设计数据湖以实现增长和扩展 AWS Cloud

AWS 规范性指导

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS 规范性指导: 设计数据湖以实现增长和扩展 AWS Cloud

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon 的商标和商业外观不得用于任何非 Amazon 的商品或服务，也不得以任何可能引起客户混淆、贬低或诋毁 Amazon 的方式使用。所有非 Amazon 拥有的其他商标均为各自所有者的财产，这些所有者可能附属于 Amazon、与 Amazon 有关联或由 Amazon 赞助，也可能不是如此。

Table of Contents

简介 1
目标业务成果 2
常见的扩展挑战 3
初始数据湖部署 3
数据使用者增加 4
数据生产者增加 5
参考架构 6
参考架构组件 8
数据生产者 9
集中式目录 11
AWS Lake Formation 12
AWS Glue 12
数据使用者 13
入职和授予访问权限 16
入职数据制作者 16
引导数据使用者 16
在数据使用者账户中授予 Select 访问权限 17
常见问题解答 19
此数据湖参考架构是否更适用于企业组织？ 19
如果我的组织只有一个数据创建器，我还能使用这个参考架构吗？ 19
我的数据湖直接将一个数据创建器与多个数据使用者连接起来。本指南的数据湖参考架构仍然相关吗？ 19
我的组织是否应该在不进行更改的情况下遵循入职和访问权限授予工作流程？ 19
资源 20
文档历史记录 21
..... xxii

设计数据湖以实现增长和扩展 AWS Cloud

Wei Shao，Amazon Web Services (AWS)

2021 年 10 月（文档历史记录）

作为其数据现代化战略的一部分，各组织越来越多地在 Amazon Web Services (AWS) 云上设计和构建数据湖。数据湖可以是存储库，用于存储任何规模的结构化和非结构化数据，并可供广泛的内部和外部用户使用。

但是，要将数据摄取到可满足您不断增长的业务用例的数据湖中，可能需要花费大量时间和精力。为了帮助降低成本并最大限度地提高数据产生的价值，许多组织计划一次性地将数据引入其数据湖，然后多次使用这些数据。设计可随数据生产、共享和消费而扩展的数据湖架构，对于在数据湖增长时为业务利益相关者创造价值至关重要。

可扩展的数据湖架构为您的组织提供了坚实的基础，让您可以从数据湖中获得价值，同时将更多数据带入其中。通过持续获得数据见解，而不会因为可扩展性限制而减慢或中断，可扩展的数据湖还可以帮助您的组织保持竞争力。

通常，数据湖有数据生产者和数据使用者。数据生产者通过收集、处理和存储其数据域中的数据来创建数据资产。这些集体数据资产构成了数据湖的内容。数据生产者可以选择有选择地与数据湖的数据使用者共享其数据资产。

数据使用者需要来自数据生产者的数据来满足其业务用例，偶尔也可以将这些数据与自己的数据结合起来。数据生产者和数据使用者通常是组织的一部分，但并非总是如此。重要的是，他们可以同时成为数据生产者或数据消费者。

可扩展的数据湖架构可帮助您实现以下成果：

- 大规模加入数据生产者，无需他们维护整个数据共享流程。这可以帮助数据生产者将其数据载入数据湖，并专注于从其数据域收集、处理和存储数据。
- 使数据使用者能够访问来自多个数据生产者的数据，而不会增加您的总体成本和管理开销。

本指南描述了组织在发展数据湖时可能出现的常见扩展挑战，提供了数据湖参考架构，并概述了数据生产者和数据使用者入职和授予访问权限的方法。本指南中的数据湖参考架构利用了提供的AWS Lake Formation不同特性和功能。该指南适用于负责在上设计数据湖的团队 AWS Cloud，包括企业数据架构师、数据平台架构师、设计师或数据领域负责人。

目标业务成果

在设计用于增长和扩展的数据湖之后，您应该期望获得以下三个结果 AWS Cloud：

- 降低组织中多条业务线的数据共享和数据消耗开销。
- 一种安全、一致的方法，可帮助您的组织包括外部数据生成者并在您的数据湖中与他们共享数据。
- 持续获取数据见解，而不会因为可扩展性限制而减慢或中断。

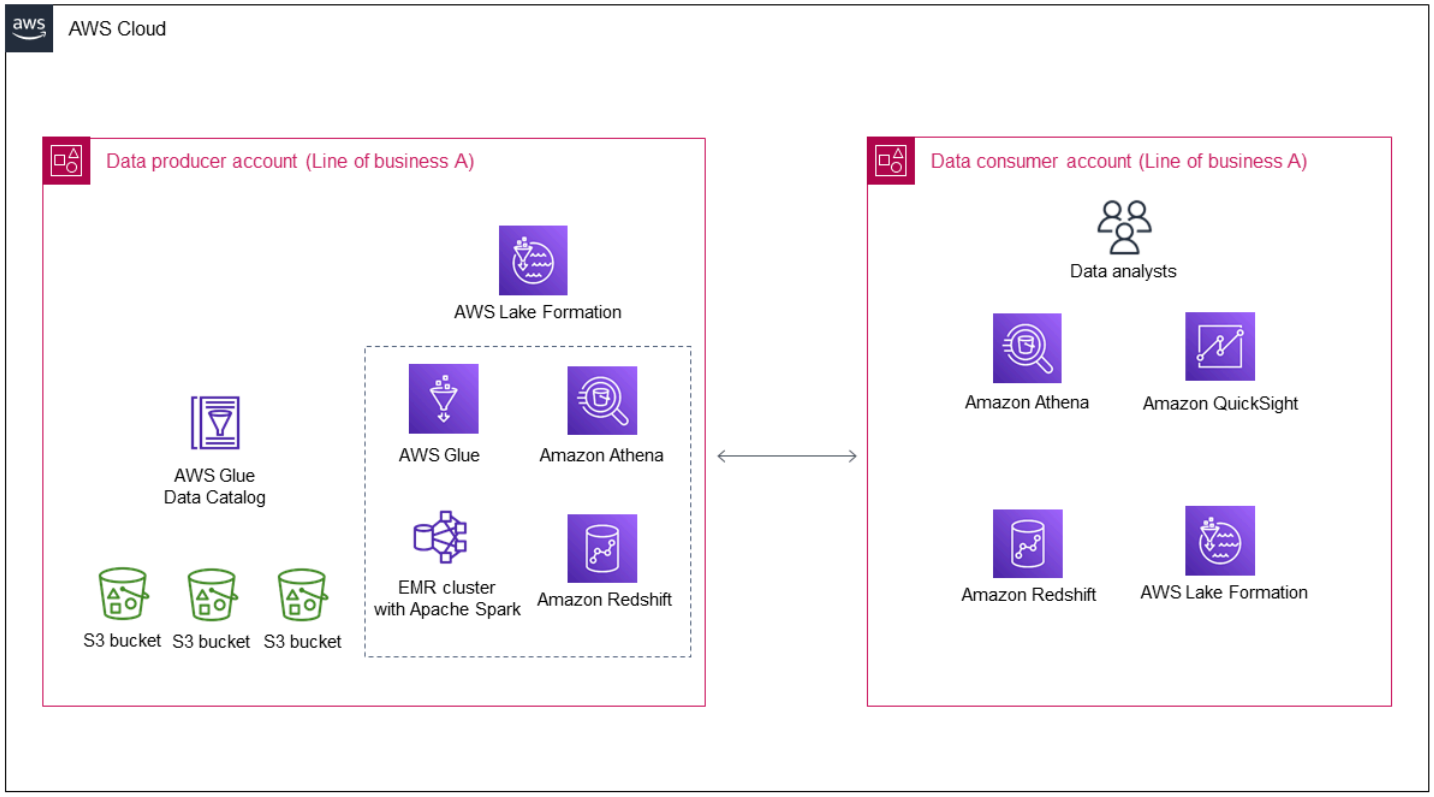
常见的扩展挑战

数据湖在初始部署后数据增长时会经历几个阶段。如果您没有使用可扩展架构来设计数据湖，您的组织可能会遇到挑战，并可能因数据湖的增长而处于不利地位。

以下各节说明了典型的数据湖的增长如何导致扩展挑战。

初始数据湖部署

下图显示了业务线 A 对数据湖进行初始部署后的架构。



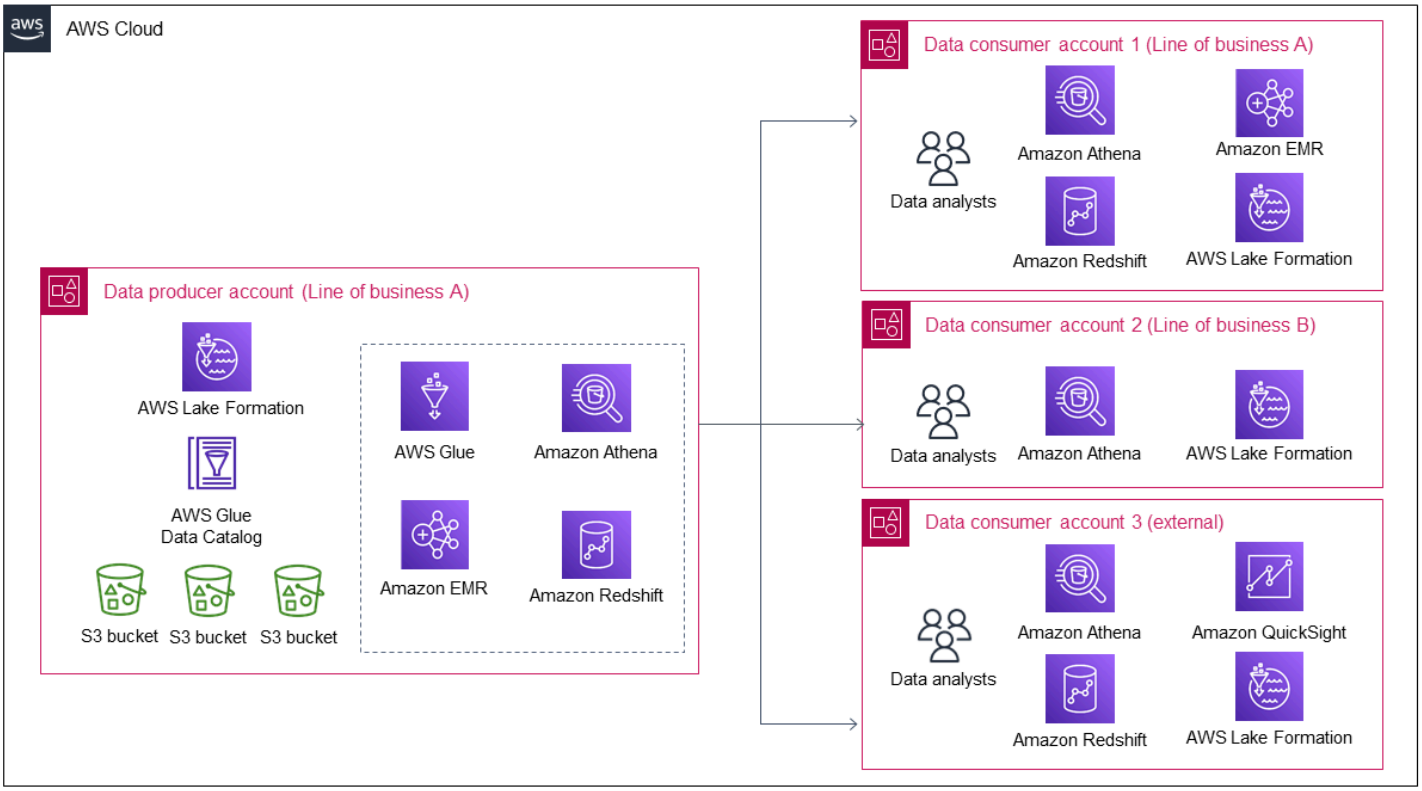
该图显示了以下组件：

- 数据生产者账户收集和处理数据，存储处理后的数据，并为其做好使用准备。
- 数据创建者账户中的数据存储在亚马逊简单存储服务 (Amazon S3) S imple S3 存储桶中，该存储桶可以有多个数据层。
- 您可以使用 AWS 服务进行数据处理（例如，AWS Glue和 Amazon EMR）。

- 数据创建者不仅在数据湖中生成和存储数据，还需要决定与数据使用者共享哪些数据以及如何共享这些数据。 AWS Lake Formation 管理数据生产者账户中的数据湖，此外还管理从数据创建者到数据使用者的跨账户数据共享。
- 数据使用者账户使用来自数据生产者账户的共享数据，用于特定的业务用例。

数据使用者增加

下图显示，当业务线 A 的数据增长时，会有更多的数据被带入数据湖。然后，数据湖吸引了更多的数据使用者来利用数据并从中获得价值。

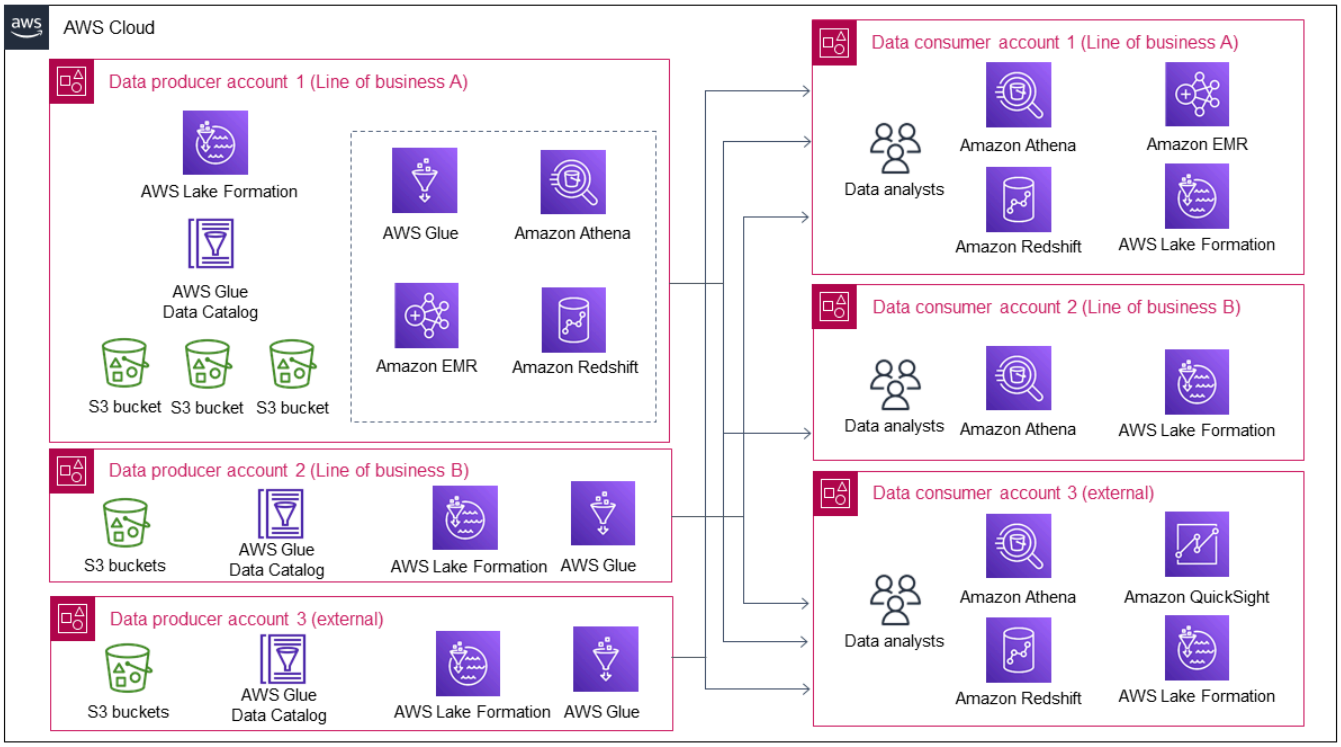


该图显示了组织如何从现有数据资产中产生近乎持续的价值，以及如何吸引更多的数据使用者。但是，当数据使用者增加时，数据生产者只有以下两个选择来适应这种增长：

- 手动管理个人数据使用者的数据共享和访问，这不是一种可扩展的方法。
- 开发用于数据共享和管理数据访问的自动化或半自动流程。尽管这可能是一个可扩展的选项，但由于内部和外部数据使用者的安全控制要求不同，因此需要大量的时间和精力来设计和构建。将来，任何解决方案改进都需要额外的时间和精力。

数据生产者增加

下图显示了多条业务线作为数据生产者加入时的数据湖架构。



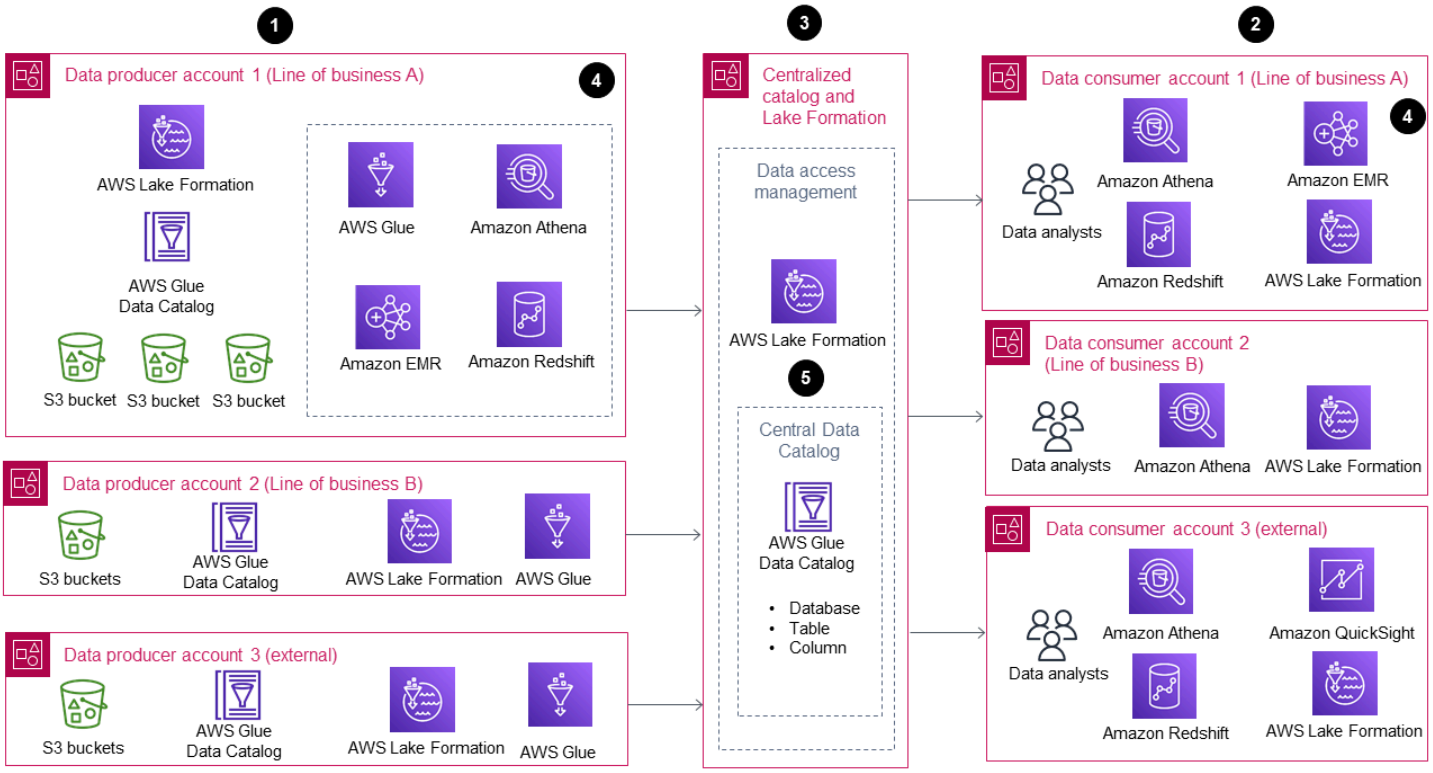
即使只有三个数据生成器和三个数据使用者，数据湖的架构也变得越来越复杂。

每个数据创建者都需要为多个数据使用者处理数据共享和数据访问管理。期望所有数据生产者都开发一种用于数据共享和数据访问管理的自动化或半自动流程是不现实的。一些数据生产者可能会选择不共享数据，从而避免负担不起的管理开销。同样，每个数据使用者都需要与多个数据生产者进行交互，以了解他们不同的数据消费流程。这意味着个人数据使用者在处理不同的数据共享模式时面临着不断增加的管理开销。

在许多组织中，这种数据湖会导致瓶颈，无法增长或扩展。这可能意味着您的组织必须重新设计和重建其数据湖以消除瓶颈，这可能会花费大量时间、资源和金钱。

参考架构

下图显示了本指南用于在上扩展和扩展数据湖的参考架构 AWS Cloud。



该图显示了以下组件：

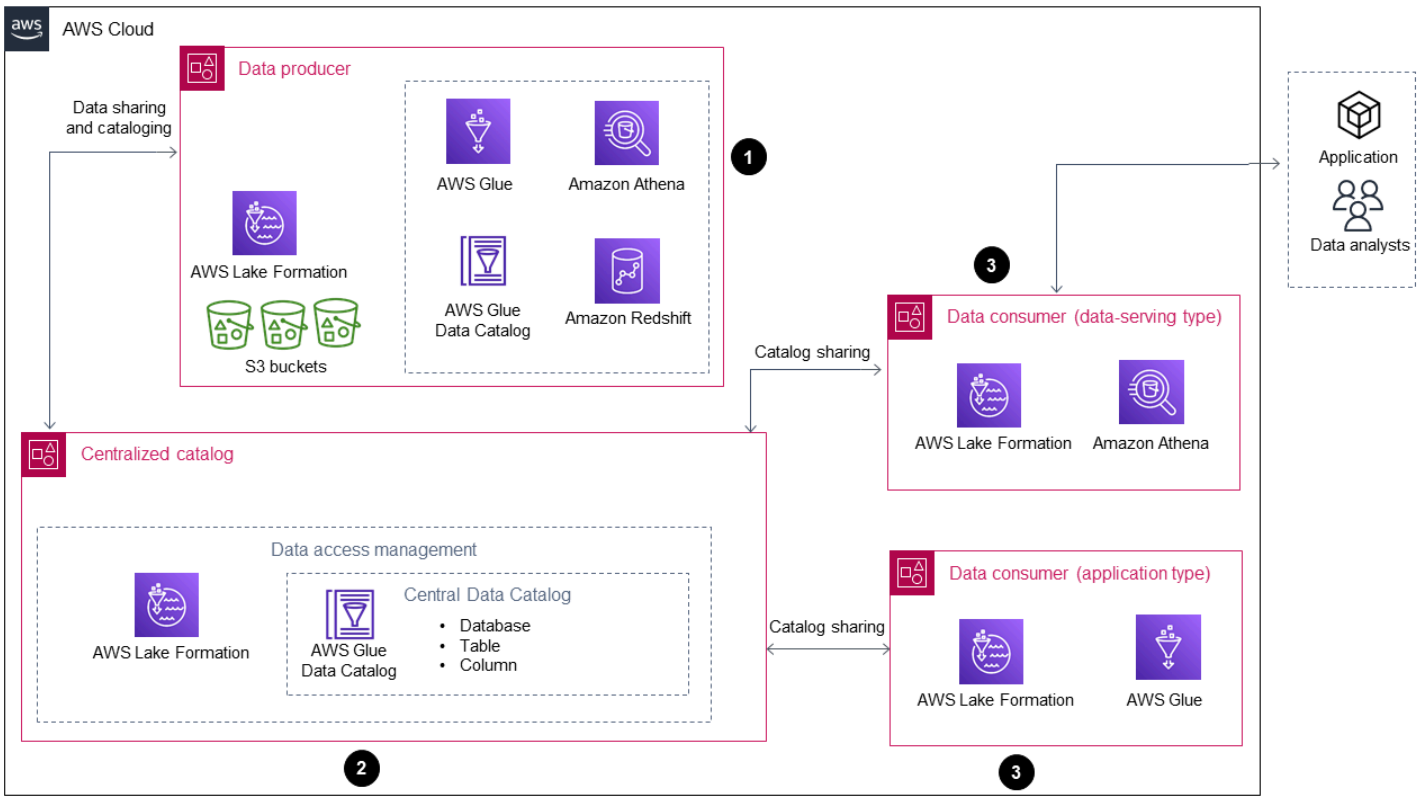
1. 不同的数据创建器层 AWS 账户。
2. 数据使用者层与众不同 AWS 账户。
3. 中的集中式目录 AWS 账户。
4. 尽管每个业务领域只有一个数据创建者和一个数据使用者，但该指南的参考架构支持每个业务领域的多个数据创建者和数据使用者。通常会将一个数据创建器与一个或多个数据使用者结合起来，这些使用者包括数据服务和应用程序类型。有关这方面的更多信息，请参阅本指南的参考架构组件部分。
5. 集中式目录是数据生产者和数据使用者用来共享和使用数据的接口。

参考架构的方法使标准化数据共享和使用成为可能，并在不增加管理开销的情况下独立扩展数据生成器和数据使用者。参考架构还支持在不同的数据生产者之间生成和分发数据。任何数据生产者都可以成为数据湖的一部分，共享他们的数据，并为数据湖提供的整体价值做出贡献。

这种方法使您的组织能够在整个业务线和外部数据所有者中获得数据价值，而不会因为将数据收集和处理限制在单个管道中而造成瓶颈。

参考架构组件

下图显示了本指南参考架构的三个组成部分：数据生成器、数据使用者和集中式目录。



该图显示了以下组件：

1. 数据生产者账户。
2. 集中式目录。
3. 两个数据使用者帐户，包括数据服务和应用程序类型。

> **Note**
>
> 该图仅显示了一个数据创建器和两个数据使用者，但是当数据湖增长时，本指南的参考架构支持多个数据生产者和数据使用者。

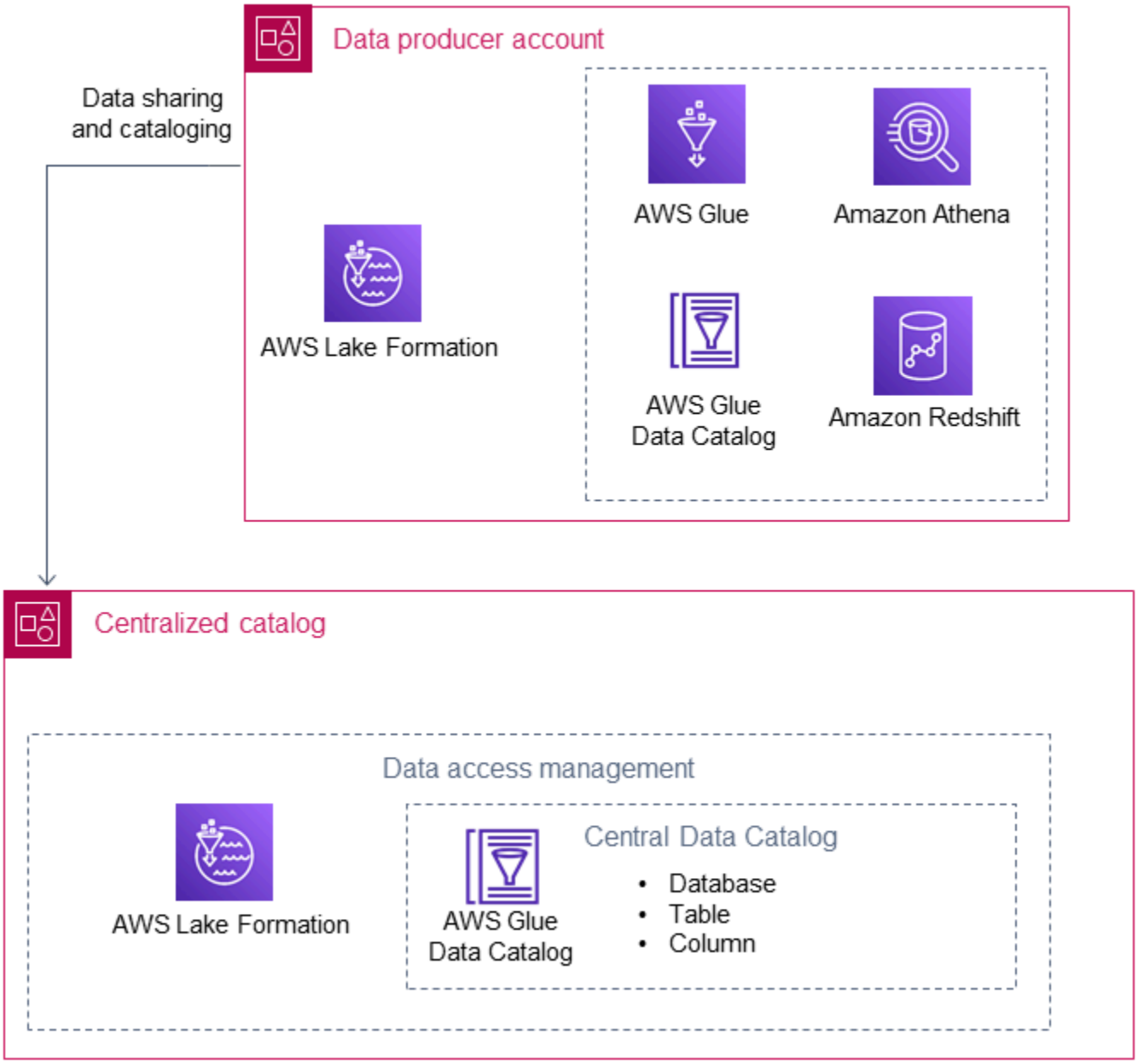
了解每个组件有助于您的组织实施可重复使用的模式，并通过独立且持续地引导新的数据生成者和数据使用者来扩展数据湖。

以下各节讨论了参考架构的组件：

- 数据生产者
- 集中式目录
- 数据使用者

数据生产者

数据生产者除了监控和确保其数据资产的质量外，还会收集、处理和存储其数据域中的数据。下图显示了作为本指南参考架构组成部分的数据生产者帐户。



每个数据创建者都有一个由管理 AWS Lake Formation 的私有数据目录 AWS 账户 ，供其内部数据处理使用。数据生成者为集中式目录提供对其数据的选择性权限，这意味着集中式目录账户中的 Lake Formation 可以访问数据创建者想要共享的数据。

这意味着数据生产者不直接与数据使用者互动。相反，数据生产者帐户及其数据存储位置是完全抽象的，对数据使用者是隐藏的。这种方法通过消除数据使用者增加的数据生产者不必要的开销来降低成本。

如果新的数据位置是由集中式目录注册的，则更改数据创建者的数据位置不会影响数据使用者。如果数据创建者想要停止共享特定数据，他们可以移除集中式目录的权限。这样可以防止数据使用者访问数据，也无需手动撤消每个数据使用者的访问权限。

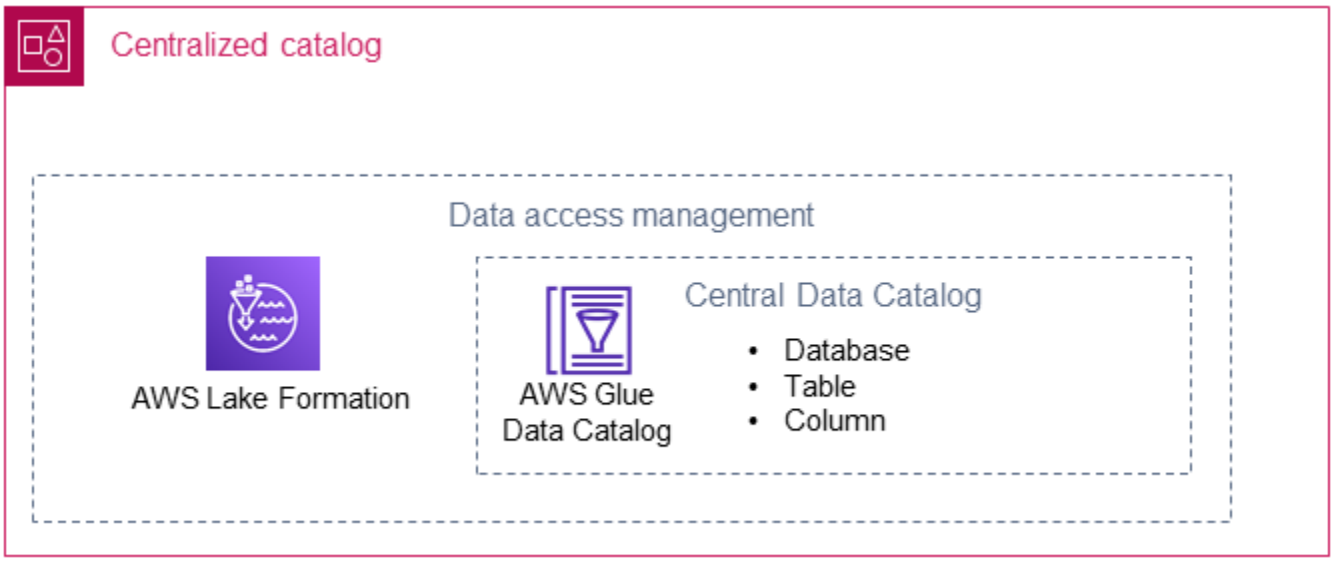
通过使用公共和私有数据目录，数据生产者可以选择与数据使用者共享的内容，同时通过私有数据目录独立管理内部数据访问权限。

下表描述了数据创建者用来与集中式目录共享数据的两种 AWS 服务。

Amazon Simple Storage Service（Amazon S3）	调整 S3 存储桶的存储桶策略，为集中式目录中的 AWS Identity and Access Management (IAM) 角色提供数据访问权限。 数据创建者还可以使用 Amazon S3 作为中间数据层，共享存储在其他数据应用程序或服务中的数据。
AWS Key Management Service (AWS KMS)	为集中式目录中的 IAM 角色提供 AWS 托管密钥的权限，以及 AWS KMS keys 用于加密数据创建者账户中共享的 Amazon S3 数据的权限。

集中式目录

下图显示了集中式目录如何连接数据湖中的数据生产者和数据使用者。



集中式目录存储和管理数据创建者账户的共享数据目录。集中式目录还托管共享数据的技术元数据（例如表名和架构），也是数据使用者访问数据的地方。

数据使用者可以在集中式目录中访问来自多个数据生产者的数据，然后可以将这些数据与自己的数据混合以进行进一步处理。使用集中式目录使数据使用者无需直接与不同的数据生产者建立联系，并减少了运营开销。

由于集中式目录可以查看数据创建者和使用者的数据共享和数据使用情况，因此它可能是应用集中式数据治理功能（例如访问审计）的理想位置。

以下各节介绍集中式目录如何使用 AWS Lake Formation 和 AWS Glue。

AWS Lake Formation

AWS Lake Formation帮助在 AWS Glue 数据目录中创建指向数据湖中多个数据生产者位置的数据库。在集中式目录中为 Lake Formation 创建了一个 AWS Identity and Access Management (IAM) 角色。通过使用 Lake Formation，集中式目录可以有选择地与数据使用者共享数据资源（例如数据库、表或列）。使用以下两种方法之一与数据使用者共享 Lake Formation 托管资源：

- 命名资源方法-此方法跨账户共享托管资源。必须指定数据库、表或列名，并且可以将资源共享给组织、组织单位 (OU) 或 AWS 账户。为了减少共享和管理开销，我们建议您尽可能在更高级别上共享资源（例如，在组织或 OU 中而不是 AWS 账户）。但是，您必须确保这种方法符合贵组织的数据安全控制要求。
  - 注意：此方法适用于具有应用程序类型的数据使用者，其中 AWS 服务使用来自数据创建器的数据。这类数据使用者的数据访问要求是应用程序驱动的、规范性的，而且相对静态。
- 基于 Lake Formation 标签的访问控制 (LF-TBAC) 方法 — LF-TBAC 对于具有数据服务类型的数据使用者特别有用。但是，带有 Lake Formation 标签的资源目前只能在 AWS 账户 级别共享，不能在组织或 OU 级别共享。

AWS Glue

您必须在中 AWS Glue 为集中式目录中的每个数据创建者创建数据库。由于集中式目录用于 AWS Glue 托管来自所有数据生产者的数据库，因此必须确保数据库名称在所有数据生产者中都是唯一的，并且它反映了数据创建者及其数据类型。例如，您可以使用以下数据库命名结构：`<Data_Producer>-<Environment>-<Data_Group>`

- `<Data_Producer>`— 数据生产者的姓名。
- `<Environment>`— 数据湖环境，`dev`例如开发环境、`sit`系统集成测试环境或`prod`生产环境。

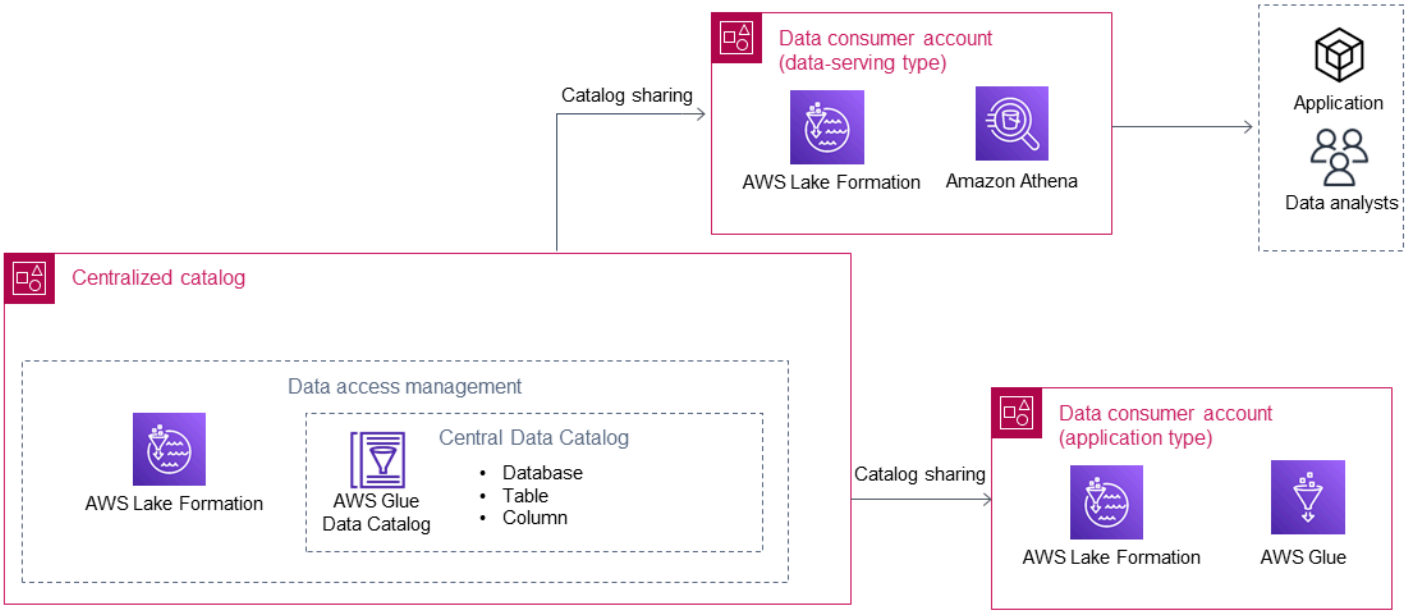
- <Data_Group>— 用于将数据来自数据创建者的数据分成逻辑组的数据组的名称。您可以使用源系统的名称、ID 或缩写作为名称。数据库描述还有助于描述数据库的内容和用途。

您可以对数据创建者的数据使用 AWS Glue 爬虫，以便在集中式目录的数据库中维护其架构。如果数据创建者定期以相同的频率创建数据，则可以使用单个 AWS Glue 抓取器。在所有其他情况下，您应该使用多个 AWS Glue 爬虫来适应不同的抓取频率。根据您的业务用例，可以将爬网程序安排为预定义的频率，也可以由事件启动。

您还可以 AWS Glue 通过调用 AWS Glue API 来创建或更新架构来维护中的表架构。尽管这可以提供灵活性，但在代码开发和维护方面还需要付出额外的努力。请务必评估用例和业务价值，然后选择符合您的要求且开销最少的选项。

数据使用者

在集中式目录使用数据创建器共享数据后，数据使用者使用这些数据 AWS Lake Formation。下图显示了数据湖中的两个数据使用者。



有两种类型的数据使用者：应用程序和数据服务。下表描述了这两种类型。

应用程序类型	应用程序数据使用者自己运行应用程序 AWS 账户。应用程序使用 AWS Identity and Access Management (IAM) 角色来访问来自数据生产者的共享数据，然后根据其逻辑对其进行处理。

	通常，这种类型的数据使用者具有规范性的数据要求来满足应用程序的需求。
Data-serving type	Data-serving 数据使用者通常适用于没有自己的个人（例如数据分析师或数据科学家）和应用程序（例如商业智能应用程序） AWS 账户。 一个组织的数据湖中可以存在多个提供数据服务的数据使用者。例如，不同的业务部门可能会选择设置自己的数据服务使用者，以帮助用户使用来自数据湖的数据。这些数据使用者在其 AWS 账户 （例如，与之关联的 IAM 角色 AWS IAM Identity Center）中配置了自己的 IAM 角色主体，数据使用者账户中的最终用户使用这些角色主体通过 AWS 服务（例如 A mazon Athena）访问共享数据。 通常，这种类型的数据使用者具有广泛且不断增长的数据需求。

AWS Lake Formation 是数据使用者用于跨账户数据共享和访问集中式目录的最重要的 AWS 服务。集中式目录共享数据库后，共享资源将在 Lake Formation 中的数据使用者帐户中可用。然后，如果需要，可以向数据使用者账户中的本地 IAM 委托人授予数据访问权限，并获得数据创建者的许可。然后，与 Lake Formation 集成的 AWS 服务（例如 Amazon Athen AWS Glue a 和）即可使用共享的数据。您可以使用以下 AWS 服务访问数据使用者账户中的共享数据：

- Amazon Athena 是一项交互式查询服务，可帮助使用标准 SQL 直接分析亚马逊简单存储服务 (Amazon S3) 中的数据。有关 Athena 和 Lake Formation 的更多信息，请参阅亚马逊 Athena 文档中的 Athena 如何访问在 Lake Formation 注册的数据。
- Amazon Redshift Spectrum 可帮助您有效地查询和检索亚马逊 S3 中的文件中的结构化和半结构化数据，而无需将数据加载到亚马逊 Redshift 表中。有关 Redshift Spectrum 和 Lake Formation 的更多信息，请参阅亚马逊 Redshift 文档中的将 Redshift Spectrum 与 Lake Formation 配合使用。
- AWS Glue是一项完全托管的提取、转换和加载 (ETL) 服务，它可以简单且经济高效地对数据进行分类、清理、丰富数据，并在不同的数据存储和数据流之间可靠地移动数据。如果 AWS Glue ETL 作业的关联 IAM 角色具有所需的访问权限，则可以访问由 Lake Formation 管理的数据湖数据。

- 亚马逊 EMR 可帮助运行大数据框架（例如 Apache Hadoop 和 Apach e Spark）来处理和分析大量数据。有关亚马逊 EMR 和 Lake Formation 的更多信息，请参阅亚马逊 EMR 文档中的将亚马逊 EMR 与 Lake Formation 集成。
- Amazon Quick 是一项可扩展、无服务器、可嵌入且由机器学习 (ML) 提供支持的商业智能服务，可用于分析和可视化来自数据湖的数据。有关 Quick 和 Lake Formation 的更多信息，请参阅快速文档中的通过 Lake Formation 授权连接。
- Amazon SageMaker AI Data Wrangler（Data Wrangler）缩短了为机器学习汇总和准备数据所需的时间。有关 Data Wrangler 和 Lake For mation 的更多信息，请参阅亚马逊 AI 文档中的使用亚马逊 A SageMaker I Data Wrangler 准备机器学习数据。 SageMaker

入职和授予访问权限

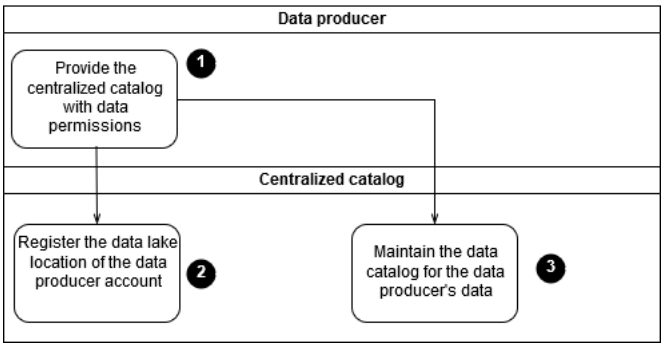
本指南的数据湖参考架构可帮助您独立扩展数据生成者和数据使用者，此外还可以定义和建立一致的入职流程并向这些数据使用者授予访问权限。

以下各节描述了数据生成者和数据使用者的入职流程以及如何在数据使用者账户中授予访问权限。本指南使用集中式目录和数据使用者之间的命名资源方法。该LF-TBAC 方法的过程相似，但略有不同。我们建议您评估和配置这些方法，以满足贵组织的数据治理实践和政策。

有关这两种方法的更多信息，请参阅本指南的集中式目录部分。

入职数据制作者

下图显示了如何将新的数据创建器载入数据湖。

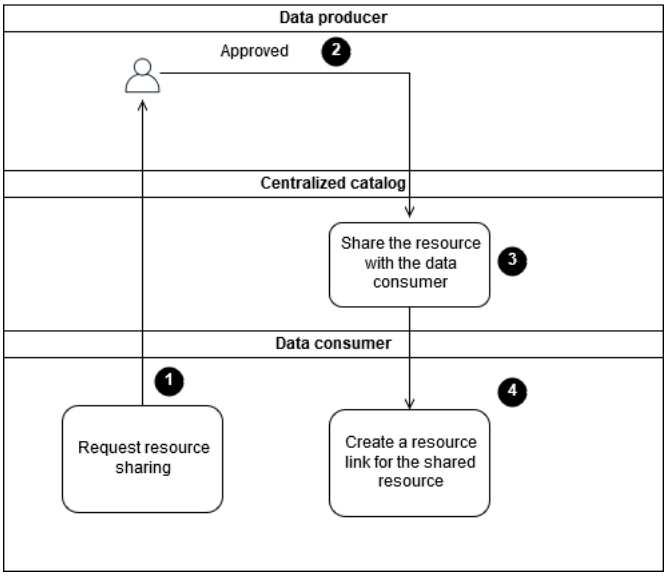


该图显示了以下入职流程：

1. 数据创建者有选择地向集中式目录提供对其数据的访问权限（例如，亚马逊简单存储服务 (Amazon S3) Service 存储桶和）。 AWS KMS key向集中式目录 AWS Identity and Access Management (IAM) 委托人提供访问权限，以在中注册数据创建者的数据湖位置 AWS Lake Formation 以及用于维护数据创建者目录的 IAM 委托人。
2. 使用集中式目录的 Lake Formation 注册数据创建者的数据湖位置（例如 S3 存储桶）。
3. 在数据目录中为来自数据创建器的新数据创建数据库、表和表架构。 AWS Glue

引导数据使用者

下图显示了如何将新的数据使用者载入数据湖。



该图显示了以下入职流程：

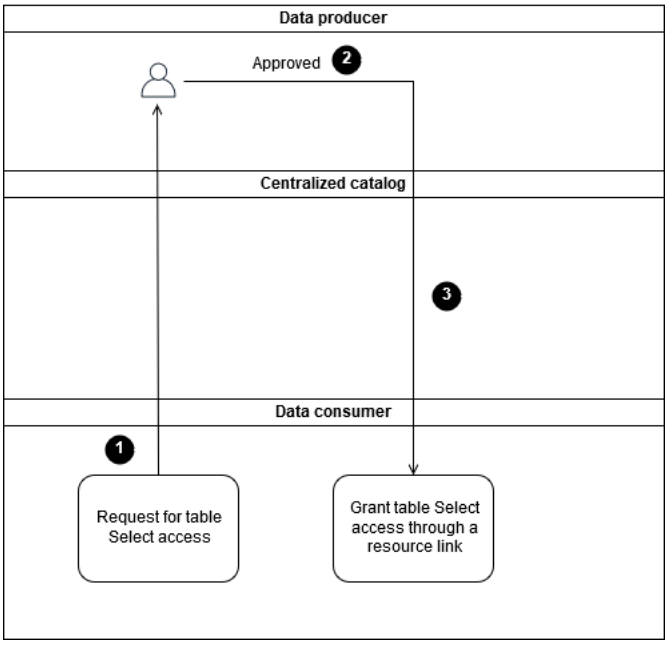
1. 数据使用者请求批准查看数据创建者的数据，并指定其需要访问的数据。
2. 数据生产者的数据管家会审查数据使用者的请求，并评估是否：
   - 共享所请求数据库中的部分或全部表。当与数据使用者共享所有表不会影响数据安全时，我们建议进行数据库级共享，这有助于避免表级共享的管理开销。
   - 在数据使用者的组织、OU 或账户级别共享。
3. 经数据创建者批准后，所需的数据目录资源将在集中式目录中与数据使用者共享。
4. 可以使用 Lake Formation 在数据使用者的账户中创建资源链接，然后指向集中式目录中的共享数据目录资源。

入职过程完成后，数据使用者的 Lake Formation 管理员可以从集中式目录和资源链接中查看数据库目录资源。在此阶段，数据使用者账户中的其他人无法访问数据生产者的数据。

在数据使用者账户中授予 Select 访问权限

下图显示了向数据使用者账户中的本地 IAM 委托人授予共享数据资源`Select`访问权限的过程。本地 IAM 委托人可以是个人用户的 IAM 角色，也可以是特定 AWS 服务使用的 IAM 角色。

> **Note**
>
> 当共享的数据敏感度较低时，您可以将访问权限授予数据使用者本身，而无需获得数据创建者的批准。这是因为他们之间已经建立了信任和共享。



此图显示以下流程：

1. 数据使用者账户中的个人 IAM 委托人请求数据使用者账户中的 IAM 委托人`Select`访问资源链接。
2. 数据生产者的数据管家会审查数据使用者的请求，并在满足所有要求时给予批准。
3. `Select`已授予访问权限，这允许 IAM 委托人使用请求的数据。

常见问题解答

本节提供了有关设计数据湖以实现增长和扩展的常见问题的答案 AWS Cloud。

此数据湖参考架构是否更适用于企业组织？

本指南的数据湖参考架构可以应用于属于任何规模组织的数据湖。该参考架构标准化了数据交换接口，降低了维护和发展数据湖的开销和成本，并且可以应用于组织数据湖发展到的任何规模。

如果我的组织只有一个数据创建器，我还能使用这个参考架构吗？

即使您的组织只有一个数据生成器，本指南的数据湖参考架构仍然具有相关性和益处。如果没有集中式目录，您的数据创建者就必须处理数据使用者的增长，这会增加复杂性和开销。您的数据湖也是您组织的长期资产，通常组织会增加更多的数据生产者。例如，出于合规性原因，或者因为您的组织收购了另一个拥有自己的数据创建器的业务部门，您可能需要一个额外的数据创建器来存储敏感数据。

我的数据湖直接将一个数据创建器与多个数据使用者连接起来。本指南的数据湖参考架构仍然相关吗？

从长远来看，数据湖参考架构将使您的组织受益。您可以使用两步方法，首先为新的数据使用者构建集中式目录。然后，您可以将现有的数据使用者连接到集中式目录。

我的组织是否应该在不进行更改的情况下遵循入职和访问权限授予工作流程？

不，该部分的主要目的是说明入职过程中所需的逻辑活动块。所有组织都应自定义流程，甚至可能有多个流程，具体取决于其数据的敏感度。

另一个考虑因素是，流程在中使用基于资源的共享方法。 AWS Lake Formation Lake Formation 还支持其他数据共享方法，例如基于标签的共享，在这种方法中，可以针对特定的共享方法量身定制流程中的差异。

资源

- AWS Glue 开发者指南
- AWS Lake Formation 开发者指南
- Lake Formation 中的跨账户访问
- 在 Lake Formation 中授予和撤消数据目录权限
- 填充 AWS Glue 数据目录

文档历史记录

下表介绍了本指南的一些重要更改。如果您希望收到有关未来更新的通知，可以订阅 RSS 源。

变更	说明	日期
一	初次发布	2021 年 10 月 1 日

本文属于机器翻译版本。若本译文内容与英语原文存在差异，则一律以英文原文为准。