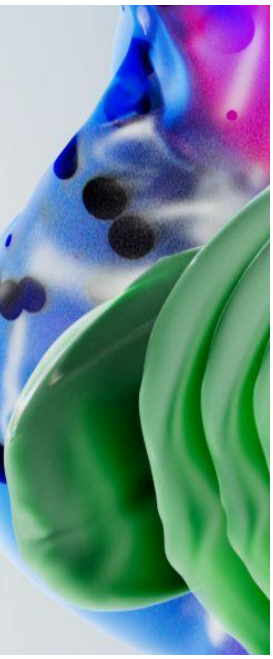
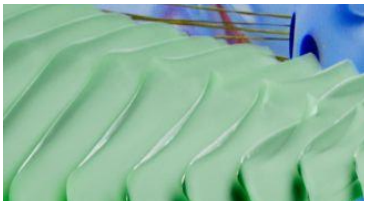
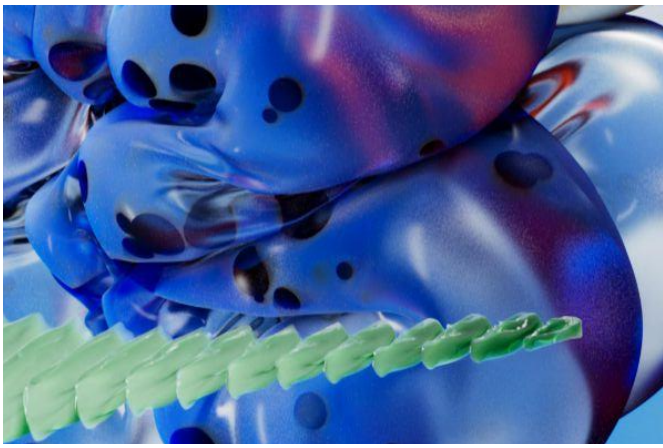
Google Cloud

数据湖迁移实用指南







释放数据潜力，推动 AI 转型

作者：Dana Soltani，
Group Product Manager，Google Cloud 组合产品经理

目录

摘要

如果您计划将数据湖迁移到云端，本指南可以帮助您成功完成这个旅程中的各个阶段。过去五年来，我们深入研究了无数企业的数据湖迁移案例，**本指南以此为依据，旨在帮助您优化这个迁移旅程**。

了解为什么要进行迁移

探索数据湖的演变历程，了解为什么一定要采用基于云的方法进行现代化改造。您将了解老旧系统面临的挑战，以及迁移到兼具扩缩能力和成本效益的创新型云环境的诸多优势。

做出明智的经济决策

深入了解云迁移的经济效益，包括成本优化策略以及大幅节省成本的公司的真实案例。您将了解如何让迁移策略与企业服务等级协议 (SLA) 保持一致，以及如何根据自身需求选择合适的方法。

制定迁移计划

按照指南中列出的五个迁移阶段，有条不紊地执行发现、评估、规划、执行和优化流程。我们针对每个阶段列出了详实的核对清单和主要注意事项，助您打造结构化且高效的迁移旅程。

自信从容地完成迁移

采用最佳实践迁移数据湖的各个部分，包括数据、元数据、工作负载、治理和工作流。您将了解如何确保数据完整性，以及如何尽可能减少停机时间和优化云端性能。

充分利用 Google Cloud 的支持服务

探索 Google Cloud 提供的一整套服务、工具和专业知识，包括来自客户支持团队的个性化支持、自动迁移工具、咨询服务、技术援助以及庞大的合作伙伴生态系统。

通过遵循本指南中的指导和最佳实践，您可以放心地将数据湖迁移到Google Cloud，从而实现扩缩能力和成本效益，同时获享先进的数据分析和AI/机器学习功能。

第 1 章

数据湖的演变历程

近年来，组织收集、存储和分析数据的方式发生了巨大变化。从数据库到数据仓库，再到数据湖，这种演变历程与互联网、大数据分析以及如今的 AI 等创新相辅相成。为了帮助您了解我们当前所处的阶段，我们先来回顾一下数据湖的发展历程。

一切始于大数据

在大数据发展的早期阶段，传统数据仓库已经难以应对数据量的爆发式增长和数据类型的多样化形势。传统数据库仓库为结构化数据而构建，这导致非结构化和半结构化数据未能得到充分利用。21 世纪 10 年代初出现的数据湖旨在弥合这一差距。数据湖是一个庞大的仓库，它以原始格式存储原始数据，不考虑数据的结构或预期用途，这让组织有机会收集和分析所有类型的数据。

Hadoop 在这一领域做出了开创性贡献。根据 Google 发布的 MapReduce 论文[1]，Hadoop 分布式文件系统 (HDFS) 可以存储海量数据，其处理框架 (MapReduce) 还可以并行分析这些数据。后来出现了 Apache Spark，它是一种更快速而高效的处理引擎。不久之后，组织开始在自有自营数据中心构建和维护本地数据湖。这样，他们便可以安全且经济高效地存储和处理专有数据，依靠 Hadoop 或 Spark 等架构完成工作。

传统数据湖面临挑战

本地数据湖提供了控制力和安全性，但难以让组织充分利用其数据资产。面临的挑战包括：

扩缩能力受到限制

受制于硬件基础设施的物理容量，通过扩容满足不断增长的数据容量和处理需求非常耗时且成本高昂。这导致了性能瓶颈，削弱了处理激增的数据或工作负载的能力。

高昂的前期 费用和维护费用

构建和维护本地数据湖需要在硬件、软件许可和 IT 基础设施方面进行大量前期投资。持续的维护，如硬件升级、软件更新和安全修补，也会增加总拥有成本。

管理开销

执行硬件预配、软件安装、配置、性能调整和安全管理等任务需要具有专业知识和专用 IT 资源，这会占用数据分析和洞见生成等高附加值活动的时间。

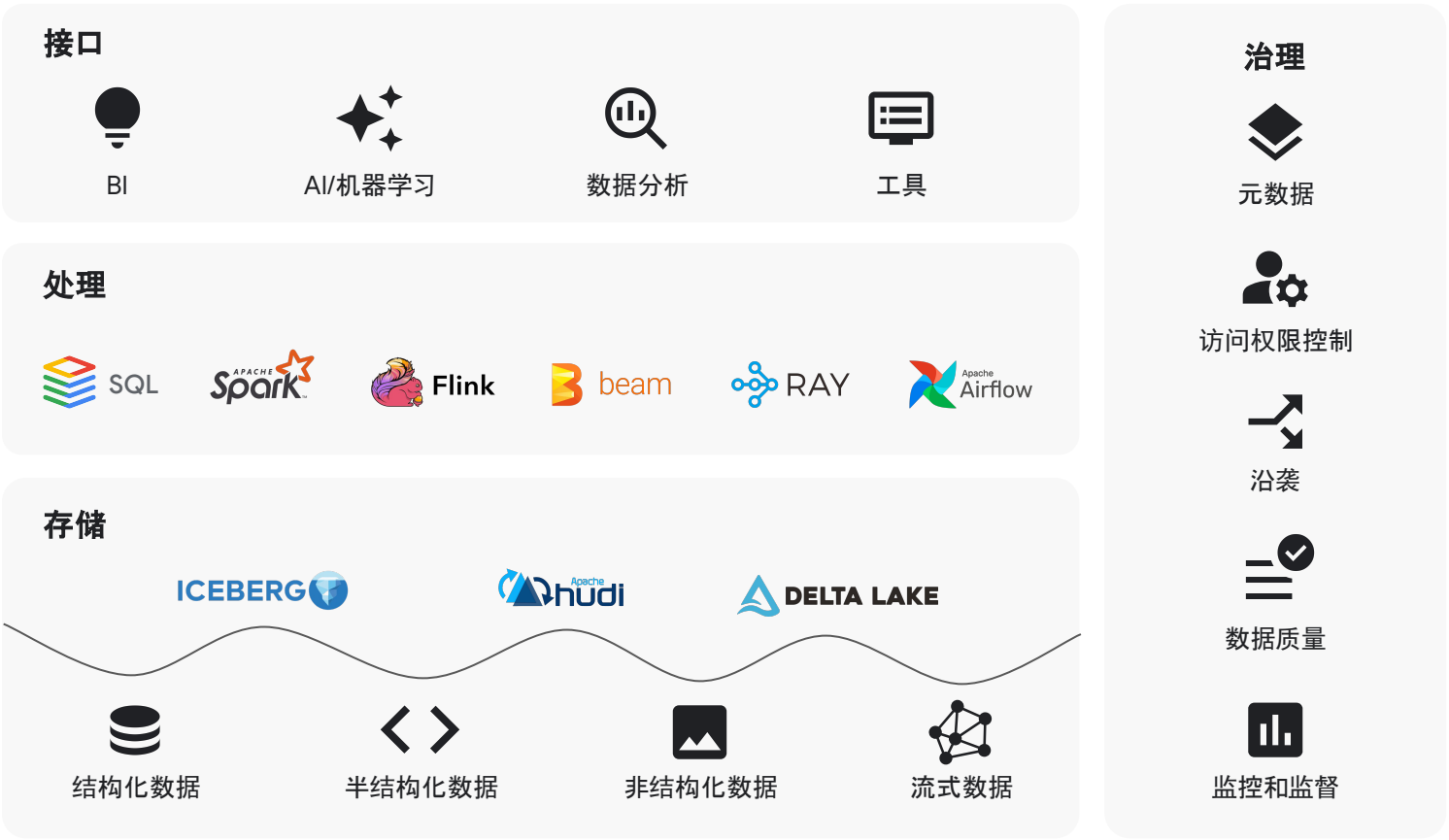
影响创新

分析、AI 以及云原生服务中的最新成果（如无服务器计算、AI/机器学习平台和高级分析工具）与本地数据湖不兼容。这不仅有碍创新，还使组织难以从数据中获得商业优势。

1. Dean, J、Ghemawat, S，2004 年《MapReduce：简化大型集群上的数据处理》

云计算如何改变了数据湖

云计算将数据湖提升到了一个新的高度。如今的 现代化云端数据湖旨在解决 传统数据湖部署通常面 临的重大挑战。相较于静态、孤立的本地存 储和计算设置，现代云端数据湖具有极高的 弹性和短 暂性，它由四个 层级构建而成，旨在帮助组织从数据中 发掘新的机会。



存储

基础层将结构化、半结构化和非结构化数据作为文件存储在通用云存储中。Apache Parquet、Avro 和 ORC 是最常见的文件格式。虽然 Hive 表格式通常用于查询 SQL 中的数据，但它正在被 Apache Iceberg 等更现代的格式所取代。除了原子性、一致性、隔离性和持久性(ACID) 事务支持外，Iceberg 还提高了 PB 级表的效率，并提供了架构和分区演变、时间旅行以及物化视图等高级功能。

处理

存储和计算相分离的特性从本地数据湖沿用下来，并在云端得到加强。云端的数据处理具有短暂性，计算资源仅在必要时才进行预配，以最大限度降低成本。根据不同的应用场景，可能会使用多个处理引擎（如 Apache Spark、Flink 或 Ray）来批量或实时处理数据。

工具和接口

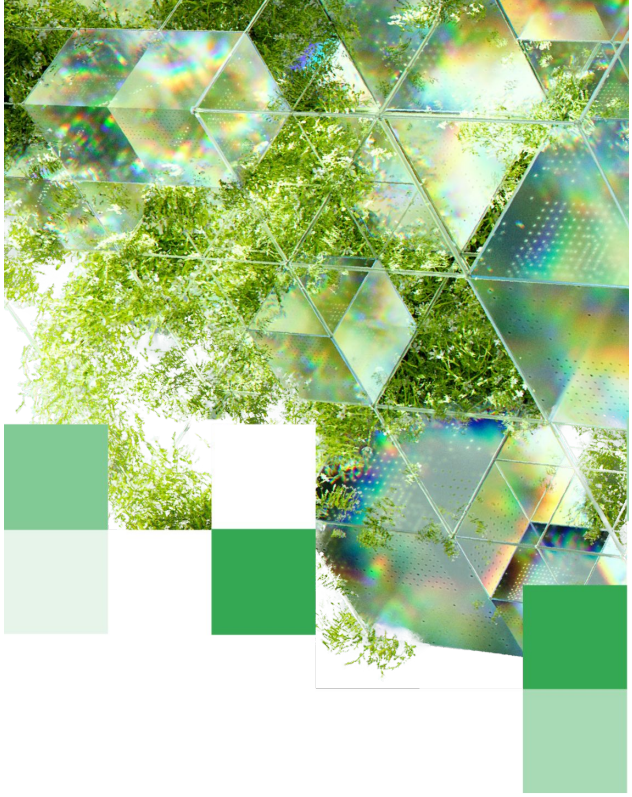
这是不同用户与数据湖进行接触的层级，应用场景包括：临时分析，以及在业务应用中构建用于持续预测的机器学习流水线等。用户通过 BI 工具、数据科学笔记本、SQL Workbench 以及 API 等方式与数据湖进行交互。

治理

数据湖通常包含敏感数据和关键任务流程，因此需要对其进行严格治理。元数据管理、数据质量控制、沿袭跟踪以及访问权限控制等功能有助于降低运营成本和数据泄露风险，避免组织遭受重大损失。

这种结构缩短了数据分析的价值实现时间，并能够在整个组织内更快速地交付分析洞见，其投资回报率明显高于许多组织目前仍在使用的本地数据湖。

云数据湖释放数据和 AI 潜力



许多组织开始意识到，他们当前的数据基础设施无法满足 AI 发展需求。传统数据湖无法高效存储和处理 AI 所需的大量非结构化数据。数据孤岛在仓库、数据湖和云端非常普遍。此外，许多组织也没有足够的计算资源来构建和部署 AI 模型。

为了突破这些局限，传统数据湖下一步的发展自然是进行现代化改造，也就是将数据湖迁移到云端，并对架构进行现代化改造，以便将所有数据整合到一个统一的平台上。

云数据湖能够让组织：

加快数据驱动型决策

在云端，数据管理和处理实现了统一，这让组织能够更轻松地利用数据获得竞争优势，为最终客户创造价值。

加速 AI 采用

面向结构化和非结构化数据的实验、监控和治理工具随时可用，还有机器学习模型助力加速 AI 创新。

降低复杂度

云数据湖通常提供多层存储选项、多种处理引擎以及丰富的机器学习和 AI 模型，这有助于简化数据和 AI 应用并降低费用。

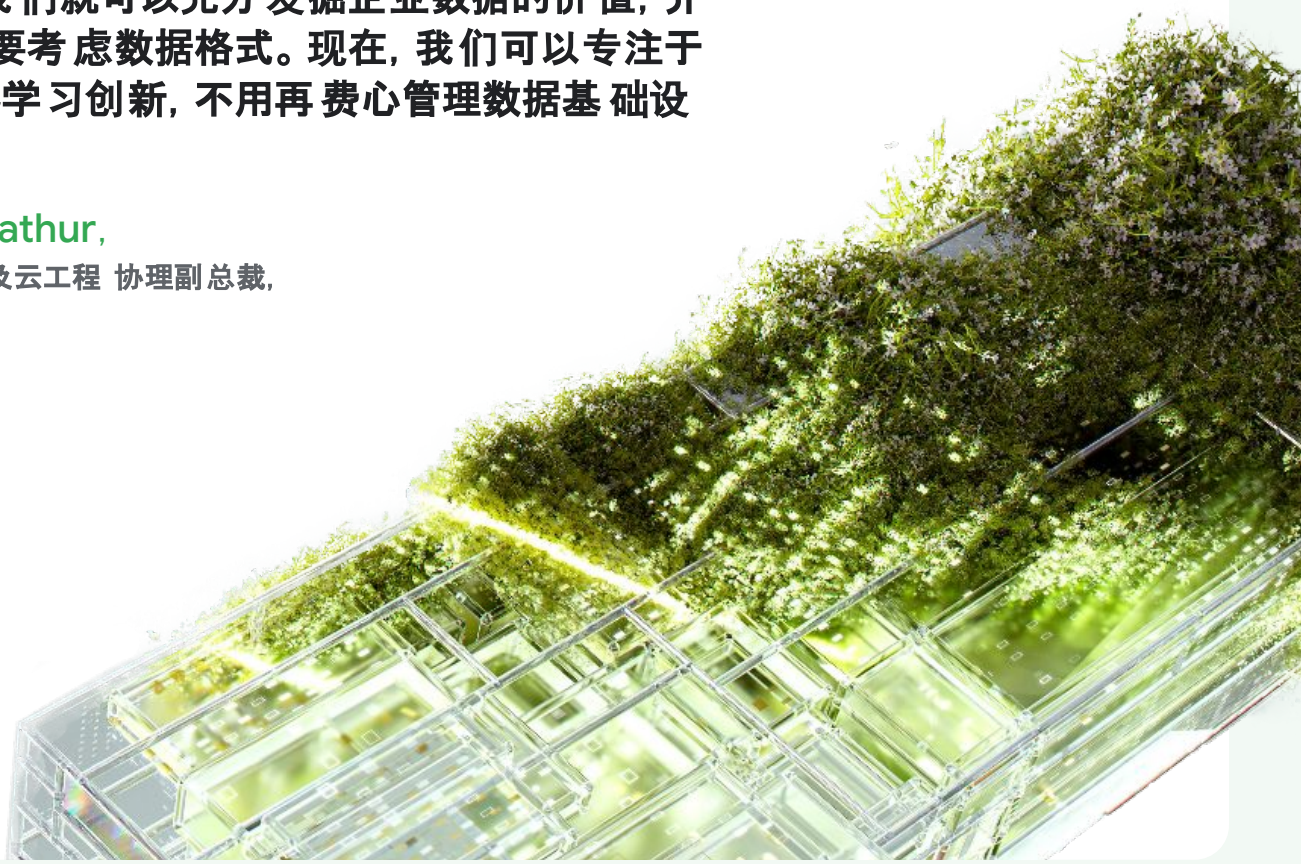
加强治理

在云端，更容易确保数据和分析洞见的质量、可信度和法规遵从性。

Definity 迁移到 BigQuery 以节省成本并提升性能

Definity 是加拿大一家拥有 150 多年历史的财产和意外险公司。作为行业领先企业，他们深知只有对数据基础设施进行现代化改造，才能应对日新月异的行业发展。公司老旧的本地Cloudera 平台正在影响其扩缩、创新以及充分利用数据和AI 的能力。在短短10 个月内，Definity 就成功迁移到了Google Cloud，在 BigQuery 上构建了数据存储区，并将 Vertex AI 设置成了分析和AI 平台。公司每年的基础设施和运营成本节省了 30% 以上，部署时间缩短了 63%，基础设施设置速度提升至原来的10 倍。

> **“我们之所以选择 BigQuery，是因为它具有扩缩能力和成本效益，并且 还能够与 Vertex AI 无缝连接，这完美契合了我 们对统一数据分析平台的需求。BigQuery 支持所有数据 类型，有了 这个平台，我 们就可以充分 发掘企业数据的价 值，并且不需要考 虑数据格式。现在，我 们可以专注于 AI/机器学 习创新，不用再 费心管理数据基 础设施。”**
>
> Nitin Mathur,
> **数据平台及云工程 协理副 总裁，**
> Definity



第 2 章

数据湖迁移的经济效益

将数据湖迁移到云端有诸多优势。转向云模式不仅能够节省大量成本，还有助于提高效率，加快分析洞见获取速度，以及增强业务决策和利用最新 AI 创新的能力。有了这些优势，组织便能够获得更多创收机会和竞争优势。

20-62%

成本降低幅度

实现方式：使用托管式云数据湖，而不是本地或自行管理的数据湖[2]

80-90%

成本降低幅度

实现方式：采用湖仓一体架构，将数据湖和仓库整合到托管式云服务中[2]

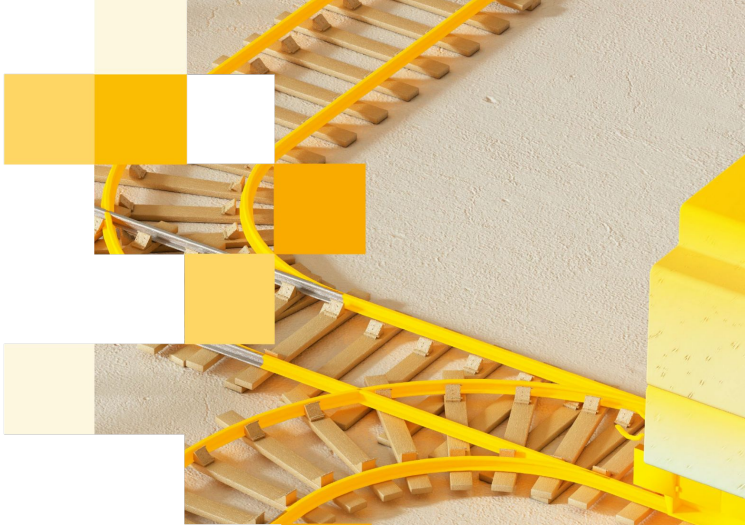
成本是推动转型的关键因素

毫无疑问，本地数据湖的成本效益越来越低。首先，构建和维护基础设施都需要成本，这会导致高昂的资本支出和运营支出。其次，为了应对不断增长的数据容量和种类，本地数据湖往往需要添置昂贵的硬件。最后，管理和保护基础设施需要具备专业的 IT 知识，这进一步增加了人员成本。

除此之外，组织还会错失很多机会。本地解决方案不如云端数据湖灵活敏捷，这会影响快速创新和 AI 的采用，使组织更难应对不断变化的业务需求。因此，许多企业意识到本地数据湖的发展模式难以为继。

2.ESG，2022 年，《Google Cloud Dataproc 经济效益分析》

云计算改变了组织的成本结构



云数据湖的成本结构类似于随用随付的公共事业模式，费用随使用量发生变化。

主要组成部分包括：

存储

这是一项基本费用，按所存储的千兆字节 (GB) 数计费，同时受多种因素的影响，包括数据量、访问频率和所选存储层（热存储、冷存储、归档存储）。

计算

处理和分析数据时产生的费用，计费取决于执行提取、转换和加载 (ETL)、查询和机器学习等任务所用的计算实例的类型和持续时间。

托管式服务

使用托管式分析和数据仓储引擎、BI 工具以及流式服务等的费用。

数据传输

将数据移入、移出云环境或在云内部移动时产生的费用。

编排和管理

与调度和监控数据及 AI 流水线相关的费用。

数据治理、安全性及合 规性

访问权限控制、加密和审核等功能会增加总体成本。

了解和优化各个部分的成本 对于管理云数据湖的总拥有成本至关重要。

Eureka 借助 Google Cloud 减少停机时间并加速应用部署

科技公司 Eureka 将 AI 和机器学习应用于大型数据集，帮助组织大规模挖掘与智能化相关的信息。为了更好地运行数据分析服务，该公司迁移到了 Google Cloud，并开始采用 BigQuery 和 Dataproc，这帮助他们减少了服务器停机时间，同时加快了新应用的部署速度。他们还节省了大量成本，BigQuery 存储让公司的成本降低了近18%。

> **数据分析能力的 显著提升令人振 奋。与之前使用的本地解决方案相比，自从迁移到 Google Cloud 后，停机 时间和服务中断都大幅减少了。**

Michael Hawkins
首席营销官 (CMO)，
Eureka

阅读完整案例 →

规划前期迁移费用

将本地数据湖迁移到云端涉及一系列前期费用。建议您从一开始就了解清楚这些费用，以避免意外支出或预算超支。

前期迁移费用通常包括：

云基础设施

在迁移期间，当本地数据湖仍在使用且无法停用时，预配虚拟机、存储和网络组件所产生的费用。

专业服务

大多数企业组织需要利用咨询、实施支持和培训服务来协助进行云迁移。

数据传输

这可能包括从当前数据中心迁出数据的出站流量费用，以及潜在的带宽费用。

数据转换和清理

将数据适配到云环境并确保兼容性可能产生大量开支。

许可和订阅费用

您可能需要负担数据湖迁移所需的云端软件和工具费用。

组织可以优化云端成本

优化云数据湖成本需要积极主动和持之以恒。您可以在迁移前、迁移中和迁移后采取下面这些措施，在确保性能和扩缩能力的同时大幅降低成本。

合理调整计算和存储资源规模

分析您的工作负载模式，并根据实际需求调整计算容量。使用自动扩缩功能按需动态调整资源。根据数据访问频率选择适合的存储层（热存储、冷存储、归档存储）。

采用经济高效的定价模式

利用预留实例或 Spot 实例处理可预测或容错型工作负载。探索云提供商提供的持续使用折扣和其他定价方案。

优化数据存储

压缩数据以减少存储占用空间，利用数据生命周期管理政策将老化数据自动移入更实惠的存储层。定期删除或归档不使用的数据。

优化数据流水线

设计高效的数据注入和处理流水线。尽可能减少跨区域/可用区的数据移动，以降低网络费用。使用无服务器计算执行数据转换任务，仅按实际使用量付费。

监控和分析云支出

利用云提供商的费用管理工具来跟踪支出、找出费用源头并设置预算提醒。定期查看费用报告并分析支出模式，找到适合优化的费用项目。

定期检查和优化

持续监控数据湖环境，根据需求变化调整策略。持续关注最新云产品和成本优化最佳实践。

实施数据治理政策

制定清晰的数据保留政策并执行数据质量标准，尽可能降低存储费用和提高数据易用性。

使用云原生工具和服务

利用 BigQuery 或 Dataproc 上的无服务器 Spark 等托管式服务执行 ELT/ETL 流程，获得比自行管理的解决方案更佳的成本效益。

LiveRamp 将 Hadoop 迁移到 Google Cloud

随着业务规模不断扩大，LiveRamp 在本地环境遇到数据中心空间和电力方面的限制，这影响了公司实现业务目标的能力。为此，他们制定了战略性决策，即利用弹性环境并从 Hadoop 迁移到 Dataproc，这让他们获得了丰厚的回报。

> **“通过合理搭配使用按需虚拟机和 Spot 虚拟机，某些集群的成本降幅达到了 30% 左右。我们之所以能够大幅节省费用，是因为我们的工程师构建了高效的 A/B 测试框架，该框架帮助我们以多种配置运行集群 /作业，从而获得最可靠、最易维护且最具成本效益的配置。此外，其中一个应用的速度现在提升了 10 倍以上。”**
>
> Mithun Bondugula
> 高级工程经理，
> LiveRamp

阅读完整案例 →

组织必须谨慎管理迁移风险

在将数据湖迁移到云端的 过程中，组织主要面临三个方面的 风险。以下是需要注意的事项和降低这些风险的方法。

01 以数据为中心的风险

数据丢失

网络问题、提取/加载错误，甚至在新环境中的误删操作，都有可能导致数据在传输过程中丢失。解决方法包括实施可靠的数据验证和备份/恢复机制，以及使用数据质量工具来确保完整性。

数据损坏

数据可能会在传输或转换过程中损坏，这可能是由格式不兼容、编码问题或迁移脚本错误所导致的。迁移前后必须执行全面的数据剖析和架构映射验证。

数据安全

维护数据安全应贯穿于迁移过程中的每个阶段。潜在风险包括未经授权的访问，数据在传输中泄露，或者云环境配置不当导致数据泄露。迁移项目的每个阶段都应实施加密、强大的访问权限控制和安全审核。

02 项目管理和执行风险

时间安排不合理

低估迁移的复杂度和所需时间会导致决策仓促、错误增加，最终导致项目失败。周密规划是关键，例如合理安排时间，以及为应对意外问题留出缓冲时间。

技能差距

缺乏云专业知识会严重影响迁移，包括能否熟练运用云平台、数据迁移工具和安全性最佳实践。培训现有员工或聘请经验丰富的云专业人员很有必要。

沟通不畅

迁移项目会涉及多个利益相关方。沟通不畅会导致误解、项目延误和各种冲突。请建立清晰的沟通渠道，确保所有利益相关方都了解最新进展并达成一致意见。

测试不充分

测试是非常重要的一个环节，它有助于发现并解决潜在问题，避免这些问题影响后续生产。如果测试仓促或不充分，在迁移后就可能出现一系列问题，让组织蒙受巨大损失。因此，必须制定周全的测试策略，包括性能和安全性测试。

03 云环境特有的风险

受制于供应商

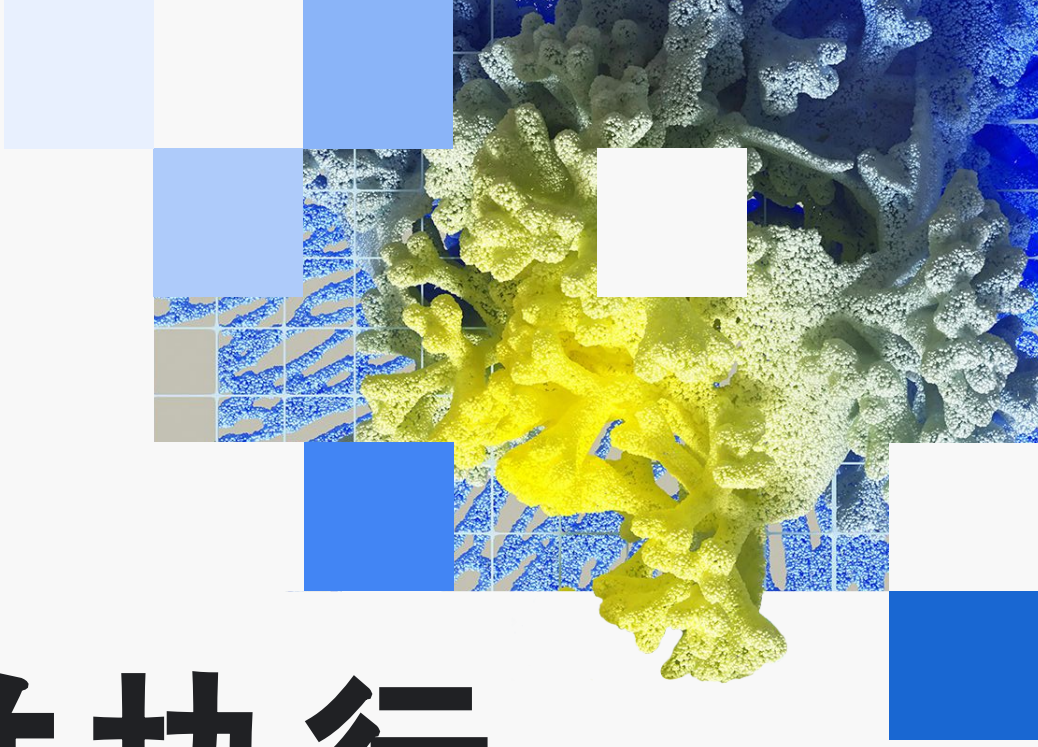
过于依赖单一技术或提供商可能会限制未来的灵活性，还可能导致更高的成本。要减少这方面的风险，请评估多个技术方案和提供商，并考虑注重互操作性的策略。

预期以外的云费用

云价格可能很复杂。如果没有妥善规划，与数据传输、存储或计算相关的这些不可预见的费用可能会让您大吃一惊。全面的成本估算、持续的成本监控和优化策略至关重要。

集成挑战

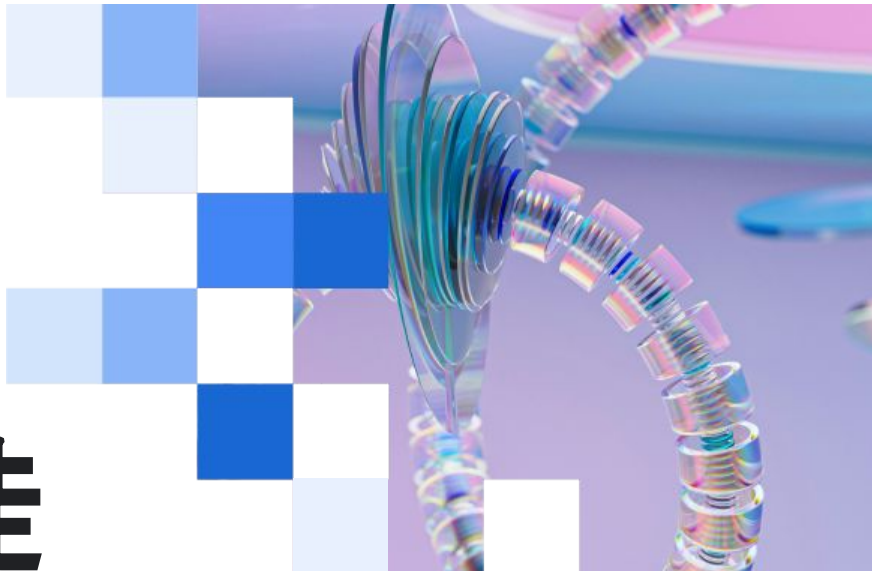
将云数据湖与现有本地系统、应用或数据源集成可能十分复杂。要确保混合环境中数据的无缝流动和兼容性，就需要精心规划并进行集成测试。



第 3 章

规划并执行成功的数据湖迁移

将数据湖迁移到云端可能是一个复 杂的过程，需要周密规划和严格执行。本章 节将带您了解所涉及的关键步骤，包括评估当前环境、制定云策略、编排迁移过程以及持 续优化，从而取得 长久性成功。



数据湖迁徙的 5 个阶段

第 1 阶段

发现

清点所有数据资产、工作负载、用户、权限和工作流，全面了解数据湖的当前状态。这些信息对于确定迁移范围和识别潜在问题至关重要。

第 2 阶段

评估

确定最佳云架构，包括选择适合的云服务、制定治理政策、明确成本控制机制，以及列出数据管理和安全方面的流程及协议。

第 3 阶段

规划

制定详细的迁移计划，包括定义任务和子任务、分配角色和职责、制定时间表以及设置预算。

第 4 阶段

执行

将表和数据、用户和权限以及工作负载和工作流迁移到云端。测试并验证迁移的数据和应用，确保它们在新环境中按预期运行。

第 5 阶段

优化

迁移完成后，继续微调性能、确保成本效益，并加强监督和治理流程。在这个阶段，您可以使用云原生功能和服务来增强扩缩能力、可靠性和可维护性。

Squarespace 利用在 Google Cloud 上构建的分析型湖仓一体架构将上报数量减少 87%

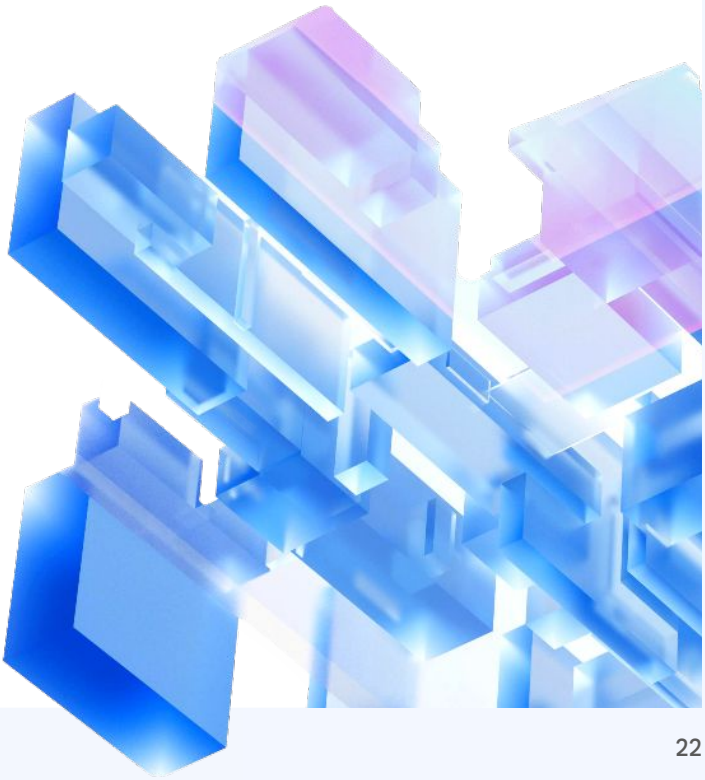
为增强扩缩能力、减少维护开销并促进数据操作创新，Squarespace 将本地 Hadoop 基础设施迁移到了 Google Cloud，并且主要利用了 BigQuery、Dataflow 和 Cloud Composer 来构建统一的数据平台。

> **“将 Hadoop 生态系统成功迁移到 Google Cloud 后，我们终于摆脱繁重的基础设施维护工作了。对比迁移前和迁移后的几个月，我们的上报数量减少了 87%。以前，数据平台和数据基础设施团队需要投入大量精力来监控各种服务/文件系统的运行状况；而现在，他们可以专注于开发新功能和改进软件体验，以满足内部用户的需求，推动业务蓬勃发展。”**

Douglas O’Connor
高级软件工程经理，
Squarespace

Constantinos Sevdinoglou
高级资深软件工程师，
Squarespace

阅读完整案例 →



迁移前的组织规划至关重要

为确保数据湖迁移成功，您应首先获得高管层的支持，并组建一支由各部门代表组成的跨职能团队，相关部门包括 IT、数据工程、数据科学、BI 以及业务部门。清晰界定每个团队成员的角色和职责，包括专门负责底层云基础设施的平台团队。

其次，建立清晰的沟通渠道和报告机制，让每个人都能了解项目进展并促进协作。

定期会议、进度报告和集中的沟通平台有助于保持透明度和及时清除障碍。鼓励团队在整个迁移过程中紧密协作、分享专业知识和相互学习。

为弥补技能差距，请提供云技术和最佳实践方面的培训和技能提升机会，这有助于团队成功驾驭新的云环境，并充分利用其功能。考虑与经验丰富的合作伙伴协作，以补充团队技能并加速迁移进程。

最后，主动解决组织内部对迁移的担忧或抵触情绪。您可以让团队成员了解迁移到云端的诸多好处，并在他们适应新环境的过程中提供全力支持。

第 1 阶段

发现

在发现阶段，您的目标是全面了解当前数据湖环境。这可以帮助您确定迁移范围、识别潜在挑战和风险，以及预估成功迁移所需的资源和时间。

同时，还应对组织需求进行整理分类，包括企业的服务等级协议 (SLA)，以确保无论是在技术层面还是在策略层面，迁移项目都能获得成功。服务等级协议 (SLA) 定义了数据和工作负载的预期性能、可用性及安全等级，这些应该是迁移策略的核心。

您应绘制当前状态的哪些元素？这份清单应包含：

数据资产

识别所有数据的来源、类型、容量、位置以及依赖这些数据的下游消费者（系统和用户），并对这些信息进行整理分类。数据发现工具可以帮助您自动执行这个过程。您要格外注意关键数据，包括支持核心业务功能和使用最严苛的服务等级协议 (SLA) 的数据。这样，您就可以安排优先迁移这些数据，尽可能减少对关键运营造成的中断。此外，还要对元数据进行整理分类。元数据是用于描述数据的信息，例如架构、数据沿袭、数据质量指标以及业务情境。

治理

详细记录当前的数据治理政策，包括数据分类、数据保留、访问权限控制规则以及合规性要求。识别所有有权访问数据湖的用户和群组，以及他们各自的角色和权限。最后，检查现有数据治理工具和流程是否与所选云平台兼容，因为您可能需要对其进行一定调整，以符合云提供商的安全和治理框架。

工作负载

识别与数据湖交互的所有流程和应用，包括 ETL 流水线、数据转换作业、分析查询以及机器学习模型。评估它们对其他系统和数据源的依赖情况，并检查它们与所选云平台的兼容情况，以便规划必要的代码修改或架构更改。与处理数据资产的方式一样，识别并优先迁移关键业务工作负载。

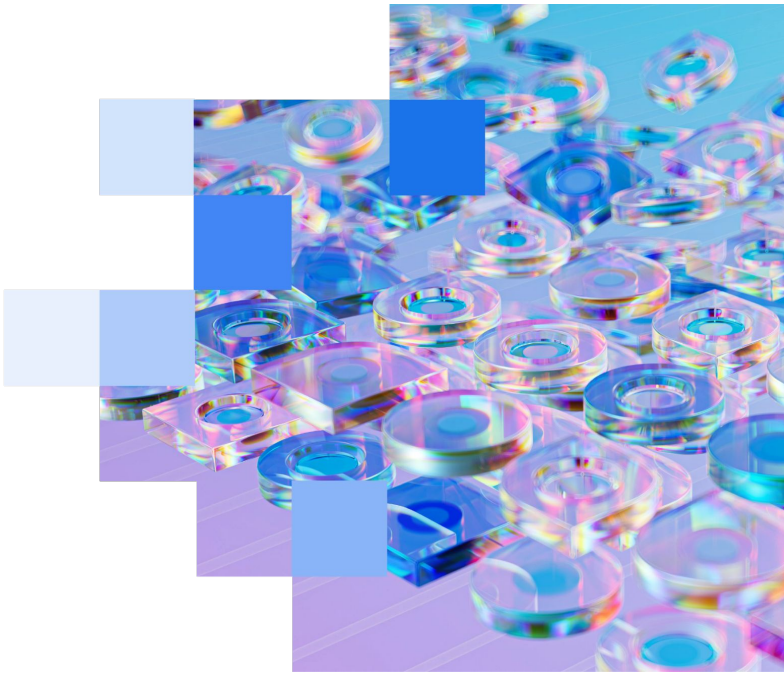
工作流

识别与数据湖相关的所有工作流和有向无环图 (DAG)，包括 ETL 流水线、数据转换作业以及机器学习工作流。分析它们的依赖关系、时间安排和资源要求。与处理工作负载和治理的方式一样，检查当前工作流编排工具与所选云平台的兼容性，以清楚了解需要进行的任何更改。

第 2 阶段

评估

对您的数据湖有比较全面的了解后，接下来就可以开始确定最优云架构。这包括选择适合的云服务、制定治理政策、确定成本控制措施，以及落实与数据管理和安全相关的流程和协议。



制定迁移政策

选择合适的迁移策略需要考虑多个因素，包括数据湖复杂程度、预算和时间限制、所需的云优化级别以及风险容忍度。清晰明确的策略有助于确保迁移过程顺畅高效，最大限度减少中断并充分发掘云投资的价值。

组织通常会选择以下三种数据湖迁移策略之一：

直接原样迁移

将现有数据湖“按原样”迁移到云端，尽可能减少更改。该策略适合时间紧迫或希望尽可能减少初始中断的用例，但它可能会影响您充分利用云优势，例如扩缩能力、成本效益和高级分析。

迁移并优化

先迁移到云端，再逐步改进数据湖架构，利用托管式云服务提升效率。该策略力求在速度和优化之间取得平衡，帮助您尽早利用部分云优势，同时以迭代方式优化费用和性能。

现代化改造

重新设计数据湖架构，以便您充分利用各种云原生服务和功能。这种综合型策略带来的长期效益最为可观，但前期投资巨大，实施时间也更长。

设计云端数据湖架构

为优化新环境中的性能、扩缩能力和成本效益，需密切关注云架构的设计。为此，请确保其标准满足或超出您的企业服务等级协议 (SLA)，包括尽可能利用可提供内置服务等级协议 (SLA) 保证的云原生服务，例如托管式数据仓库或无服务器数据处理引擎。

云端数据湖架构包含六个关键要素。

01 存储

对象存储

云对象存储区提供具备扩缩能力和成本效益的存储空间，用于存储原始数据和处理后的数据，特别是图像、视频和日志文件等非结构化数据。

表格式

相较于 Apache Hive 等传统表格式，Apache Iceberg 等现代表格式具有显著优势。它们提供 ACID 属性、架构演变和时间旅行功能，这对于更新频繁且架构不断演变的数据湖至关重要。

02 计算

无服务器流式处理

云提供商提供 Pub/Sub 和 Dataflow 等无服务器流处理平台，以及适用于 Apache Flink 和 Apache Kafka 的托管式服务。这些产品非常适合用于事件驱动的实时数据处理和转换。

托管式引擎

云提供商提供托管式无服务器 Spark 和 Ray 服务，以支持大规模数据处理、分析和机器学习。这些服务能够简化基础设施管理和提供优化的性能。

数据仓储

对于结构化数据和分析型工作负载，请考虑使用云数据仓储服务，以便为复杂查询提供高性能和扩缩能力。您还可以考虑使用 BigQuery 这样的统一数据平台，以利用 Apache Iceberg 等开放格式，针对结构化和非结构化数据运行分析和机器学习工作负载。

03 网络

VPC

使用虚拟私有云 (VPC) 隔离数据湖环境并控制网络访问。

数据传输

利用云提供商提供的专用连接服务，优化本地环境与云环境之间的数据传输。

04 安全

访问权限控制

使用 Identity and Access Management (IAM) 角色和政策实施精细的访问权限控制，限制对敏感数据的访问。

加密

使用云提供的加密服务，加密静态数据和传输中的数据。

威胁检测

使用云原生安全工具进行威胁检测和监控。

05 数据治理和元数据统一

数据治理

制定清晰的数据治理政策，并利用基于云的工具实施精细的访问权限控制和合规机制。对现代化策略进行长期规划时，请考虑您的云提供商是否具备从数据到 AI 的全面治理能力。

基于云的元存储

云提供商可能会为您的数据湖提供一个集中式元数据仓库，使您能够在一个统一、集成的组件中执行数据发现、治理、沿袭跟踪和质量监控。

06 迁移工具

数据迁移服务

云提供商提供各种数据迁移服务，其中包含数据验证、转换和架构映射等功能。探索适用于其他迁移场景的开源工具或商业 ETL 解决方案，同时权衡所迁移数据的容量和类型，以及期望的时间节点和可接受的停机时间。

批量数据传输

对于大型数据集，考虑使用云提供商提供的批量数据传输工具。增量数据迁移（仅传输更改的内容）有助于尽可能减少停机时间和保持数据一致性。

第 3 阶段

规划

周密规划对于顺利完成迁移至关重要。接下来，制定一个详实的项目计划，并将其用作活动文档，以便在执行迁移期间随时更新，记录实际发生的情况。

项目计划的关键要素包括：

时间表和里程碑

定义迁移过程中的关键里程碑，以及达到每个里程碑的预期时间表。采用分阶段的方法有助于尽可能减少中断。记录应首先迁移的关键数据和工作负载，并根据紧急程度和重要程度规划后续迁移阶段。

任务和依赖关系

将迁移过程分解为可管理的步骤，并记录每个阶段的依赖关系。

角色和职责

明确迁移过程中的主要利益相关方及其承担的角色。记录利益相关方的沟通渠道，并确定必要的上报程序。

预算

列出与迁移相关的预估费用，包括云资源、迁移工具和专业服务。

回滚方案

记录出现问题并影响关键业务流程时应采取的步骤。

第 4 阶段

执行

现在，就可以开始将数据湖的各个部分传输到云端了。这个阶段需注意的关键事项包括：

- 如何确保迁移阶段的数据完整性
- 如何尽可能减少停机时间和关键业务流程中断
- 如何验证迁移的数据和工作负载
- 如何处理可能出现的任何问题

在整个迁移过程中，请持续监控性能和可用性，以确保遵守服务等级协议 (SLA)。使用在规划阶段确立的沟通渠道，提醒利益相关方可能出现的问题或中断。让整个迁移工作与企业服务等级协议 (SLA) 保持一致，以顺利完成过渡、尽可能减少停机时间，并确保云环境中关键数据和应用的完整性。

我们来逐一了解迁移工作的主要组成部分。

数据

为确保数据准确和完整，请在迁移前、迁移中和迁移后验证数据完整性。这可能涉及比较校验和、针对架构验证数据，以及核对记录数量和其他统计信息。选择适合的数据迁移工具同样重要，因为云提供商会提供各种数据迁移服务，其中又包含数据验证、转换和架构映射等功能。最后要牢记，数据迁移不只是传输数据，它还涉及转换数据以适应新的云环境，以及针对云原生工具和服务对其进行优化。这可能涉及数据清理、架构更改和格式转换，以便在新的数据湖中有效使用数据。

元数据

迁移元数据非常重要，这可以帮助您持续执行数据发现和数据治理，以及实现数据在新环境中的价值和目标。充分利用云提供商提供的云原生元数据管理工具，这些工具通常提供自动执行元数据发现、分类和沿袭跟踪等功能。在迁移期间，保持来源元数据和目标元数据之间清晰的映射关系，这有助于保留数据沿袭，确保数据在云端可以准确解释和使用。最后，别忘了验证所迁移元数据的完整性和准确性。

治理

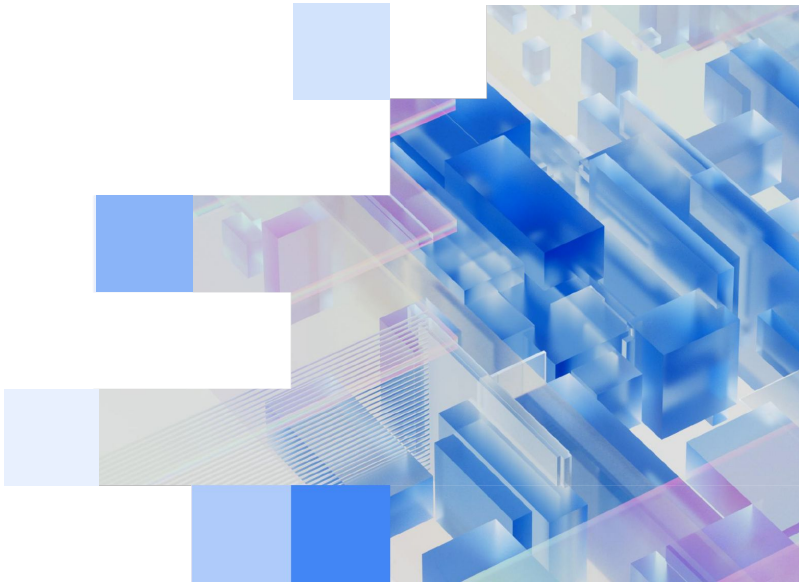
为了定义各种角色、创建用户群组并为云端数据湖执行访问权限控制政策，同时也为了顺利完成过渡和尽可能减少中断，请使用可提供对用户访问权限实施精细化控制的 IAM 服务。请务必在整个迁移过程中始终如一地执行数据访问政策，并根据您的治理政策来保护敏感数据；定期审核用户访问和权限，以识别和消除任何潜在的安全风险。通过在迁移过程中采取这些步骤，您可以打造一个更安全、更合规的环境，同时让授权用户能够访问所需数据以高效完成工作。

工作流迁移

迁移并调整工作流和 DAG 有助于确保数据处理和机器学习流水线的连续性，以充分利用云端编排服务的扩缩能力和效率。请考虑使用 Cloud Composer 这类工作流编排服务，以充分利用可以定义、调度和监控复杂数据流水线的托管式服务。这些服务通常提供可视化 DAG 编辑器、版本控制和监控等功能。同时，您还可以考虑重构工作流，以针对云环境对其进行优化。全面测试云环境中迁移的工作流和 DAG，确保其按预期运行并能满足性能要求。

工作负载

重构工作负载可帮助您充分利用云原生服务和优化性能。您应考虑使用托管式 Spark、Flink 和 Ray 服务进行无服务器计算，或针对大规模数据处理使用无服务器数据仓储。根据业务关键度和依赖关系确定工作负载迁移优先级，通过分阶段迁移来降低风险和尽可能减少停机时间。全面测试云环境中迁移的工作负载，确保其功能、性能和数据完整性；利用云监控工具来帮助识别任何瓶颈或优化机会。



第 5 阶段

优化

成功迁移数据湖后，工作还没有结束。迁移到云端后，您可以就成本和性能对新环境进行持续优化。以下是主要优化方法：

性能优化

持续监控和微调数据湖性能，利用云原生工具识别瓶颈和优化查询。

费用优化

节省费用的方法包括：采用无服务器产品，使用适当的实例类型，利用 Spot 实例，以及优化数据存储。

加强监督和治理

使用云原生工具进行数据质量监控、沿袭跟踪和合规性自动化，进一步加强数据治理。

使用云原生服务

持续探索和采用新的云服务和功能，进一步增强数据湖的可扩缩性、可靠性、安全性和处理能力。

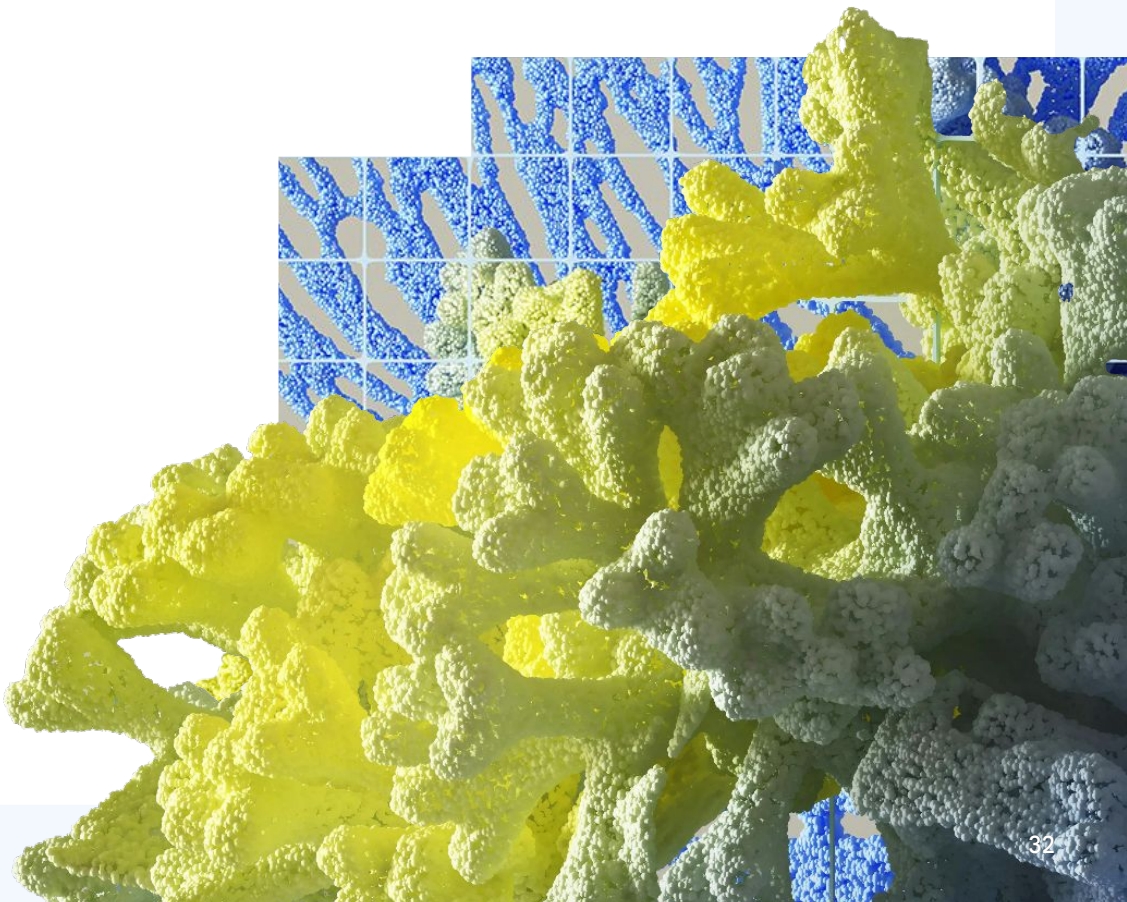
Accenture 助力 General Mills 将迁移数据湖到 Google Cloud 的速度提升 30%

General Mills 的本地数据和分析生态系统已无法满足业务发展需求，为加速 AI 创新，公司决定将数据湖迁移到BigQuery，并携手 Accenture 来帮助他们将这一愿景变为现实。

在 Accenture 的支持下，原计划需要近三年才能完成的转型项目，实际上仅用了 21 个月就顺利落地，速度提升达30% 以上。

> **“在执行迁移策略的过程中，Accenture 发挥了关键性作用。无论是 Google Cloud，还是我们要采用的更广泛的架构，他们都有非常丰富的实践经验，这方面的洞见对我们大有帮助。”**
>
> Jason Staloch
> **数字核心部副总裁，**
> General Mills

阅读完整案例 →



帮助您制定规划的核对清单

本清单全面概述了将数据湖迁移到云端的主要注意事项，请注意根据您的实际需求和要求对其进行调整。

下载核对清单 →

第 4 章

Google 如何提供帮助

Google Cloud 提供了一整套服 务和资源，可以帮助您的组织实施数据湖迁移 项目，并加速 实现数据驱动的创新。

专属 Google Cloud 客户支持团队

Google Cloud 客户支持团队提供个性化指导和支持，助力您顺利完成整个迁移旅程。他们将帮助您评估当前环境，并制定云策略和量身定制的迁移计划。他们还充当单一联系人，负责协调各种 Google Cloud 资源，确保您的项目按计划推进。

Google 自动迁移工具

Google Cloud 提供自动迁移工具，以简化将数据和工作负载迁移到云端的过程。例如，BigQuery 迁移服务可以简化各种数据源（包括本地数据湖和数据仓库）的发现过程，以及将数据从这些来源传输到 Google Cloud 的过程。

Google 咨询服务

对于复杂的迁移场景，Google 咨询服务可以提供专家指导和实操支持。Google Cloud 的顾问拥有丰富的实战经验，他们帮助各行各业的客户成功完成了许多复杂且规模庞大的数据湖迁移项目。

无论是对迁移项目的最佳实践，还是对其中所涉及的挑战和细节，他们都有深刻的理解，能够帮助您评估需求，设计云架构，迁移数据和工作负载，还可以优化您的数据湖以实现最优性能和成本效益。同时，他们还能帮助您降低与复杂迁移相关的风险，加快迁移进程，并尽可能降低整体迁移费用。最后，Google 咨询服务可以帮助您成功完成迁移，使其契合您的业务目标，从而获得更理想的业务成果。

Migration Black Belt 计划

为了解决迁移过程中的技术难题并加快迁移进程，Google Cloud 会安排经验丰富的工程师提供技术支持，包括检查架构以优化云数据湖设计，从而实现最佳的性能、扩缩能力和成本效益。Google Cloud 的 Migration Black Belt 工程团队还可以为云端数据和工作负载迁移、安全和治理提供最佳实践指导。

财务激励

为了帮助您减少前期迁移费用，Google Cloud 提供了各种财务激励计划，包括为云服务的使用提供了赠金，以帮助您抵消同时运行两种系统的费用，还包括为您承担部分前期迁移费用。

Google Cloud 合作伙伴计划

Google Cloud 合作伙伴计划为您提供了一个庞大的合作伙伴网络，这些合作伙伴都已经过认证，在数据湖迁移和云技术领域拥有深厚的专业知识。这些合作伙伴提供一系列服务，包括迁移规划、迁移实施、持续支持以及托管式服务。

您是否符合加入迁移激励计划的资格？

在此查看答案 →

为什么要将数据湖迁移到 Google Cloud？

相较于其他云提供商，将数据湖迁移到 Google Cloud 能够给您带来更多显著优势。

卓越的数据分析能力

Google Cloud 在数据分析和AI 领域一直处于领先地位，能够为您提供一系列强大的工具，例如融入了AI 技术的统一数据平台 BigQuery，以及极具成本效益的全托管式 Spark 服务 Dataproc。这些工具能够实现更快速、更高效的数据处理和分析，并且适用于所有数据类型，让您能够从数据中发掘富有价值的分析洞见。

开放灵活的生 态系统

Google Cloud 积极拥抱开源技术，并提供灵活的生态系统，可支持各种数据湖架构和工具。这意味着您可以充分利用现有投资，同时根据您的特定需求选择最佳解决方案。

创新和 AI/机器学 习

Google Cloud 一直走在 AI 和机器学习创新前沿，提供预训练模型、自定义模型开发工具和专用 AI/机器学习基础设施，让您可以利用这些工具来构建智能应用。

经济高效

Google Cloud 提供极具竞争力的价格以及各种费用优化工具和策略，帮助您管理云开支并充分发挥投资价值。

安全与合 规性

Google Cloud 对安全与合规性做出了坚定承诺，提供一整套安全功能和认证，以保护您的数据并满足监管要求。

全球基 础设施

Google Cloud 的全球基础设施为数据湖提供高可用性、低延迟和可扩缩性，确保可靠的性能，让您能够在世界任何地方访问您的数据。

准备好迈出下一步了吗？

要开启数据湖迁移旅程，请联系 Google Cloud 团队。诚邀您参与我们的限时激励计划，加速您的迁移之旅！

立即填写此表单，**借助 Google Cloud 释放数据的全部潜力。**

贡献者：

Angela Soares
产品营销经理，
Google Cloud

Adnan Hasan
AI 分析数据平台
GTM 策略师，
Google Cloud

Sajal Agarwal
Senior Outbound
Product Manager，
Google Cloud

Jill Hardy
Product Marketing
Manager，
Google Cloud

核对清单一览

数据迁移核对清单

第 1 阶段

发现

清点数据资产

- ▢ 对所有数据源整理分类（数据库、表格、文件等）
- ▢ 确定数据容量和类型
- ▢ 评估数据质量并识别潜在问题
- ▢ 记录数据沿袭和依赖关系

分析工作负载

- ▢ 确定访问数据湖的工作负载和流程
- ▢ 确定数据访问模式和频率
- ▢ 分析每个工作负载的性能要求

记录用户和权限

- ▢ 识别有权访问数据湖的所有用户和群组
- ▢ 记录每个用户/群组的访问级别和权限

绘制工作流

- ▢ 记录数据流水线和 ETL 流程
- ▢ 确定常用数据转换和扩充步骤
- ▢ 分析系统之间的数据沿袭和依赖关系

第 2 阶段

评估

选择云提供商

根据以下标准评估云提供商：

- ▢ 提供的产品（存储、计算、分析、AI）
- ▢ 定价模式和费用优化方案
- ▢ 安全功能和合规性认证
- ▢ 扩缩能力和性能
- ▢ 地理位置可用性和数据驻留要求

制定迁移策略

- ▢ 直接原样迁移
- ▢ 迁移并优化
- ▢ 现代化改造

确定云架构

- ▢ 设计云端目标数据湖架构
- ▢ 选择适合的云存储解决方案（例如对象存储、数据仓库）
- ▢ 确定数据处理所使用的计算资源（例如无服务器产品、虚拟机）

制定治理政策

- ▢ 确定数据安全政策和访问权限控制机制
- ▢ 建立数据质量标准和验证程序
- ▢ 确保遵守相关法规（PCI、GDPR、HIPAA 等）
- ▢ 实施数据保留和删除政策

确定成本控制机制

- ▢ 估算存储、计算和其他服务的云费用
- ▢ 实施费用优化策略（例如，合理调整资源和预留实例的规模）
- ▢ 建立预算监控和提醒机制

第 3 阶段

规划

制定迁移计划

- 确定数据和工作负载迁移优先级
- 列出迁移任务和依赖关系
- 确定时间表和里程碑

分配角色和职责

- 明确迁移过程中的主要利益相关方及其承担的角色
- 明确沟通渠道和上报程序

设置预算

- 为云资源、迁移工具和专业服务分配预算
- 针对预算跟踪和监控迁移费用

第 4 阶段

执行

迁移数据

- 使用适当的工具和流程迁移数据
- 确保迁移期间的数据完整性和一致性
- 验证迁移后的数据与源数据

迁移用户和权限

- 配置云端用户身份验证和授权
- 将用户角色和权限迁移到新环境

迁移工作负载

- 迁移访问数据湖的工作负载和流程
- 配置云端数据流水线和 ETL 流程
- 测试并验证迁移的工作负载

监控和问题排查

- 持续监控迁移过程
- 解决迁移过程中出现的问题或错误
- 记录迁移进度和遇到的任何问题

第 5 阶段

优化

优化性能

- 微调数据存储和处理能力以获得最佳性能
- 利用云原生服务增强性能
- 监控并分析性能指标以识别瓶颈

管理费用

- 持续监控云费用并识别优化机会
- 根据需求合理调整资源规模和使用量
- 使用云提供商提供的费用管理工具和服务

增强安全性

- 实施数据保护安全性最佳实践
- 使用云原生安全功能
  （加密、访问权限控制、威胁检测）
- 定期检查并更新安全政策和程序