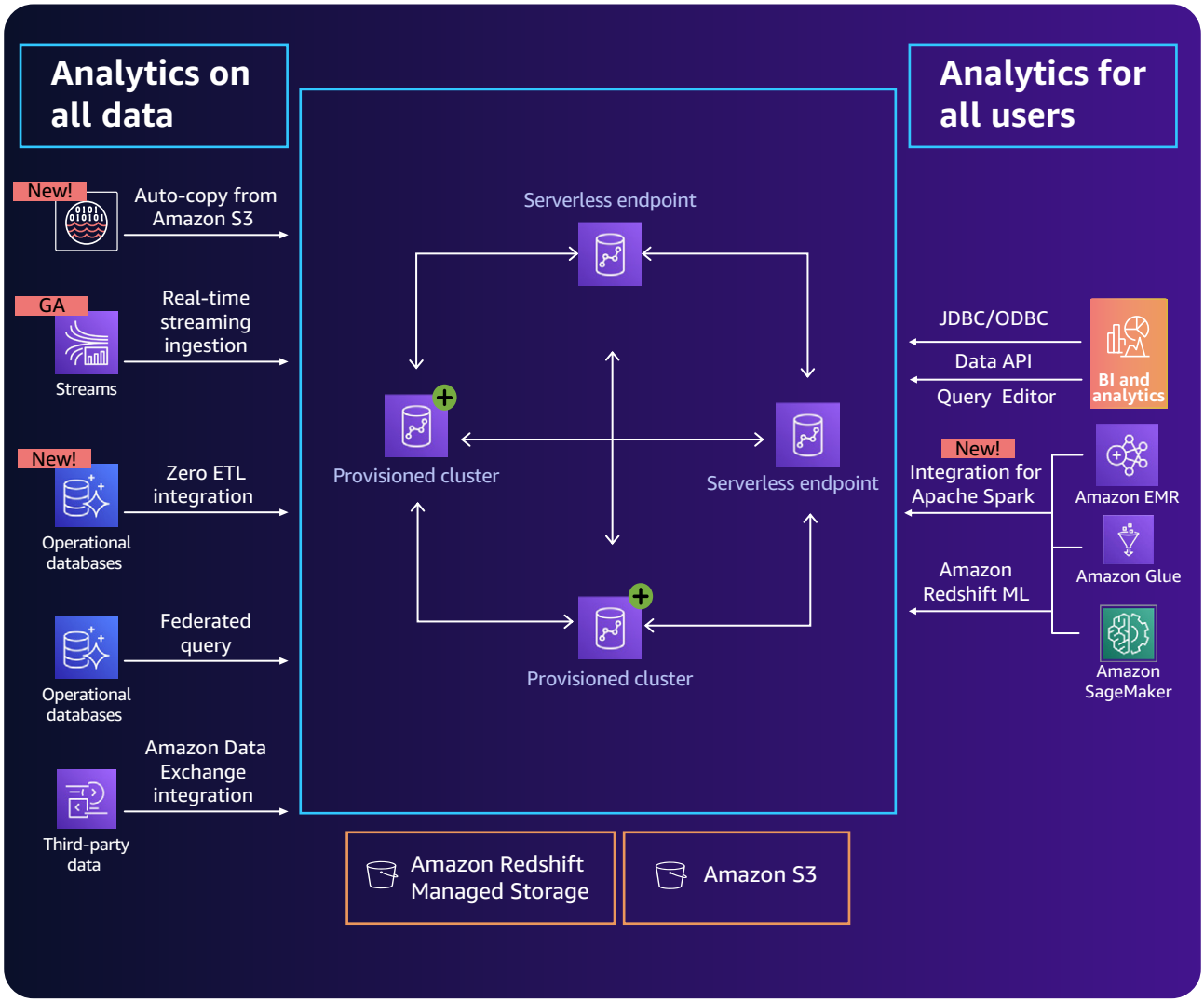
2.以云原生数据仓库 Amazon Redshift 为核心的一系列 Zero ETL 摄入功能

要想数据分析到位，首先要保证有稳定、可靠的数据摄入通道，来实现端到端的第一环（其实还有第零环是业务在数据源侧的规划），而这一块也是大部分数据工程中遇到最头疼的问题之一。首先，数据源就包含很多种，最常见的数据源包括关系型数据库、数据湖和实时的流数据。其次，不管是手动还是自动的 ETL 流水线，都需要专业的数据工程团队来构建和维护，并且经常要处理或介入数据结构的变更等情况。这次，Amazon Redshift 连发多个功能特性来帮助客户解决或者消除这类问题。



图：多数据源入库的复杂程度

而经过了十年的创新，Amazon Redshift 目前已经具备了联邦查询、流式数据自动摄取、关系型数据库 Zero ETL，与机器学习（Amazon SageMaker）联动，Serverless 等一系列强大功能，为用户分析所有数据，提供支持与保障。



图：一图了解 Amazon Redshift 的创新

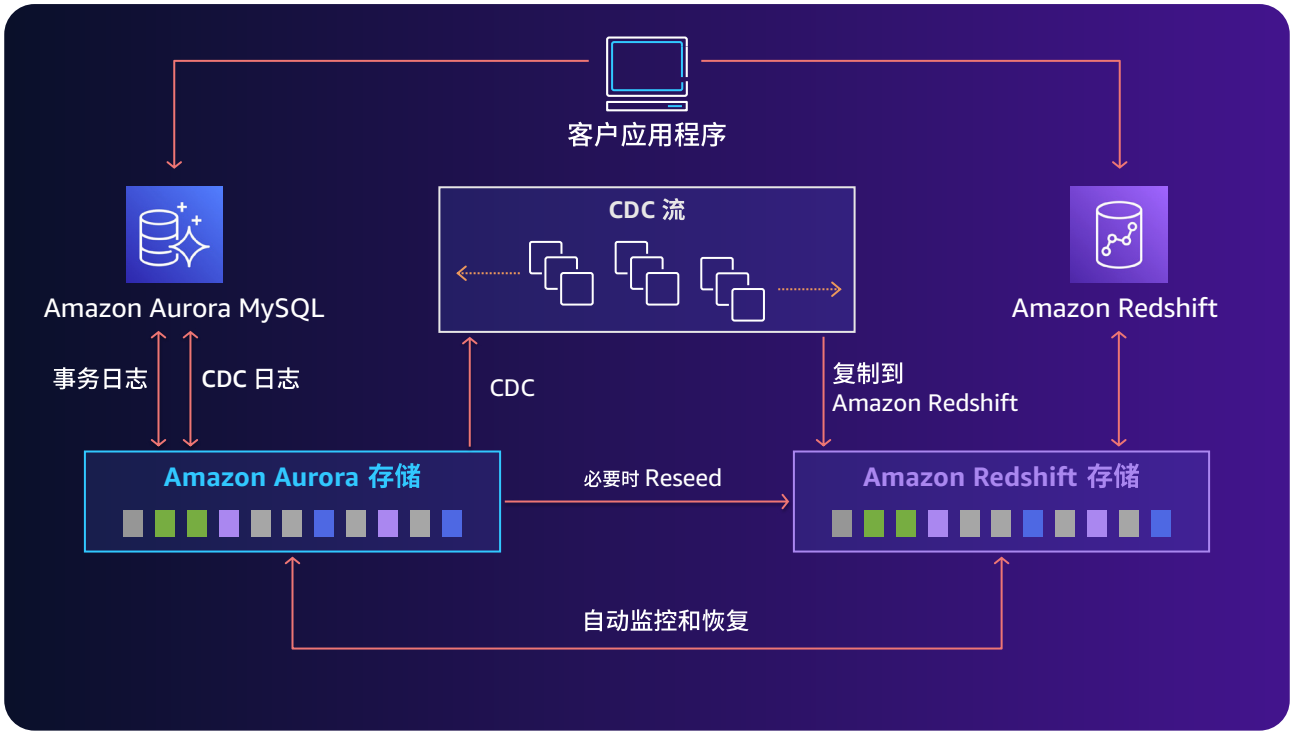
2.1 Amazon Aurora Zero ETL to Amazon Redshift 助力 PB 级分析交易数据进行近实时分析

首先是最常见的关系型数据库，也就是经典的 OLTP 向 OLAP 的数据传递。如果是为了更快或者更实时地获取线上业务的事务数据来做分析，通常可以通过开启数据库的 Binlog 来捕捉 CDC 变更，然后再使用解析 CDC 的工具如 Amazon DMS、Debezium 等来实现，这些都需要客户进行不断地监控、配置和优化。此外，不同的数据库和数据表可能会有不同的需求，这样就再加倍了数量级的维护成本。

图例：在云上构建一个 ETL 工作流

企业希望更好地了解核心业务驱动因素，制定战略以增加销售额、降低成本、获得竞争优势，因此，近乎实时地从购买、预订和金融交易等交易数据中获得洞察的需求不断增加。

Amazon Redshift 通过与 Amazon Aurora 数据库深度集成，在事务型数据写入 Amazon Aurora 后，数据在底层被持续地复制到 Amazon Redshift，完成行式数据存储到列式数据存储的转换，彻底消除了自己构建和维护复杂数据管道的工作。没有 Hybrid OLTP 和 OLAP，仍然是熟悉专门构建、极致性能的 Purpose-Build 工具，各司其职，并以最高性价比解决最实际的问题。同时，客户的应用程序架构保持不变，读写端点指向 Amazon Aurora，分析端点指向 Amazon Redshift，但是底层已经不再是一大串接一大串的数据抽取、转换和加载，直接无缝衔接并且达到近实时的效果。



图例：Amazon Aurora Zero ETL to Amazon Redshift 功能实现

Amazon Aurora Zero ETL to Amazon Redshift 功能，交易数据在写入 Amazon Aurora 后的几秒钟内可以自动连续复制，使其在 Amazon Redshift 中即时可用。一旦数据在 Amazon Redshift 中可用，客户立即可以开始分析数据，并且应用数据共享和 Amazon Redshift ML 等高级功能获得全面的预测性洞察。客户可以将数据从多个 Amazon Aurora 数据库集群复制到同一个 Amazon Redshift 实例，跨多个应用程序获得洞察。如此，客户可以使用 Amazon Aurora 支持交易数据库需求，使用 Amazon Redshift 进行分析，无需构建或维护复杂的数据管道。

2.2 Amazon Redshift 支持 Amazon S3 自动复制

在智能湖仓架构中，数据仓库 Amazon Redshift 开始支持从 Amazon S3 数据湖中自动复制，原本需要手动才能完成的工作，现在只需要配置即可实现。

之前，如果想要拷贝数据都需要手动或者定时执行 COPY 命令，现在 Amazon Redshift 新添加了 COPY JOB 命令自动检测指定路径的新文件，跳过已经加载完毕的旧文件。自己编写的定时任务脚本可以退役了，再也不用担心手抖重复执行，数据工程师的工作变得美好了起来。

2.3 Amazon Redshift streaming ingestion 流式数据接入功能已上线

如果业务需求是实时的，那么通过 Amazon S3 作为 Staging 存储在 COPY 的方式就跟不上节奏了，Amazon Redshift 流式摄入目前已支持 Amazon Kinesis Data Streams，今年新增支持了 Amazon Managed Streaming for Apache Kafka(MSK)，同时流式摄入也正式推出，告别预览。从上面的图中可以看出，流式摄入合并了数据消费的过程，直接在 Amazon Redshift 中实现并持续加载到数据仓库。在 Amazon Redshift 中，流式摄入是通过物化视图的方式实现的（查找官方文档是在物化视图章节），用户还可以在这个物化视图基础上再配合其他数据叠加物化视图提高查询效率。另外，记得可以给流式摄入开启自动刷新功能哈～从此，客户可以更简单地完成实时数据分析，包括 IoT 物联网设备、点击流、应用程序监控、欺诈检测和游戏实时排行榜等。

3.超过百种外部数据连接，助力构建 Zero ETL 未来

纵观全局，亚马逊云科技数据服务已经可以连接超过 100 种外部数据源，像 Adobe，Salesforce 等各类 SaaS 应用，也包括了各类 on-premise 数据源类型，因此您能更全面的利用所有数据的力量。

1. 大多数环境下，数据分散在多个系统和数据存储中（无论是本地还是在云中），Amazon AppFlow 在本地系统和应用程序、SaaS 应用程序和亚马逊云科技服务之间提供双向数据集成。使用低代码或无代码、经济实惠的解决方案，只需单击几下，Amazon AppFlow 即可在 Salesforce、SAP 或 ServiceNow 等 SaaS 应用程序与您的 Amazon S3 数据湖和 Amazon Redshift 之间安全地传输数据，帮助客户打破数据孤岛，随着业务需求的变化，该解决方案可在几分钟内轻松重新配置。

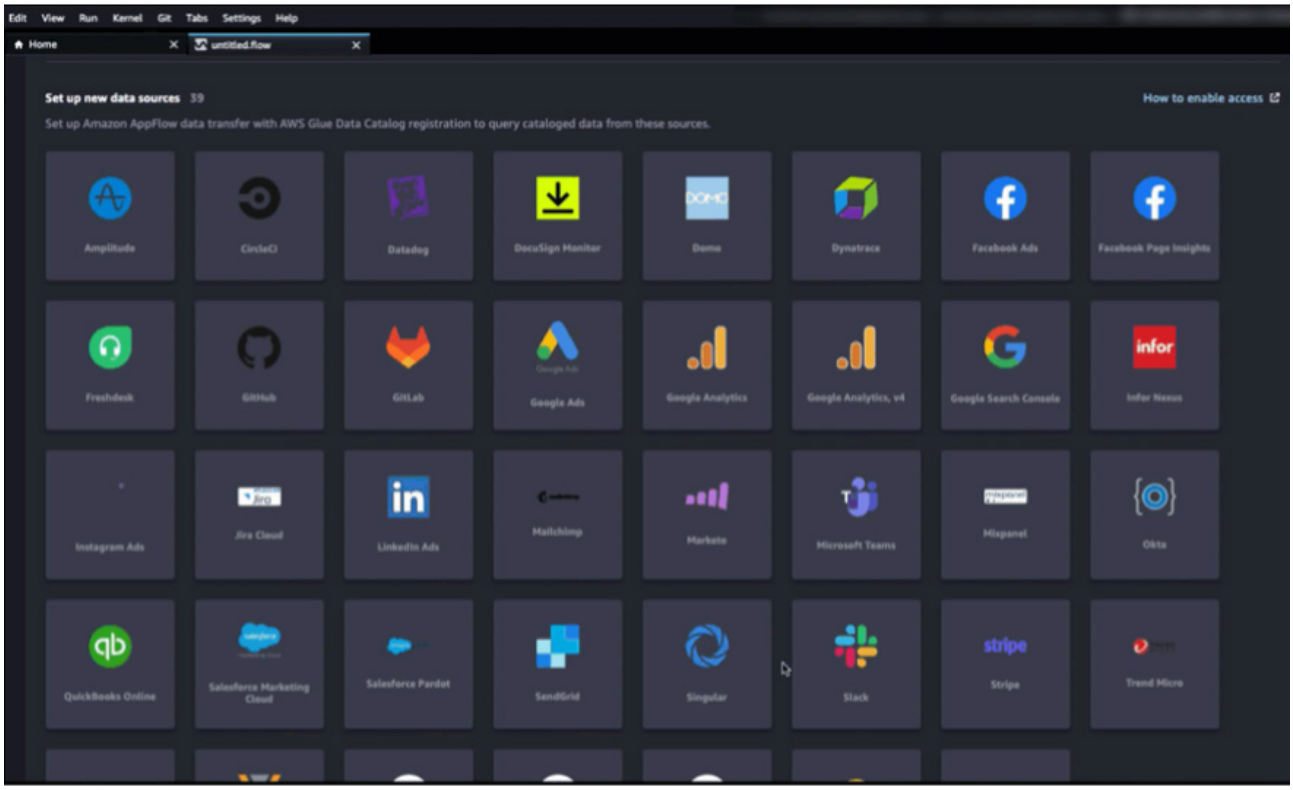
营销连接器（例如，Facebook 广告、谷歌广告、Instagram 广告、领英广告）。

用于客户服务和参与的连接器（例如 MailChimp、Sendgrid、Zendesk Sell 或 Chat 等）。

业务运营（Stripe、QuickBooks 在线和 GitHub）。

在 re:Invent 2022 上我们发布了新增 22 个新的连接器，现在 Amazon AppFlow 已经支持超过 50 种连接器。

2 Amazon SageMaker Data Wrangler 支持基于 Amazon AppFlow 的 SaaS 数据源。从今天开始，您可以在 Amazon SageMaker Data Wrangler 中聚合 ML 的外部 SaaS 应用程序数据，为 ML 准备数据。数据源通过 Amazon AppFlow 注册到 Amazon Glue Data Catalog 后，您就可以使用 Amazon Data Wrangler SQL explorer 浏览这些数据源中的表和模式。该功能使用 Amazon AppFlow 在 SaaS 应用程序和 Amazon SageMaker Data Wrangler 之间提供无缝数据集成。



图例：Amazon SageMaker Data Wrangler 数据源选择

04

客户故事

Adobe:

从个人和小型企业到政府机构和全球品牌，Adobe 使每个人都能创造和交付卓越的数字体验。“Adobe 的使命是通过数字体验改变世界，在当今世界，这意味着拥有能够提供深刻和实时洞察的分析工具。”Adobe Acrobat Sign 首席科学家 Jack Lull 表示，“作为 Amazon Aurora 的客户，我们非常欢迎 Amazon Redshift 集成的 Amazon Aurora Zero ETL 功能。它将为我们不断扩大的 Acrobat Sign 客户群提供新的洞察和更快的分析能力，并随着他们用量的增加而同步增长。所有这些都不需要我们自己的团队做日常维护。”

Infor:

Infor 是商业云软件和特定行业 ERP 解决方案的全球领导者。“在 Infor，我们使用亚马逊云科技构建和部署现代化的工具，帮助客户转型其业务并加速创新，其中包括我们最新提供的面向客户行业云数据的托管数据仓库服务，以帮助客户通过高级分析和机器学习更快地做出决策。”Infor 云服务高级副总裁 Jim Plourde 表示，“我们很高兴使用 Amazon Redshift 集成的 Amazon Aurora Zero ETL 功能，它将让 Amazon Aurora 中的交易数据近乎实时地提供给 Amazon Redshift，减轻我们的运营负担。现在，我们既可以受益于 Amazon Aurora 用作关系数据库管理系统的性能，又可以轻松利用 Amazon Redshift 的分析和机器学习功能实现新的托管数据仓库服务。”

高盛集团：

高盛集团是一家领先的全球金融机构，为包括企业、金融机构、政府和个人在内的庞大而多元化的客户群提供投资银行、证券、投资管理和消费者银行业务等广泛的金融服务。“我们的重点是为高盛内所有用户提供自助式数据访问。当在整个金融服务行业开展协作时，我们通过开源数据管理和治理平台 Legend 可以助力用户开发以数据为中心的应用程序，并且获得数据驱动的洞察。”高盛首席数据官 Neema Raphael 表示，“通过面向 Apache Spark 的 Amazon Redshift 集成功能，我们的数据平台团队以最少的定制化操作就可以访问 Amazon Redshift 数据，实现零代码 ETL，使我们更有能力在工程师收集完整及时的信息时，让他们更容易专注于完善其工作流。由于我们的用户现在可以轻松访问 Amazon Redshift 中的最新数据，我们将能实现更高的应用程序性能和更强的安全性。”

亚马逊云科技

扫码了解云原生数据战略

助力 150 万家企业成为数据驱动型企业