Entwicklung eines Data Lakes für Wachstum und Skalierung auf der AWS Cloud

AWS Prescriptive Guidance

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

AWS Prescriptive Guidance: Entwicklung eines Data Lakes für Wachstum und Skalierung auf der AWS Cloud

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

Die Handelsmarken und die Handelsaufmachung von Amazon dürfen nicht in einer Weise in Verbindung mit nicht von Amazon stammenden Produkten oder Services verwendet werden, durch die Kunden irregeführt werden könnten oder Amazon in schlechtem Licht dargestellt oder diskreditiert werden könnte. Alle anderen Handelsmarken, die nicht Eigentum von Amazon sind, gehören den jeweiligen Besitzern, die möglicherweise zu Amazon gehören oder nicht, mit Amazon verbunden sind oder von Amazon gesponsert werden.

Table of Contents

Einführung 1
Gezielte Geschäftsergebnisse 2
Allgemeine Skalierungsherausforderungen 3
Erste Bereitstellung eines Data Lake 3
Die Zahl der Datennutzer nimmt zu 4
Die Zahl der Datenproduzenten steigt 5
Referenzarchitektur 7
Komponenten der Referenzarchitektur 9
Datenproduzenten 10
Zentralisierter Katalog 12
AWS Lake Formation 13
AWS Glue 14
Datenkonsumenten 15
Onboarding und Zugangsgewährung 19
Onboarding von Datenproduzenten 19
Einbindung von Datennutzern 20
Gewähren Sie Select-Zugriff in einem Datenverbraucherkonto 21
Häufig gestellte Fragen 23
Ist diese Data-Lake-Referenzarchitektur eher für Unternehmensorganisationen geeignet? 23
Kann ich diese Referenzarchitektur trotzdem verwenden, wenn meine Organisation nur über einen Datenproduzenten verfügt? 23
My Data Lake verbindet einen Datenproduzenten direkt mit mehreren Datenverbrauchern. Ist die Data-Lake-Referenzarchitektur dieses Handbuchs noch relevant? 23
Sollte mein Unternehmen den Arbeitsablauf für das Onboarding und die Zugriffsgewährung einhalten, ohne Änderungen daran vorzunehmen? 24
Ressourcen 25
Dokumentverlauf 26
..... xxvii

Entwicklung eines Data Lakes für Wachstum und Skalierung auf der AWS Cloud

Wei Shao, Amazon Web Services ()AWS

Oktober 2021 (Verlauf der Dokumente)

Organizations entwerfen und bauen im Rahmen ihrer Datenmodernisierungsstrategie zunehmend Data Lakes in der Amazon Web Services (AWS) Cloud. Data Lakes können Repositorys sein, in denen Ihre strukturierten und unstrukturierten Daten in beliebiger Größenordnung gespeichert und einer Vielzahl interner und externer Benutzer zur Verfügung gestellt werden.

Es kann jedoch viel Zeit und Mühe kosten, Daten in Data Lakes aufzunehmen, die Ihren wachsenden Geschäftsanwendungsfällen dienen. Um die Kosten zu senken und den aus den Daten generierten Wert zu maximieren, planen viele Unternehmen eine einmalige Aufnahme von Daten in ihren Data Lake und verwenden diese Daten anschließend mehrfach. Die Entwicklung einer Data-Lake-Architektur, die sich an die Produktion, gemeinsame Nutzung und Nutzung von Daten anpasst, ist entscheidend, um den Beteiligten im Unternehmen einen Mehrwert zu bieten, wenn Ihr Data Lake wächst.

Eine skalierbare Data-Lake-Architektur bietet Ihrem Unternehmen eine solide Grundlage, um aus Ihrem Data Lake Nutzen zu ziehen und gleichzeitig mehr Daten in ihn einzubeziehen. Durch die kontinuierliche Gewinnung von Dateneinblicken, ohne aufgrund von Skalierbarkeitsbeschränkungen gebremst oder unterbrochen zu werden, trägt ein skalierbarer Data Lake auch dazu bei, dass Ihr Unternehmen wettbewerbsfähig bleibt.

In der Regel gibt es in einem Data Lake Datenproduzenten und Datenkonsumenten. Datenproduzenten erstellen Datenbestände, indem sie Daten aus ihrer Datendomäne sammeln, verarbeiten und speichern. Diese kollektiven Datenbestände bilden den Inhalt Ihres Data Lakes. Datenproduzenten können sich dafür entscheiden, ihre Datenbestände selektiv mit den Datennutzern des Data Lakes zu teilen.

Datenkonsumenten benötigen die Daten der Datenproduzenten, um ihre geschäftlichen Anwendungsfälle zu erfüllen, und können diese Daten gelegentlich auch mit ihren eigenen Daten kombinieren. Datenproduzenten und Datenkonsumenten sind in der Regel, aber nicht immer, Teil Ihres Unternehmens. Wichtig ist, dass sie beide gleichzeitig Datenproduzenten oder Datenkonsumenten sein können.

Eine skalierbare Data-Lake-Architektur hilft Ihnen dabei, die folgenden Ergebnisse zu erzielen:

- Integrieren Sie Datenproduzenten in großem Umfang, ohne dass sie den gesamten Datenaustauschprozess aufrechterhalten müssen. Auf diese Weise können Datenproduzenten ihre Daten in den Data Lake integrieren und sich darauf konzentrieren, Daten aus ihrer Datendomäne zu sammeln, zu verarbeiten und zu speichern.
- Ermöglichen Sie Datenkonsumenten den Zugriff auf Daten von mehreren Datenproduzenten, ohne Ihre Gesamtkosten und Ihren Verwaltungsaufwand zu erhöhen.

In diesem Leitfaden werden die häufigsten Skalierungsherausforderungen beschrieben, die auftreten können, wenn Unternehmen ihren Data Lake erweitern, und es wird eine Referenzarchitektur für den Data Lake bereitgestellt. Außerdem werden Ansätze für das Onboarding und die Gewährung des Zugriffs für Datenproduzenten und Datenkonsumenten beschrieben. Die Data-Lake-Referenzarchitektur in diesem Leitfaden nutzt die verschiedenen Funktionen und Fähigkeiten von. AWS Lake Formation Der Leitfaden richtet sich an Teams, die für die Gestaltung von Data Lakes auf der Website verantwortlich sind AWS Cloud, darunter Unternehmensdatenarchitekten, Datenplattformarchitekten, Designer oder Data Domain-Leads.

Gezielte Geschäftsergebnisse

Wenn Sie einen Data Lake für Wachstum und Skalierung konzipiert haben, sollten Sie die folgenden drei Ergebnisse erwarten AWS Cloud:

- Geringere Gemeinkosten für Datenaustausch und Datenverbrauch in mehreren Geschäftsbereichen Ihres Unternehmens.
- Ein sicherer und konsistenter Ansatz, der Ihrem Unternehmen hilft, externe Datenproduzenten einzubeziehen und Daten mit ihnen in Ihrem Data Lake zu teilen.
- Gewinnen Sie kontinuierlich Einblicke in Ihre Daten, ohne aufgrund von Einschränkungen der Skalierbarkeit gebremst oder unterbrochen zu werden.

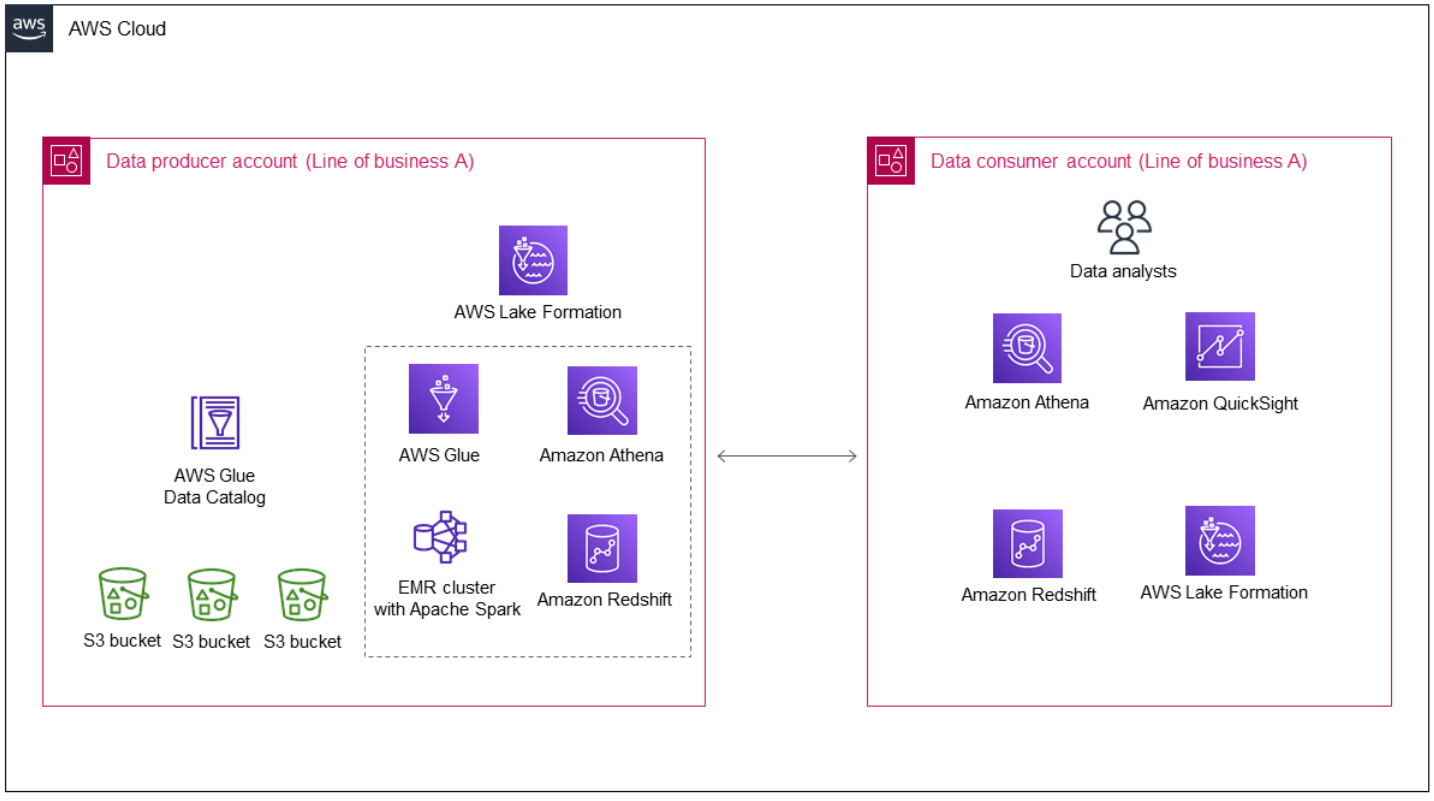
Allgemeine Skalierungsherausforderungen

Ein Data Lake durchläuft mehrere Phasen, in denen seine Datenmenge nach der ersten Bereitstellung anwächst. Wenn Sie bei der Gestaltung Ihres Data Lakes keine skalierbare Architektur verwendet haben, könnte Ihr Unternehmen vor Herausforderungen stehen und durch das Wachstum des Data Lakes benachteiligt werden.

In den folgenden Abschnitten wird erklärt, wie das Wachstum eines typischen Data Lakes zu Skalierungsproblemen führen kann.

Erste Bereitstellung eines Data Lake

Das folgende Diagramm zeigt die Architektur eines Data Lakes nach seiner ersten Bereitstellung durch den Geschäftsbereich A.



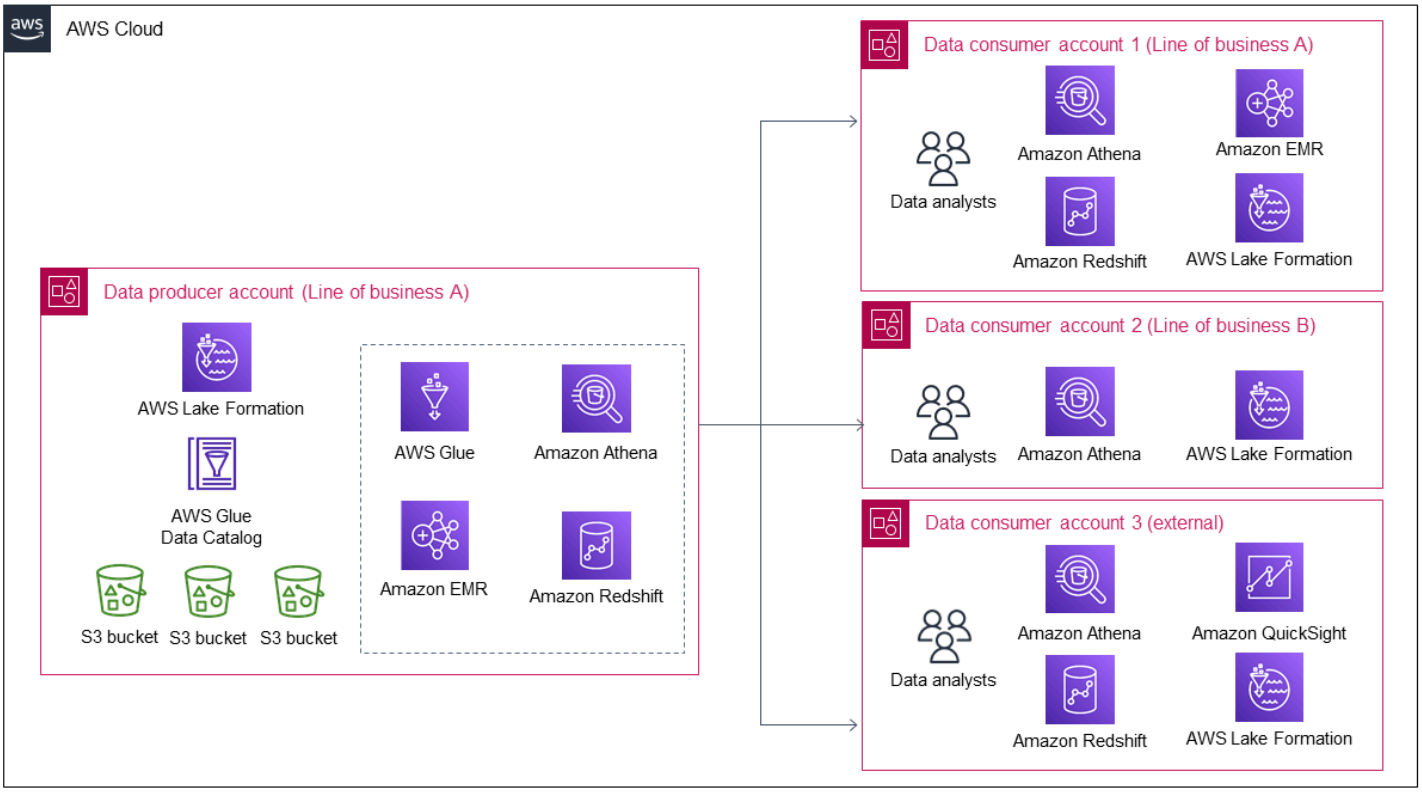
Das Diagramm zeigt die folgenden Komponenten:

- Das Konto des Datenproduzenten sammelt und verarbeitet Daten, speichert die verarbeiteten Daten und bereitet sie für den Verbrauch auf.

- Daten im Data Producer-Konto werden in Amazon Simple Storage Service (Amazon S3) -Buckets gespeichert, die mehrere Datenschichten haben können.
- Sie können AWS Dienste für die Datenverarbeitung verwenden (z. B. AWS Glueund Amazon EMR).
- Der Datenproduzent erzeugt und speichert nicht nur Daten im Data Lake, sondern muss dann auch entscheiden, welche Daten er mit einem Datenverbraucher teilen möchte und wie er sie teilt. AWS Lake Formation verwaltet den Data Lake im Datenproduzentenkonto und verwaltet zusätzlich den kontenübergreifenden Datenaustausch vom Datenproduzenten bis zum Datenverbraucher.
- Das Datenverbraucherkonto nutzt gemeinsam genutzte Daten aus dem Datenproduzentenkonto für spezifische geschäftliche Anwendungsfälle.

Die Zahl der Datennutzer nimmt zu

Das folgende Diagramm zeigt, dass mehr Daten in den Data Lake gelangen, wenn die Datenmenge von Geschäftsbereich A zunimmt. Der Data Lake zieht dann mehr Datenkonsumenten an, um die Daten zu nutzen und Mehrwert aus ihnen zu ziehen.

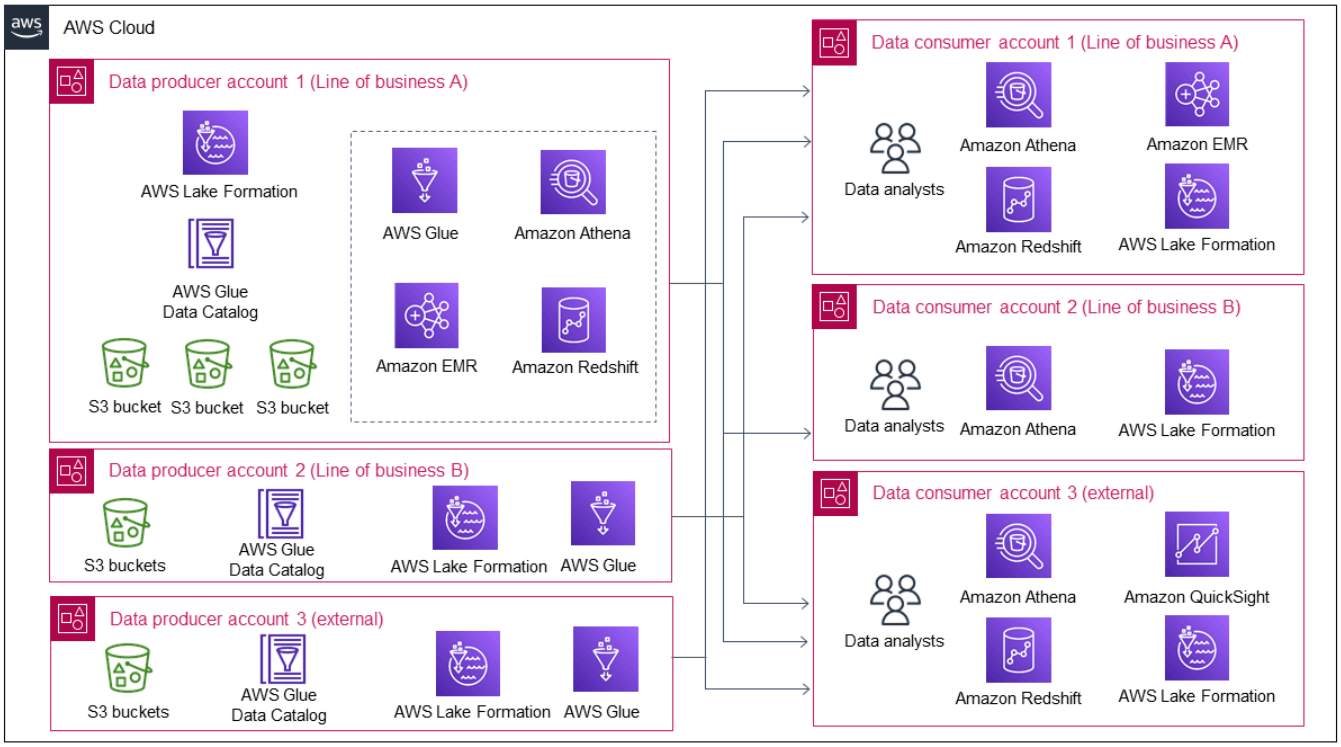


Das Diagramm zeigt, wie ein Unternehmen nahezu kontinuierlich Wert aus einem vorhandenen Datenbestand generiert und dass dies mehr Datennutzer anzieht. Bei einer Zunahme der Datenverbraucher stehen dem Datenproduzenten jedoch nur die folgenden zwei Optionen zur Verfügung, um diesem Wachstum Rechnung zu tragen:

- Manuelles Verwalten der gemeinsamen Nutzung von Daten und des Zugriffs durch einzelne Datenverbraucher, was kein skalierbarer Ansatz ist.
- Entwickeln Sie einen automatisierten oder halbautomatischen Prozess für die gemeinsame Nutzung von Daten und die Verwaltung des Datenzugriffs. Dies könnte zwar eine skalierbare Option sein, erfordert jedoch einen erheblichen Zeit- und Arbeitsaufwand bei der Planung und Erstellung, da interne und externe Datenverbraucher unterschiedliche Anforderungen an die Sicherheitskontrolle haben. In future wären auch zusätzliche Zeit und Mühe für alle Lösungsverbesserungen erforderlich.

Die Zahl der Datenproduzenten steigt

Das folgende Diagramm zeigt die Data-Lake-Architektur, wenn sich mehrere Geschäftsbereiche als Datenproduzenten zusammenschließen.



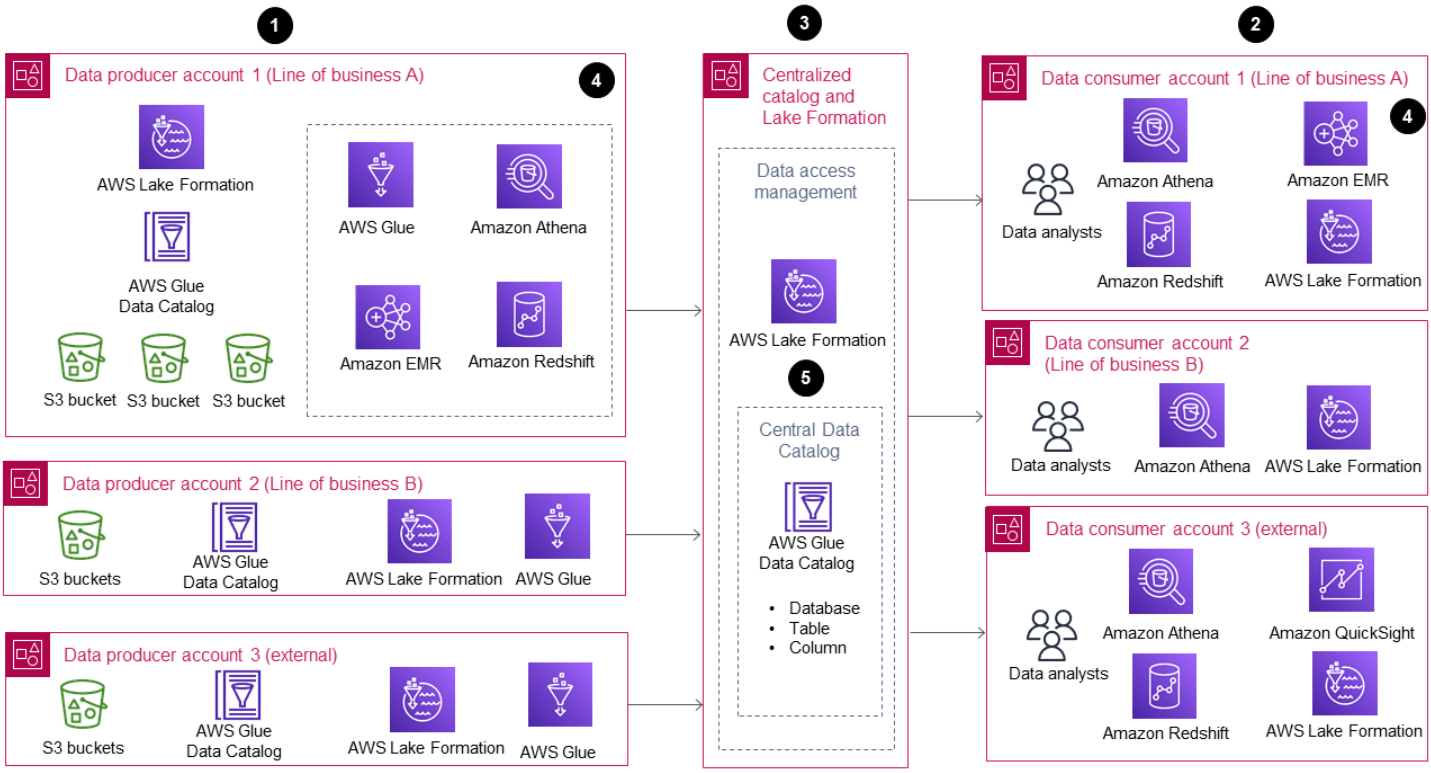
Die Architektur eines Data Lakes wird immer komplizierter, selbst wenn es nur drei Datenproduzenten und drei Datenkonsumenten gibt.

Jeder Datenproduzent muss sich um die gemeinsame Nutzung von Daten und das Datenzugriffsmanagement für mehrere Datenverbraucher kümmern. Es ist unrealistisch zu erwarten, dass alle Datenproduzenten ein automatisiertes oder halbautomatisiertes Verfahren für die gemeinsame Nutzung von Daten und das Datenzugriffsmanagement entwickeln. Einige Datenproduzenten entscheiden sich möglicherweise dafür, ihre Daten nicht weiterzugeben, um unerschwinglichen Verwaltungsaufwand zu vermeiden. In ähnlicher Weise muss jeder Datenverbraucher mit mehreren Datenproduzenten interagieren, um deren unterschiedliche Datennutzungsprozesse zu verstehen. Das bedeutet, dass einzelne Datenkonsumenten mit einem zunehmenden Verwaltungsaufwand konfrontiert sind, wenn es um den Umgang mit unterschiedlichen Mustern des Datenaustauschs geht.

In vielen Unternehmen führt dieser Data Lake zu Engpässen und kann weder wachsen noch skalieren. Dies kann bedeuten, dass Ihr Unternehmen seinen Data Lake neu entwerfen und neu aufbauen muss, um den Engpass zu beseitigen, was viel Zeit, Ressourcen und Geld kosten kann.

Referenzarchitektur

Das folgende Diagramm zeigt die Referenzarchitektur dieses Handbuchs für das Wachstum und die Skalierung eines Data Lakes auf dem AWS Cloud.



Das Diagramm zeigt die folgenden Komponenten:

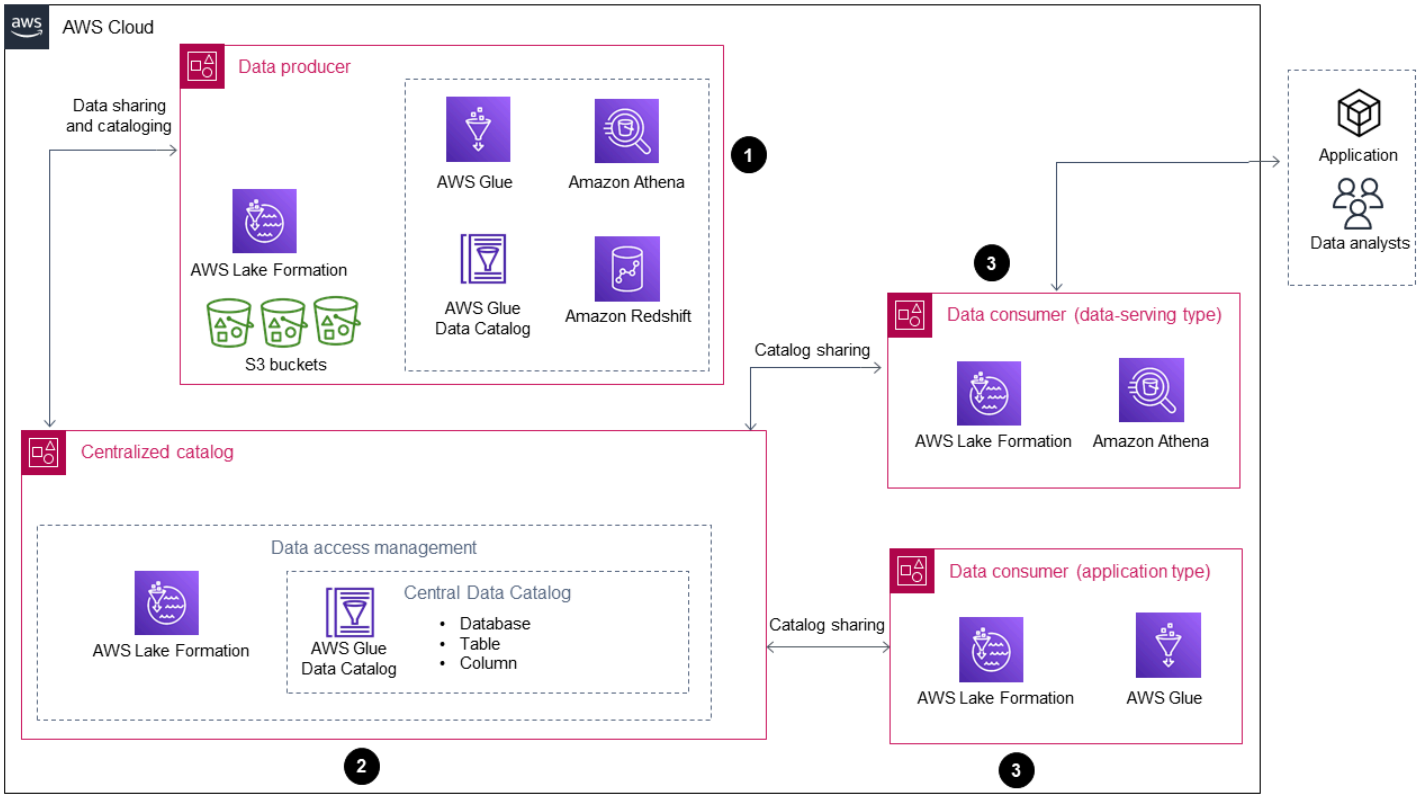
1. Eine Datenproduzentenebene ist unterschiedlich AWS-Konten.
2. Eine Datenverbraucherebene ist anders AWS-Konten.
3. Ein zentralisierter Katalog in einem AWS-Konto.
4. Obwohl es in jedem Geschäftsbereich nur einen Datenproduzenten und einen Datenverbraucher gibt, unterstützt die Referenzarchitektur des Leitfadens mehrere Datenproduzenten und Datenkonsumenten für jeden Geschäftsbereich. Es ist üblich, einen Datenproduzenten mit einem oder mehreren Datenverbrauchern zu verbinden, zu denen sowohl Datenbereitstellungs- als auch Anwendungstypen gehören. Weitere Informationen dazu finden Sie im Komponenten der Referenzarchitektur Abschnitt dieses Handbuchs.
5. Der zentralisierte Katalog ist die Schnittstelle, die von Datenproduzenten und Datenkonsumenten verwendet wird, um Daten auszutauschen und zu konsumieren.

Der Ansatz der Referenzarchitektur ermöglicht es, die gemeinsame Nutzung und Nutzung von Daten zu standardisieren und Datenproduzenten und Datenverbraucher unabhängig voneinander zu skalieren, ohne Ihren Verwaltungsaufwand zu erhöhen. Die Referenzarchitektur ermöglicht auch die Datenproduktion und -verteilung zwischen verschiedenen Datenproduzenten. Jeder Datenproduzent kann Teil des Data Lakes sein, seine Daten teilen und so zum Gesamtwert beitragen, den der Data Lake bietet.

Dieser Ansatz ermöglicht es Ihrem Unternehmen, Datenwerte in allen Geschäftsbereichen und externen Dateneigentümern zu nutzen, ohne dass dadurch Engpässe entstehen, indem die Datenerfassung und -verarbeitung in einer einzigen Pipeline eingeschränkt wird.

Komponenten der Referenzarchitektur

Das folgende Diagramm zeigt die drei Komponenten der Referenzarchitektur dieses Handbuchs: Datenproduzenten, Datenkonsumenten und den zentralisierten Katalog.



Das Diagramm zeigt die folgenden Komponenten:

1. Ein Datenproduzentenkonto.
2. Ein zentralisierter Katalog.
3. Zwei Datenverbraucherkonten, zu denen Datenbereitungs- und Anwendungstypen gehören.

> **Note**
>
> Das Diagramm zeigt nur einen Datenproduzenten und zwei Datenkonsumenten, aber die Referenzarchitektur dieses Handbuchs unterstützt mehrere Datenproduzenten und Datenkonsumenten, wenn der Data Lake wächst.

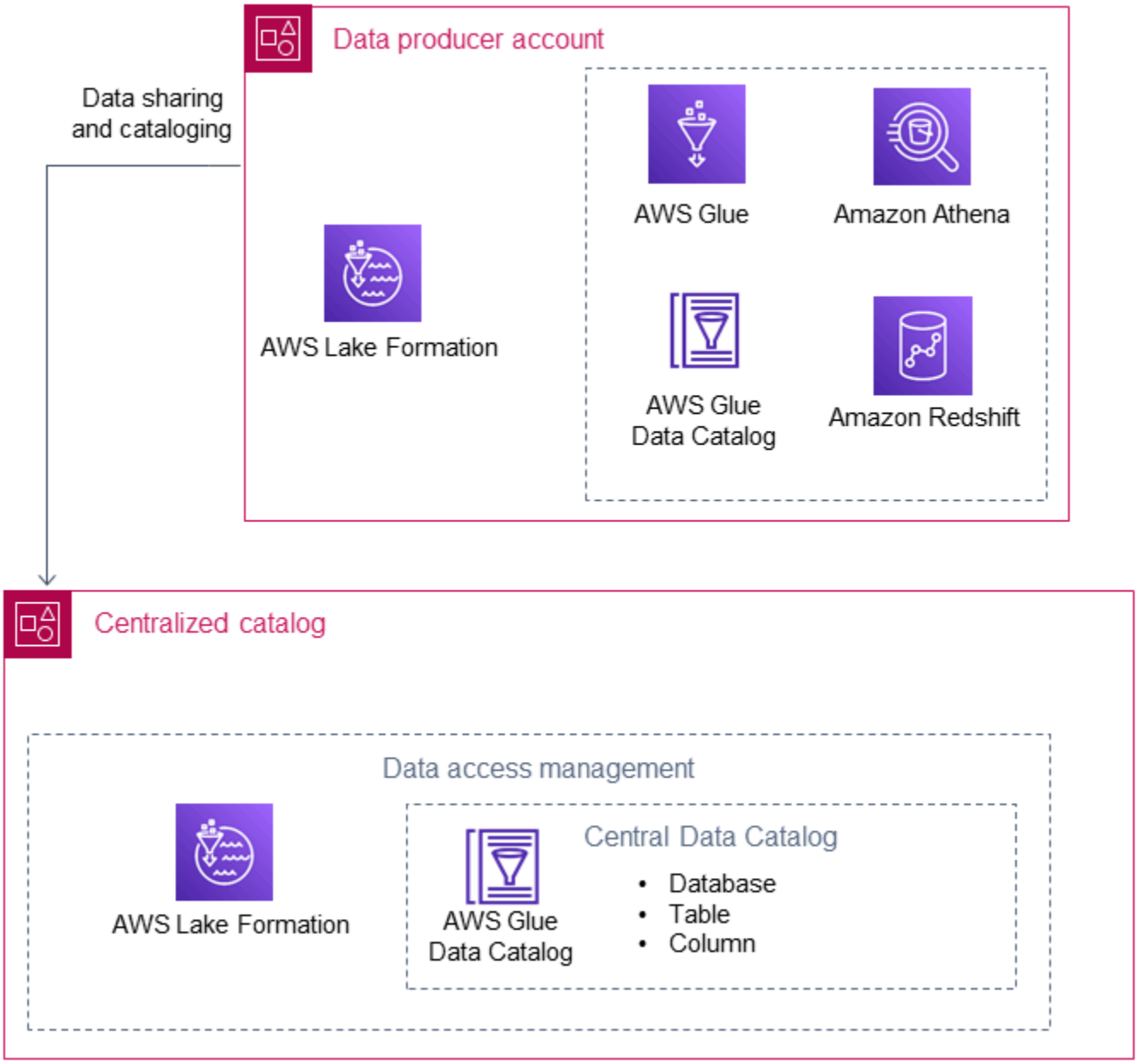
Das Verständnis der einzelnen Komponenten hilft Ihrem Unternehmen, wiederverwendbare Muster zu implementieren und Ihren Data Lake zu skalieren, indem neue Datenproduzenten und Datenkonsumenten unabhängig und konsistent eingebunden werden.

In den folgenden Abschnitten werden die Komponenten der Referenzarchitektur behandelt:

- Datenproduzenten
- Zentralisierter Katalog
- Datenkonsumenten

Datenproduzenten

Ein Datenproduzent sammelt, verarbeitet und speichert Daten aus seiner Datendomäne und überwacht und gewährleistet zusätzlich die Qualität seiner Datenbestände. Das folgende Diagramm zeigt das Datenproduzentenkonto als Bestandteil der Referenzarchitektur dieses Handbuchs.



Jeder Datenproduzent verfügt über einen eigenen Datenkatalog AWS-Konto , der von seinem internen Datenprozess verwaltet wird. AWS Lake Formation Datenproduzenten gewähren dem zentralisierten Katalog selektive Berechtigungen für ihre Daten, was bedeutet, dass Lake Formation im zentralisierten Katalogkonto auf Daten zugreifen kann, die der Datenproduzent teilen möchte.

Das bedeutet, dass Datenproduzenten nicht direkt mit Datenkonsumenten interagieren. Stattdessen werden das Konto des Datenproduzenten und sein Datenspeicherort vollständig abstrahiert und vor dem Datenverbraucher verborgen. Dieser Ansatz senkt die Kosten, indem unnötiger Aufwand für Datenproduzenten vermieden wird, die eine Zunahme ihrer Datenkonsumenten verzeichnen.

Eine Änderung des Datenstandorts des Datenproduzenten hat keine Auswirkungen auf den Datenverbraucher, wenn der neue Datenstandort im zentralen Katalog registriert wird. Wenn der Datenproduzent die gemeinsame Nutzung eines bestimmten Datenelements beenden möchte, kann er dem zentralisierten Katalog die Berechtigungen entziehen. Dadurch wird verhindert, dass Datenkonsumenten auf die Daten zugreifen können, und es ist nicht mehr erforderlich, den Zugriff für jeden Datenverbraucher manuell zu widerrufen.

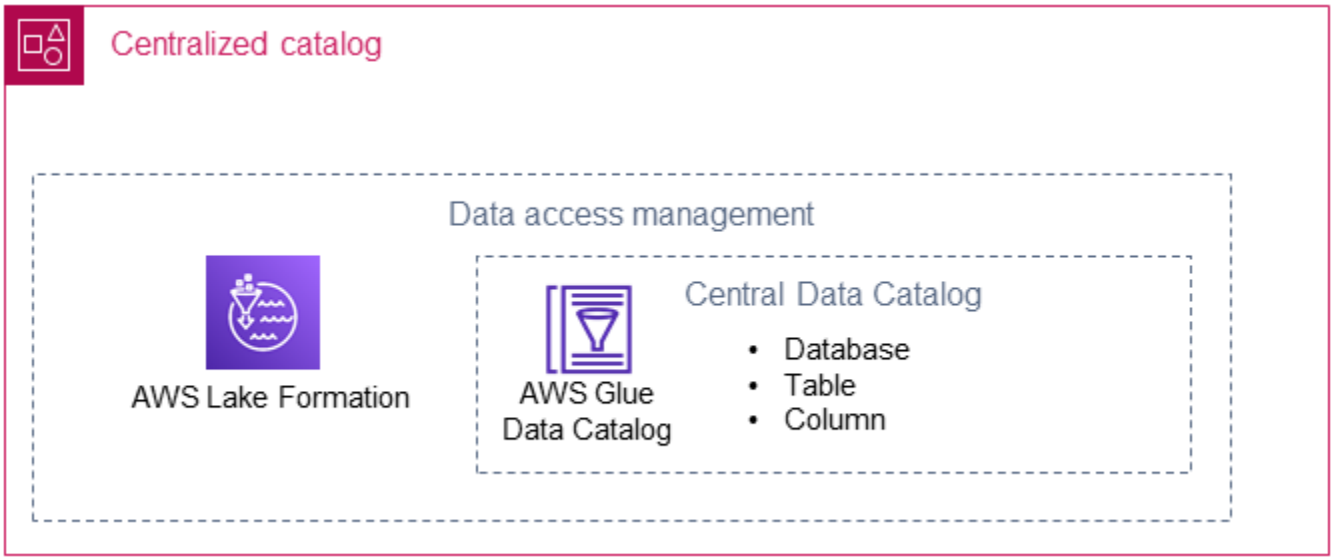
Durch die Verwendung von öffentlichen und privaten Datenkatalogen können Datenproduzenten wählen, was sie mit den Datenkonsumenten teilen möchten, und gleichzeitig den internen Datenzugriff über einen privaten Datenkatalog unabhängig verwalten.

In der folgenden Tabelle werden die beiden AWS Dienste beschrieben, die Datenproduzenten verwenden, um Daten mit dem zentralen Katalog gemeinsam zu nutzen.

Amazon Simple Storage Service (Amazon-S3)	Passen Sie die Bucket-Richtlinie für S3-Buckets an, um Datenzugriff auf die AWS Identity and Access Management (IAM-) Rollen im zentralisierten Katalog zu ermöglichen. Datenproduzenten können auch Daten, die in anderen Datenanwendungen oder Diensten gespeichert sind, gemeinsam nutzen, indem sie Amazon S3 als Zwischendatenschicht verwenden.
AWS Key Management Service (AWS KMS)	Stellen Sie Berechtigungen für die AWS verwalteten Schlüssel für die IAM-Rollen im zentralisierten Katalog und für die AWS KMS keys Verschlüsselung der gemeinsam genutzten Amazon S3 S3-Daten in den Datenproduzentenkonten bereit.

Zentralisierter Katalog

Das folgende Diagramm zeigt, wie der zentralisierte Katalog Datenproduzenten und Datenkonsumenten im Data Lake verbindet.



Der zentrale Katalog speichert und verwaltet den gemeinsamen Datenkatalog für die Konten der Datenproduzenten. Der zentrale Katalog enthält auch die technischen Metadaten der gemeinsam genutzten Daten (z. B. Tabellenname und Schema) und ist der Ort, an dem Datenkonsumenten auf Daten zugreifen.

Datenverbraucher können auf Daten von mehreren Datenproduzenten im zentralisierten Katalog zugreifen und diese Daten dann zur weiteren Verarbeitung mit ihren eigenen Daten mischen. Durch die Verwendung eines zentralisierten Katalogs müssen Datenkonsumenten keine direkte Verbindung zu verschiedenen Datenproduzenten herstellen, und der Betriebsaufwand wird reduziert.

Da der zentralisierte Katalog Einblick in die gemeinsame Nutzung und Nutzung von Daten durch Datenproduzenten und -verbraucher bietet, kann er ein idealer Ort sein, um Ihre zentralen Datenverwaltungsfunktionen (z. B. Zugriffsprüfung) anzuwenden.

In den folgenden Abschnitten wird beschrieben, wie der zentralisierte Katalog AWS Lake Formation und verwendet AWS Glue.

AWS Lake Formation

AWS Lake Formationhilft bei der Erstellung von Datenbanken in einem AWS Glue Datenkatalog, die auf die Standorte mehrerer Datenproduzenten in Ihrem Data Lake verweisen. Eine AWS Identity and Access Management (IAM-) Rolle wird für Lake Formation im zentralisierten Katalog erstellt. Durch die Verwendung von Lake Formation kann der zentralisierte Katalog Datenressourcen (z. B. Datenbanken, Tabellen oder Spalten) selektiv mit Datenverbrauchern teilen. Die von Lake

Formation verwalteten Ressourcen werden mithilfe einer der folgenden beiden Methoden mit Datenverbrauchern gemeinsam genutzt:

- Methode mit benannter Ressource — Bei dieser Methode werden verwaltete Ressourcen von mehreren Konten gemeinsam genutzt. Datenbanken, Tabellen oder Spaltennamen müssen angegeben werden, und eine Ressource kann für eine Organisation, Organisationseinheit (OU) oder gemeinsam genutzt werden AWS-Konto. Um den Gemeinschafts- und Verwaltungsaufwand zu reduzieren, empfehlen wir, Ressourcen nach Möglichkeit auf höheren Ebenen gemeinsam zu nutzen (z. B. in einer Organisation oder Organisationseinheit statt in einer AWS-Konto). Sie müssen jedoch sicherstellen, dass dieser Ansatz die Anforderungen Ihrer Organisation an die Datensicherheitskontrolle erfüllt.
  - Hinweis: Diese Methode eignet sich gut für Datenkonsumenten mit einem Anwendungstyp, bei dem AWS Dienste Daten vom Datenproduzenten verwenden. Die Datenzugriffsanforderungen für diese Art von Datenkonsumenten sind anwendungsabhängig, präskriptiv und relativ statisch.
- Tag-basierte Zugriffskontrollmethode (LF-TBAC) von Lake Formation — LF-TBAC ist besonders nützlich für Datenkonsumenten mit einem Datenbereitungstyp. Ressourcen, die mit Lake Formation gekennzeichnet sind, können derzeit jedoch nur auf AWS-Konto Ebene und nicht auf Organisations- oder OU-Ebene gemeinsam genutzt werden.

AWS Glue

Sie müssen Datenbanken AWS Glue für jeden Datenproduzenten in Ihrem zentralisierten Katalog erstellen. Da der zentralisierte Katalog AWS Glue Datenbanken aller Datenproduzenten hostet, müssen Sie sicherstellen, dass der Datenbankname für alle Datenproduzenten eindeutig ist und dass er den Datenproduzenten und dessen Datentyp widerspiegelt. Sie können beispielsweise die folgende Datenbankbenennungsstruktur verwenden: `<Data_Producer>-<Environment>-<Data_Group>`

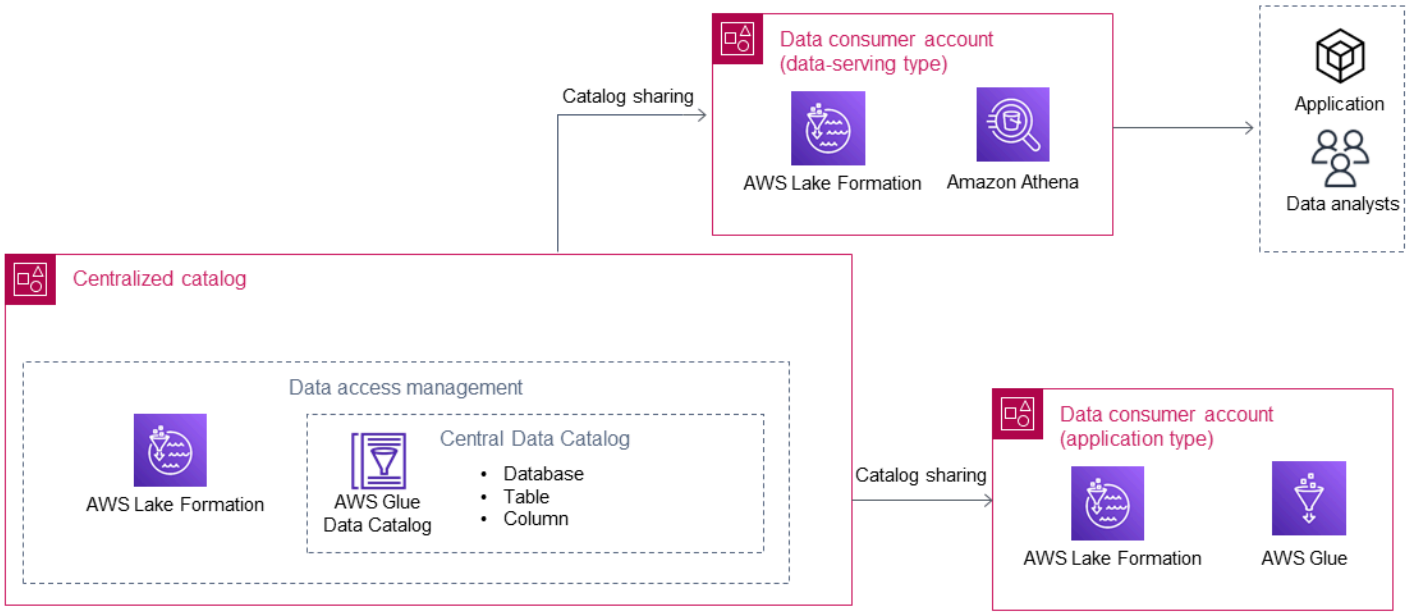
- `<Data_Producer>`— Der Name des Datenproduzenten.
- `<Environment>`— Die Data-Lake-Umgebung, z. B. `dev` für eine Entwicklungsumgebung, `sit` für eine Systemintegrationstestumgebung oder `prod` für eine Produktionsumgebung.
- `<Data_Group>`— Der Name der Datengruppe, die verwendet wird, um Daten von einem Datenproduzenten in logische Gruppen zu unterteilen. Sie können den Namen, die ID oder die Abkürzung des Quellsystems als Namen verwenden. Eine Datenbankbeschreibung hilft auch dabei, den Inhalt und den Zweck der Datenbank zu beschreiben.

Sie können einen AWS Glue Crawler für die Daten des Datenproduzenten verwenden, um dessen Schema in der Datenbank des zentralen Katalogs zu verwalten. Wenn Daten regelmäßig mit derselben Frequenz von einem Datenproduzenten erstellt werden, können Sie einen einzelnen AWS Glue Crawler verwenden. In allen anderen Fällen sollten Sie mehrere AWS Glue Crawler verwenden, um unterschiedliche Crawling-Frequenzen zu berücksichtigen. Je nach geschäftlichem Anwendungsfall kann der Crawler entweder für eine vordefinierte Frequenz geplant oder durch Ereignisse initiiert werden.

Sie können das Tabellenschema auch verwalten, AWS Glue indem Sie die AWS Glue API aufrufen, um das Schema zu erstellen oder zu aktualisieren. Dies kann zwar für Flexibilität sorgen, es ist jedoch zusätzlicher Aufwand für die Codeentwicklung und -wartung erforderlich. Stellen Sie sicher, dass Sie den Anwendungsfall und den geschäftlichen Nutzen bewerten und dann die Option wählen, die Ihren Anforderungen entspricht und den geringsten Aufwand verursacht.

Datenkonsumenten

Datenkonsumenten konsumieren die Daten vom Datenproduzenten, nachdem sie vom zentralen Katalog gemeinsam genutzt wurden AWS Lake Formation. Das folgende Diagramm zeigt zwei Datenverbraucher im Data Lake.



Es gibt zwei Arten von Datenkonsumenten: Anwendung und Datenbereitstellung. In der folgenden Tabelle werden diese beiden Typen beschrieben.

Art der Anwendung	Nutzer von Anwendungsdaten führen Anwendungen eigenständig aus AWS-Konten. Die Anwendungen nutzen die AWS Identity and Access Management (IAM-) Rollen, um auf die gemeinsam genutzten Daten eines Datenprod uzenten zuzugreifen und sie dann entsprech end ihrer Logik zu verarbeiten. Typischerweise hat diese Art von Datenkons umenten vorgeschriebene Datenanforderungen , um die Anforderungen einer Anwendung zu erfüllen.
Data-serving Typ	Data-serving Datenkonsumenten sind in der Regel für Einzelpersonen (z. B. Datenanalysten oder Datenwissenschaftler) und Anwendungen (z. B. eine Business Intelligence-Anwendung) gedacht, für die es keine eigenen gibt. AWS-Konten Im Data Lake eines Unternehmens können mehrere Datenkonsumenten existieren, die Daten bereitstellen. Beispielsweise könnten sich verschiedene Geschäftsbereiche dafür entscheiden, ihre eigenen Datenverbraucher einzurichten, um Benutzern zu helfen, Daten aus dem Data Lake zu nutzen. Für diese Datenverbraucher sind eigene IAM-Rolle nprinzipale konfiguriert AWS-Konto (z. B. zugeordnete IAM-Rollen AWS IAM Identity Center), die von Endbenutzern im Datenverb raucherkonto für den Zugriff auf gemeinsam genutzte Daten über AWS Dienste (z. B. Amazon Athena) verwendet werden.

	Typischerweise hat diese Art von Datenkonsumenten weitreichende und kontinuierlich steigende Datenanforderungen.

AWS Lake Formation ist der wichtigste AWS Dienst, der von einem Datenverbraucher für die kontenübergreifende gemeinsame Nutzung von Daten und den Zugriff auf den zentralen Katalog genutzt wird. Nachdem die Datenbanken vom zentralisierten Katalog gemeinsam genutzt wurden, sind die gemeinsam genutzten Ressourcen in Lake Formation im Datenverbraucherkonto verfügbar. Der Datenzugriff kann dann den lokalen IAM-Prinzipalen im Datenverbraucherkonto gewährt werden, sofern erforderlich, mit Genehmigung des Datenproduzenten. Die gemeinsam genutzten Daten können dann von AWS Diensten verwendet werden, die in Lake Formation integriert sind (z. B. Amazon Athena und AWS Glue). Sie können die folgenden AWS Dienste verwenden, um auf gemeinsam genutzte Daten im Datenverbraucherkonto zuzugreifen:

- Amazon Athena ist ein interaktiver Abfrageservice, mit dem Daten in Amazon Simple Storage Service (Amazon S3) mithilfe von Standard-SQL direkt analysiert werden können. Weitere Informationen zu Athena und Lake Formation finden Sie in der Amazon Athena-Dokumentation unter So greift Athena auf bei Lake Formation registrierte Daten zu.
- Amazon Redshift Spectrum hilft Ihnen dabei, strukturierte und halbstrukturierte Daten effizient aus Dateien in Amazon S3 abzufragen und abzurufen, ohne die Daten in Amazon Redshift Redshift-Tabellen laden zu müssen. Weitere Informationen zu Redshift Spectrum und Lake Formation finden Sie unter Using Redshift Spectrum with Lake Formation in der Amazon Redshift Redshift-Dokumentation.
- AWS Glueist ein vollständig verwalteter ETL-Service (Extrahieren, Transformieren und Laden), mit dem Sie Ihre Daten einfach und kostengünstig kategorisieren, bereinigen, anreichern und zuverlässig zwischen verschiedenen Datenspeichern und Datenströmen verschieben können. Die einem AWS Glue ETL-Job zugeordnete IAM-Rolle kann auf die von Lake Formation verwalteten Data Lake-Daten zugreifen, sofern sie über die erforderlichen Zugriffsberechtigungen verfügt.
- Amazon EMR unterstützt den Betrieb von Big-Data-Frameworks (z. B. Apache Hadoop und Apache Spark) zur Verarbeitung und Analyse großer Datenmengen. Weitere Informationen zu Amazon EMR und Lake Formation finden Sie unter Integrieren von Amazon EMR in Lake Formation in der Amazon EMR-Dokumentation.
- Amazon Quick ist ein skalierbarer, serverloser, einbettbarer und auf maschinellem Lernen (ML) basierender Business Intelligence-Service, mit dem Sie Daten aus Ihrem Data Lake analysieren

und visualisieren können. Weitere Informationen zu Quick und Lake Formation finden Sie in der Schnelldokumentation unter Autorisieren von Verbindungen über Lake Formation.

- Amazon SageMaker AI Data Wrangler (Data Wrangler) reduziert den Zeitaufwand für die Aggregation und Vorbereitung von Daten für ML. Weitere Informationen zu Data Wrangler und Lake Formation finden Sie unter Vorbereiten von ML-Daten mit Amazon SageMaker AI Data Wrangler in der Amazon AI-Dokumentation. SageMaker

Onboarding und Zugangsgewährung

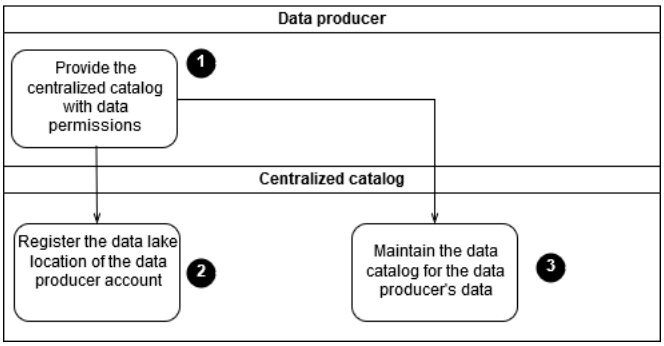
Die Data-Lake-Referenzarchitektur dieses Leitfadens hilft Ihnen dabei, Datenproduzenten und Datenkonsumenten unabhängig voneinander zu skalieren und einen konsistenten Prozess für das Onboarding und die Gewährung des Zugriffs für diese Datenkonsumenten zu definieren und einzurichten.

In den folgenden Abschnitten wird der Onboarding-Prozess für Datenproduzenten und Datenkonsumenten sowie die Gewährung des Zugriffs über ein Datenverbraucherkonto beschrieben. In diesem Handbuch wird die Methode der benannten Ressource zwischen dem zentralisierten Katalog und den Datenverbrauchern verwendet. Der Prozess für die LF-TBAC Methode ist ähnlich, unterscheidet sich jedoch geringfügig. Wir empfehlen Ihnen, diese Ansätze zu evaluieren und zu konfigurieren, damit sie den Datenverwaltungspraktiken und -richtlinien Ihres Unternehmens entsprechen.

Weitere Informationen zu diesen beiden Methoden finden Sie im Zentralisierter Katalog Abschnitt dieses Handbuchs.

Onboarding von Datenproduzenten

Das folgende Diagramm zeigt, wie Sie einen neuen Datenproduzenten in Ihren Data Lake integrieren.



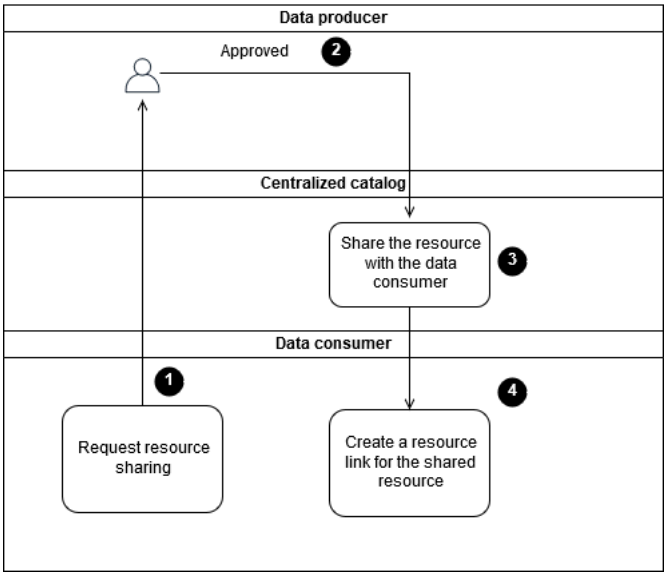
Das Diagramm zeigt den folgenden Onboarding-Prozess:

1. Der Datenproduzent gewährt dem zentralisierten Katalog selektiv Zugriff auf seine Daten (z. B. einen Amazon Simple Storage Service (Amazon S3) -Bucket und AWS KMS key). Den Principals des zentralisierten Katalogs AWS Identity and Access Management (IAM) wird Zugriff gewährt, um den Data Lake-Standort des Datenproduzenten zu registrieren, AWS Lake Formation sowie die IAM-Prinzipale, die zur Verwaltung des Katalogs des Datenproduzenten verwendet werden.

2. Registrieren Sie den Data Lake-Standort des Datenproduzenten (z. B. einen S3-Bucket), der die Lake Formation des zentralisierten Katalogs verwendet.
3. Erstellen Sie die Datenbank, Tabellen und Tabellenschemas für die neuen Daten aus dem Datenproduzenten im AWS Glue Datenkatalog.

Einbindung von Datennutzern

Das folgende Diagramm zeigt, wie Sie einen neuen Datenverbraucher in Ihren Data Lake integrieren können.



Das Diagramm zeigt den folgenden Onboarding-Prozess:

1. Der Datenverbraucher bittet um die Genehmigung, die Daten des Datenproduzenten einsehen zu dürfen, und gibt an, auf welche Daten er zugreifen muss.
2. Der Datenverwalter des Datenproduzenten prüft die Anfrage des Datenkonsumenten und bewertet, ob:
   - Geben Sie einige oder alle Tabellen in den angeforderten Datenbanken frei. Wir empfehlen die gemeinsame Nutzung auf Datenbankebene, wenn die gemeinsame Nutzung aller Tabellen für den Datenverbraucher keine Auswirkungen auf die Datensicherheit hat, wodurch der Verwaltungsaufwand vermieden wird, der durch die gemeinsame Nutzung auf Tabellenebene entsteht.
   - Teilen Sie Daten auf Organisations-, OU- oder Kontoebene des Datennutzers.

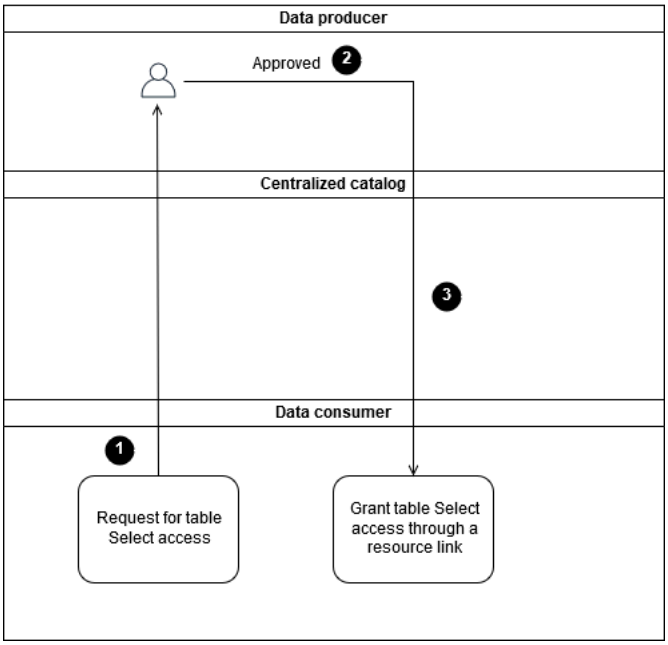
3. Nach der Genehmigung durch den Datenproduzenten werden die erforderlichen Datenkatalogressourcen im zentralisierten Katalog gemeinsam mit dem Datenverbraucher genutzt.
4. Ressourcenlinks können mithilfe von Lake Formation im Konto des Datenverbrauchers erstellt werden und dann auf die gemeinsam genutzten Datenkatalogressourcen im zentralen Katalog verweisen.

Nach Abschluss des Onboarding-Vorgangs kann der Lake Formation-Administrator des Datenkonsumenten die Datenbankkatalogressource im zentralen Katalog und im Ressourcenlink sehen. Zu diesem Zeitpunkt kann niemand anderes Mitglied des Kontos des Datenkonsumenten auf die Daten des Datenproduzenten zugreifen.

Gewähren Sie Select-Zugriff in einem Datenverbraucherkonto

Das folgende Diagramm zeigt den Prozess zur Gewährung des `SelectZugriffs auf` gemeinsam genutzte Datenressourcen mit einem lokalen IAM-Prinzipal im Datenverbraucherkonto. Der lokale IAM-Prinzipal kann die IAM-Rolle für einzelne Benutzer oder eine IAM-Rolle sein, die von bestimmten Diensten genutzt wird. AWS

> **Note**
>
> Wenn die gemeinsam genutzten Daten wenig sensibel sind, können Sie die Zugriffsgewährung an den Datenverbraucher selbst delegieren, ohne dass die Zustimmung des Datenproduzenten erforderlich ist. Dies liegt daran, dass zwischen ihnen bereits Vertrauen und gemeinsame Nutzung bestehen.



Das Diagramm zeigt den folgenden Prozess:

1. Der einzelne IAM-Prinzipal im Datenverbraucherkonto fordert vom IAM-Prinzipal im Datenverbraucherkonto `Select` Zugriff auf den Ressourcenlink an.
2. Der Datenverwalter des Datenproduzenten prüft die Anfrage des Datenkonsumenten und erteilt die Genehmigung, wenn alle Anforderungen erfüllt sind.
3. `Select`Der Zugriff wird gewährt, sodass der IAM-Principal die angeforderten Daten nutzen kann.

Häufig gestellte Fragen

Dieser Abschnitt enthält Antworten auf häufig gestellte Fragen zur Gestaltung eines Data Lakes für Wachstum und Skalierung auf dem AWS Cloud.

Ist diese Data-Lake-Referenzarchitektur eher für Unternehmensorganisationen geeignet?

Die Data-Lake-Referenzarchitektur dieses Handbuchs kann auf Data Lakes angewendet werden, die Unternehmen jeder Größe gehören. Die Referenzarchitektur standardisiert die Datenaustauschschnittstelle, senkt den Aufwand und die Kosten für die Wartung und Erweiterung des Data Lakes und kann auf jede Größenordnung angewendet werden, bis zu der der Data Lake Ihres Unternehmens wächst.

Kann ich diese Referenzarchitektur trotzdem verwenden, wenn meine Organisation nur über einen Datenproduzenten verfügt?

Die Data-Lake-Referenzarchitektur dieses Handbuchs ist immer noch relevant und nützlich, auch wenn Ihr Unternehmen nur über einen Datenproduzenten verfügt. Ohne den zentralisierten Katalog muss Ihr Datenproduzent das Wachstum der Datenkonsumenten bewältigen, was die Komplexität und den Mehraufwand erhöht. Ihr Data Lake ist auch ein langfristiger Vorteil für Ihr Unternehmen, und in der Regel fügen Unternehmen mehr Datenproduzenten hinzu. Beispielsweise benötigen Sie möglicherweise einen zusätzlichen Datenproduzenten, um sensible Daten aus Compliance-Gründen zu speichern oder weil Ihr Unternehmen eine weitere Geschäftseinheit mit eigenem Datenproduzenten erwirbt.

My Data Lake verbindet einen Datenproduzenten direkt mit mehreren Datenverbrauchern. Ist die Data-Lake-Referenzarchitektur dieses Handbuchs noch relevant?

Die Data-Lake-Referenzarchitektur würde Ihrem Unternehmen langfristig zugute kommen. Sie könnten einen zweistufigen Ansatz wählen und damit beginnen, den zentralen Katalog für neue Datennutzer zu erstellen. Anschließend könnten Sie Ihre vorhandenen Datenverbraucher mit dem zentralen Katalog verbinden.

Sollte mein Unternehmen den Arbeitsablauf für das Onboarding und die Zugriffsgewährung einhalten, ohne Änderungen daran vorzunehmen?

Nein, der Hauptzweck dieses Abschnitts besteht darin, die logischen Aktivitätsblöcke zu veranschaulichen, die während des Onboarding-Prozesses erforderlich sind. Alle Organisationen sollten den Prozess individuell anpassen und je nach Sensibilität ihrer Daten sogar mehrere Prozesse haben.

Eine weitere Überlegung ist, dass der Prozessablauf den Ansatz der ressourcenbasierten gemeinsamen Nutzung von verwendet. AWS Lake Formation Es gibt auch andere Methoden zum Datenaustausch, die von Lake Formation unterstützt werden, wie z. B. das tagbasierte Teilen, bei dem die Unterschiede im Prozess auf die jeweilige Freigabemethode zugeschnitten werden können.

Ressourcen

- AWS Glue Leitfaden für Entwickler
- AWS Lake Formation Leitfaden für Entwickler
- Kontoübergreifender Zugriff in Lake Formation
- Erteilung und Widerruf von Datenkatalogberechtigungen in Lake Formation
- Den Datenkatalog AWS Glue auffüllen

Dokumentverlauf

In der folgenden Tabelle werden wichtige Änderungen in diesem Leitfaden beschrieben. Um Benachrichtigungen über zukünftige Aktualisierungen zu erhalten, können Sie einen RSS-Feed abonnieren.

Änderung	Beschreibung	Datum
—	Erste Veröffentlichung	1. Oktober 2021

Die vorliegende Übersetzung wurde maschinell erstellt. Im Falle eines Konflikts oder eines Widerspruchs zwischen dieser übersetzten Fassung und der englischen Fassung (einschließlich infolge von Verzögerungen bei der Übersetzung) ist die englische Fassung maßgeblich.